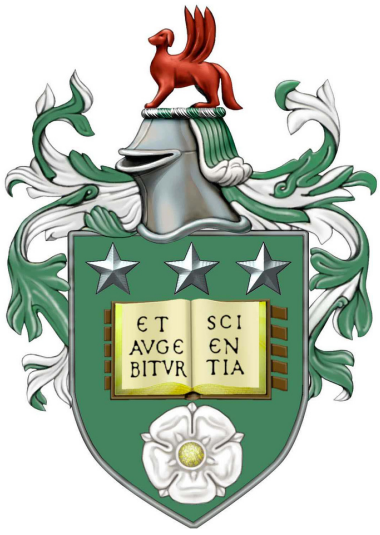
Predicting melanoma patient outcomes using digital pathology

Lucy Godson

Submitted in accordance with the requirements
for the degree of Doctor of Philosophy



The University of Leeds
School of Computing
January 2025

The candidate confirms that the work submitted is her own, except where work which has formed part of a jointly authored publication has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some parts of the work presented in Chapters 4 and 5 have been published in the following articles:

Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Julia Newton-Bishop, and Ali Gooya. "Weakly-supervised learning for image-based classification of primary melanomas into genomic immune subgroups." In Proceedings of The 5th International Conference on Medical Imaging with Deep Learning, 423–440. PMLR, December 2022.

Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Julia Newton-Bishop, Ali Gooya, and Derek Magee. "Immune subtyping of melanoma whole slide images using multiple instance learning." Medical Image Analysis 93, April 2024: 103097. ISSN: 1361-8415. DOI: 10.1016/j.media.2024.103097.

Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Julia Newton-Bishop, and Derek Magee. "Multi-level Graph Representations of Melanoma Whole Slide Images for Identifying Immune Subgroups." In Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology, edited by Seyed-Ahmad Ahmadi and Sérgio Pereira, 85–96. Cham: Springer Nature Switzerland, 2024. DOI: 10.1007/978-3-031-55088-1_8.

The above publications are primarily the work of the candidate.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis maybe published without proper acknowledgement.
©2025 The University of Leeds and Lucy Godson

Acknowledgements

Firstly, I would like to thank my supervisors, Dr. Derek Magee, Dr. Navid Alemi, Dr. Jeremie Nsengimana and Prof. Graham Cook. Derek, your ideas and encouragement kept me motivated and helped me to grow into a more skilled and independent researcher. Navid, I'm so glad you became a part of this project; your skill as a researcher and guidance during the tough parts of the pandemic helped me immeasurably. Jeremie and Graham, your wealth of knowledge about the Leeds Melanoma Cohort dataset and melanoma greatly enhanced my understanding of the research problem. I also want to thank Prof. Ali Gooya for initiating my PhD journey and pointing me in the right direction. I am grateful to Dr. Emily Clarke for her invaluable histopathological insight, interpretations, and general advice. Thank you to Prof. Julia Newton-Bishop and Prof. Tim Bishop for taking the time to explain the Leeds Melanoma Cohort to me and for helping me refine my research direction in the early stages of my PhD. I am also very appreciative of my PhD examiners, Prof. Fayyaz Minhas and Dr. Nishant Ravikumar, for their insightful and challenging questions during the viva, and to Prof. David Hogg for chairing the examination.

I am deeply thankful to the Centre for Doctoral training directors, management and students, especially Prof. David Hogg, Prof. Vania Dimitrova and Owen Johnson. Thank you for organising the many seminars, conferences and activities that gave us the opportunity to discuss ideas and address challenges throughout the PhD journey. I would also like to thank Dr. Alex Wright, Prof. Darren Treanor, and the NPIC team for organising the digital pathology seminars and for their advice and support during the write-up stages of my thesis, as well as in preparation for the viva.

I am extremely grateful to my School of Computing colleagues and friends, especially Anna Linton, Sam Llanwarne, Dr. Yash Deo and Dr. Rodrigo Bonazzola, for provoking important scientific (and not so scientific) discussions! Additionally, thank you to Volodomyr Chapman, Joe Sims and Dr. Jack Breen, who were always happy to recommend or point me in the direction of a helpful computational pathology resource.

I would like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) for funding this research [EP/S024336/1]. This work also utilised the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC [EP/T022167/1]. I am deeply appreciative of the Research Computing team at Leeds, as part of this work was undertaken on ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. Special thanks to John Hodrien and Dr. Alex Coleman for helping me navigate through dependency hell!

I am also very thankful to the Leeds Melanoma Cohort patients; this work would not have been possible without their involvement and generosity in providing data for this research. I would also like to thank the patients from the The Cancer Genome Atlas, as the results shown in this thesis are in part based upon data generated by the The Cancer Genome Atlas Research Network.

On a more personal note, my friends outside of my PhD life have been incredibly important to me and have supported me throughout this time. I am so grateful to the Final Five: Cinthu Vivehananthan, Ella Luo, Emma Woodhouse, and Tabassum Iqbal—each of you has listened to me rant or complain and has made me laugh during the stressful times. Marilena Onisiforou, Adriana Whittle, Flinn O'Hara, Alex Close, Billie Matchett-Byrd, Emily Sweeting, Chelsea Issott, Flavian Kow, Kirsty Blair and George Sellers: I am so grateful for the phone calls, memes, meet-ups or random messages of support – they kept me going throughout these years! A special thank you to Dr. Petar Hristov for all the Marmite bagels, cups of coffee, rubber-ducking sessions, laughter, late-night chats, and, most importantly, for always being there for me. Your continual love and support helped me make it to the end.

Finally, I would like to thank my family - the Godsons, Morritts, Claytons and Crothers - who have offered me unwavering support throughout my life. Thank you to my big brother, Harry, and my sister-in-law, Jo, for their constant encouragement during my PhD journey. A special thank you to my parents: my mum, Jennifer, who instilled in me a love for science and investigation, and my dad, Richard, for our dinnertime debates, which prepared me well for the viva. You have both always been there, offering advice and support during challenging times. Your love and belief in me have brought me to where I am today. Thank you!

Abstract

Melanoma is the most aggressive form of skin cancer and fifth most common cancer in the UK. Although immunotherapy has improved outcomes for patients with advanced stages of the disease, some patients do not respond to these treatments or develop resistance during therapy. Additionally, the current staging system which is used for determining prognosis and guiding the treatment of melanoma patients, shows significant variability, with some early-stage patients progressing to metastatic disease. Therefore, identifying which patients will respond to treatments and discovering new prognostic biomarkers have become crucial research areas for improving patient outcomes. The development of digital slide scanners has meant that tissue slides, which contain a wealth of phenotypic information, are being digitised to generate whole slide images (WSIs). This thesis investigates how artificial neural networks with digital pathology workflows, can be used to stratify patients based on their outcomes.

Firstly, we classified patients' WSIs into subgroups based on the inferred expression of immune cells within their tumors. These immune subgroups, derived from genetic data, present varying potential treatment targets and survival outcomes. WSIs are multi-resolution, multi-gigabyte images containing billions of pixels but often will only have a slide-level label. To handle this, we employed multiple instance learning (MIL) methods, where the image is divided into numerous patches, and the image is classified based on a set of these patches using a single supervisory label. We are among the first to investigate how factors such as patch resolution, feature extraction methods, and MIL techniques influence the classification of melanoma patients into immune subgroups. In a primary melanoma dataset, we achieve a mean area under the receiver operating characteristic curve (AUC) of 0.80 for classifying histopathology images into high or low immune subgroups and a mean AUC of 0.82 in an independent TCGA melanoma dataset. Our findings indicate that using 10x resolution patches, pathology-specific feature extraction methods, and attention-based MIL models improve classification performance.

To build on these MIL methods, we introduce a novel way to represent WSIs, using multi-resolution patch graphs, with resolution-aware node embeddings. These graphs enable us to capture long-range dependencies within the images. By employing graph neural networks to aggregate surrounding node embeddings from the WSI patches, we improved the classification performance beyond the MIL models. Here, we achieved a mean test AUC of 0.81 for classifying low and high immune melanoma subtypes, using graph representations.

Finally, we use survival graph neural networks to identify new prognostic subgroups from patch graph WSI representations. Here we show that the risk groups and immune subgroups generated from the models presented within this thesis were predictive of melanoma-specific survival (Concordance index = 0.73), even when adjusting for known prognostic factors. This suggests that these risk groups could represent novel predictive and prognostic biomarkers. Overall this thesis adds to the evidence that digital pathology workflows are an emerging tool for better understanding melanoma patient outcomes and survival.

Abbreviations

AI Artificial Intelligence

AIC Akaike Information Criterion

AJCC American Joint Committee on Cancer

ANN Artificial Neural Network

AttMIL Attention-based Multiple Instance Learning

AUC Area Under the receiver operating characteristic Curve

BA Balanced Accuracy

BATF3 Basic leucine zipper ATF-like Transcription Factor 3

BYOL Bootstrap Your Own Latent

CART Classification And Regression Trees

CI Confidence Interval

C-Index Concordance Index

CLAM Clustering-constrained Attention Multiple instance learning

CNN Convolutional Neural Network

CPH Cox Proportional Hazard

CPTAC-CM Clinical Proteomic Tumor Analysis Consortium Cutaneous Melanoma

CTLA-4 Cytotoxic T-lymphocyte associated protein 4

DASL cDNA-mediated Annealing, Selection, extension, and Ligation

DSMIL Dual Stream Multiple Instance Learning

ELU Exponential Linear Unit

FC Fully Connected

FFPE Formalin-Fixed Paraffin-Embedded

GATv2 Graph Attention Network version 2

GNN Graph Neural Network

GPU Graphics Processing Unit

H&E Haematoxylin & Eosin

HIPT Hierarchical Image Pyramid Transformer

HR Hazard Ratio

HSV Hue Saturation Value

IDO1 Indoleamine 2,3-dioxygenase 1

LAG-3 Lymphocyte-activation gene 3

LMC Leeds Melanoma Cohort

MIL Multiple Instance Learning

MLP Multilayer Perceptron

mRNA Messenger ribonucleic acid

MSI Microsatellite Instability

MSS Melanoma Specific Survival

NICE National Institute for Health and Care Excellence

NT-Xent Normalized Temperature-scaled Cross Entropy Loss

PANDA Prostate cANcer graDe Assessment

PD-1 Programmed cell Death protein 1

PD-L1 Programmed Death-Ligand 1

RAdam Rectified Adam

ReLU Rectified Linear Unit

ResNet Residual Network

RGB Red Green Blue

RNA Ribonucleic Acid

RNA-seq RNA sequencing

ROI Region Of Interest

SimCLR Simple Contrastive Learning of Representations

SOTA State Of The Art

SSL Self-Supervised Learning

SurvGNN Survival based Graph Neural Network

SVM Support Vector Machines

Tanh Hyperbolic Tangent function

TCGA The Cancer Genome Atlas

TIGIT T cell Immunoreceptor with Ig and ITIM domains

TILs Tumour Infiltrating Lymphocytes

TIM-3 T-cell Immunoglobulin and Mucin-domain containing-3

TME Tumour Microenvironment

TransMIL Transformer-based Multiple Instance Learning

VGP Vertical Growth Phase

WSI Whole Slide Image

Contents

1 Introduction **1**

1.1 Background and Motivation . . . 1
1.2 Thesis aims and contributions . . . 3
1.2.1 Contributions . . . 3
1.3 Thesis structure . . . 4
1.4 Summary . . . 5

2 Clinical and technical background **7**

2.1 Melanoma background . . . 7
2.1.1 Histological subtypes and variation in melanoma presentation . . . 7
2.1.2 Histopathological factors associated with melanoma prognosis . . . 10
2.1.3 Genetic subgroups for stratifying patients . . . 12
2.2 Technical background . . . 13
2.2.1 Artificial Neural Networks . . . 13
2.2.2 Convolutional neural networks . . . 14
2.2.3 Deep learning for computational pathology . . . 15
2.2.4 Multiple instance learning . . . 16
2.2.5 Graph representations and graph-based methods . . . 19
2.2.6 Application of deep learning to survival prediction . . . 21
2.3 Review of deep learning methods applied to melanoma whole slide images . . . 22
2.3.1 Predicting genetic subtypes of melanoma . . . 22
2.3.2 Predicting immune subtype signatures of melanoma . . . 24
2.3.3 Current challenges and opportunities . . . 26
2.4 Conclusion . . . 29

3 Datasets and image pre-processing **31**

3.1 The Leeds Melanoma Cohort . . . 31
3.1.1 H&E slide and whole slide image generation . . . 32
3.2 The Cancer Genome Atlas . . . 33
3.3 Limitations of the datasets . . . 34

3.4 Immune subgroup ground truth labels . . . 35
3.5 Image processing . . . 37
3.5.1 Segmentation and patch generation . . . 37
3.5.2 Dataset splits . . . 38
3.6 Conclusion . . . 38

4 Immune subtyping of melanoma whole slide images using multiple instance learning 41
4.1 Introduction . . . 41
4.1.1 Contributions . . . 44
4.2 Methods . . . 44
4.2.1 Segmentation and feature extraction . . . 44
4.2.2 Model architectures and training . . . 44
4.2.3 Visual attention maps . . . 48
4.2.4 Survival analysis . . . 48
4.3 Datasets . . . 49
4.4 Results . . . 50
4.4.1 Predicting high, intermediate and low immune subtypes . . . 50
4.4.2 Comparison of MIL models . . . 50
4.4.3 Comparison of feature extraction methods . . . 52
4.4.4 Comparison of resolutions . . . 53
4.4.5 Immune attention heatmaps for explainability . . . 57
4.4.6 Prognostic associations of model predictions . . . 61
4.5 Discussion . . . 66
4.5.1 Overview of findings . . . 66
4.5.2 Challenges using transcriptomic ground truth labels . . . 67
4.5.3 MIL models attend to regions with prognostic value . . . 67
4.5.4 The importance of feature inputs and model selection . . . 68
4.5.5 Limitations and future directions . . . 69
4.6 Conclusion . . . 70

5 Multi-level Graph Representation of Melanoma Whole Slide Images for Identifying Immune Subgroups 73
5.1 Introduction . . . 73
5.1.1 Contributions . . . 75
5.2 Methods . . . 76
5.2.1 Segmentation and feature extraction . . . 76
5.2.2 Graph construction . . . 77
5.2.3 Model architecture . . . 77
5.2.4 Model training . . . 79

5.3 Results . . . 80
5.3.1 GNN models outperform MIL models . . . 80
5.3.2 Increasing GATv2 layers improves performance for models trained with larger graph representations . . . 83
5.3.3 Attention mechanism generates the best AUC performance for model trained with 40x single resolution graph representations . . . 84
5.3.4 Additional metrics and prognostic associations of models . . . 87
5.4 Discussion . . . 90
5.5 Conclusion . . . 92

6 Survival prediction for melanoma patients, using graph-based whole slide image representations 93
6.1 Introduction . . . 93
6.1.1 Related works . . . 94
6.1.2 Contributions . . . 96
6.2 Methods . . . 96
6.2.1 Patient-level representations and ground truth labels . . . 96
6.2.2 Model architecture . . . 98
6.2.3 Loss function . . . 99
6.2.4 Model implementation . . . 100
6.2.5 Generating de-novo survival-based subgroups . . . 101
6.2.6 Univariate and multivariate statistics . . . 102
6.2.7 Treatment target expression analysis . . . 105
6.3 Results . . . 105
6.3.1 Analysis of high and low risk groups . . . 105
6.3.2 Analysis of high, low and intermediate risk groups . . . 110
6.3.3 Determining the prognostic ability of the risk groups in the presence of established prognostic factors . . . 113
6.3.4 The distribution of potential treatment targets and clinical, histological and immune features in the predicted immune and risk groups . . . 117
6.4 Discussion . . . 121
6.4.1 Conclusion . . . 125

7 Conclusions and future work 127
7.1 Summary and thesis contributions . . . 127
7.1.1 The importance of resolution, feature extraction method attention for classifying immune subtypes . . . 128
7.1.2 Context improves immune subtype classifications of melanoma whole slide images . . . 128

7.1.3 Melanoma whole slide images contain prognostic information beyond the established prognostic factors . . . 128
7.2 Limitations and future directions . . . 129
7.2.1 Immediate limitations and future directions specific to this work . . . 129
7.2.2 Limitations and future directions for wider digital pathology research applied to melanoma . . . 131
7.3 Conclusion . . . 133

Appendices **171**

A Multi-level graph representation additional results **173**

B Immune, clinical and clinicohistological features of the immune and risk subgroups **175**

List of Figures

2.1 The three different gradings of tumour infiltrating lymphocytes (TILs) distributions: Brisk, non-brisk and absent lymphocytes. 9

2.2 A simplified diagram of how PD-1 and PD-L1 are involved in the interaction of melanoma tumour cells and T-cells, leading to T-cell deactivation. 9

2.3 The different levels of Breslow thickness used in the American Joint committee on Cancer (AJCC) staging of melanomas. 10

2.4 Multiple instance learning being used to solve a binary diagnostic problem. . . . 17

2.5 A GNN with two message passing layers, showing how information from the neighbouring nodes of target node A from an input graph, are incorporated into the graph-representation. 20

3.1 (A) The width and height of the 667 images with corresponding immune subgroup labels. (B) A histogram showing the distribution of the areas of the images in millions of pixels^2. 33

3.2 (A) The width and height of the 285 images with corresponding immune subgroup labels. (B) A histogram showing the distribution of the areas of the images in millions of pixels^2. 34

3.3 An example of an LMC image and patches at five different resolutions. 39

3.4 An example of a TCGA image and patches at five different resolutions. 40

4.1 Datasets and experimental framework. 43

4.2 Confusion matrices for the model predictions of the 'low', 'intermediate' and 'high' cases on the 10-fold test sets. 51

4.3 Performance comparison of the six different types of MIL models when classifying the melanoma immune subtypes in the two datasets. 52

4.4 Model performance when altering the feature extraction method and resolution of input patches for models trained with the LMC dataset. 54

4.5 Model performance when altering the feature extraction method and resolution of input patches for models trained with the TCGA dataset. 55

4.6 Mean 10-fold test AUC performance with 95% CI of MIL models classifying 'high' and 'low immune' subtypes with different epoch stopping lengths during training. 56

4.7 Confusion matrices for the model predictions of the 'low', and 'high' cases on the 10-fold test sets. 56

4.8 Comparison of correctly classified LMC 'high immune' tumours using different input features 58

4.9 Comparison of correctly classified LMC 'low immune' tumours using different input features. 59

4.10 Correctly classified 'high' and 'low immune' TCGA tumours. 60

4.11 Incorrectly classified 'high immune' LMC and TCGA tumours. 61

4.12 LMC 'low immune' image that was correctly classified by one model and misclassified by another model that used different input features. 62

4.13 Prognostic associations of the 'high' and 'low immune' subtypes from the LMC dataset ($N = 230$) 63

4.14 Prognostic associations of patients predicted as 'high' and 'low immune' subtypes from the LMC dataset ($N = 401$). 65

5.1 Experimental framework for classifying the WSIs, using graph representations. . 75

5.2 Single and multi-resolution patch graph representation construction and edge feature encodings used for multi-resolution graphs. Self-loops are not visualised. . 76

5.3 Types of multi-level graph structures, that were used in the ablation studies. . . 78

5.4 Network architecture for a three GATv2 layer graph neural network. 79

5.5 Confusion matrix for the GNN model prediction of the 'low', 'intermediate' and 'high immune' cases on the 10-fold test sets. 81

5.6 Correctly classified high immune images (A-C) and incorrectly classified low immune images (D-F) from the top three performing attention GNNs. 85

5.7 High attention regions from the multi-level attention GNN. 87

5.8 Survival distributions for the 'high' v.s. 'low immune' test set predictions from GNN models which utilise an attention mechanism. 88

6.1 Survival analysis pipeline. 97

6.2 A reminder of the graph representations of the WSIs, which are inputs for the SurvGNNs. 98

6.3 The clinical factors associated with MSS. (A) Log HR values with 95% CI for the clinical factors associated with MSS. The C-index and partial AIC values are also shown for the multivariate CPH model fitted with these covariates. (B-F) Kaplan- Meier estimators with 95% CI to show the survival distributions for different clinical factors, with two sided log-rank statistics. (D) Age at diagnosis is shown here as a categorical variable to demonstrate the adverse effect on survival, however, is treated as a continuous variable in our analyses. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, Hazard ratio; MSS, melanoma specific survival. 104
6.4 (A) Log HRs with 95% CI of clinicohistological features in relation to MSS.(B) Kaplan-Meier survival analysis with 95% CI of the risk groups from the 10x SurvGNN with two-sided log-rank test. 109
6.5 The top nine high attention patches from eight different melanoma WSI cases (n=8 patients) classified, as (A) low risk or (B) high risk of melanoma death by the 40x Attention-based SurvGNN model. 110
6.6 (A) Log HRs with 95% CI of clinicohistological features in relation to MSS. B) Kaplan Meier survival analysis with CI of the risk groups from the 20x SurvGNN with two-sided log-rank test. 115
6.7 Log HRs with 95% CI for significant clinicohistological features and SurvGNN and GNN attention risk groups in relation to MSS. 117
6.8 Relative distributions of potential treatments targets and immune biomarkers for patients in the high and low risk groups as predicted by the 10x SurvGNN model, using smaller hidden layers and batch size. 119
6.9 Relative distributions of potential treatments targets and immune biomarkers for patients in the high and low immune classes as predicted by the GNN attention 40x (3 GATv2) model. 120

A.1 Additional example of high attention regions generated from the best performing multi-level attention GNN. 173

B.1 Relative distributions of potential treatments targets and immune biomarkers for patients in the three immune subgroups derived by Pozniak *et al.* (2019) [1]. . . 176

List of Tables

3.1 Summary data of the clinical and clinicohistological features of the three LMC immune subgroups found in the paper published by Pozniak *et al.* (2019) [1] . . 36

3.2 The average number of patches extracted from the segmented tissue from LMC ($N = 667$) and TCGA ($N = 285$) WSIs at multiple resolutions. 38

3.3 LMC dataset splits, showing the number of WSIs labelled with the three different immune subgroups found by [1], within the training, validation and test sets, when carrying out Monte Carlo 10-fold cross validation 38

3.4 TCGA dataset splits, showing the number of digitised WSIs labelled with the three different immune subgroups found by [1], within the training, validation and test sets, when carrying out Monte Carlo 10-fold cross validation. 38

4.1 Mean 10-fold test performance metrics with 95% CI for the three models with the highest mean test AUCs for classification of 'high' and 'low immune' LMC cases. 57

4.2 The prognostic value of the 'high immune' and 'low immune' subtypes in univariable and multivariable analyses from the LMC dataset, when using test set labels (N = 230), derived from the highest performing MIL models. 64

4.3 The prognostic value of the 'high immune' and 'low immune' subtype classifications in univariable and multivariable analyses from the LMC dataset, when using test set labels (N = 401) for the whole datset (including 'intermediate immune' subtypes), derived from the highest performing MIL models. 66

5.1 One-hot encoding dictionary for the edge features in symmetric and asymmetric multi-level graphs. 76

5.2 Mean test AUC for classification of three immune subtypes with 10 fold cross validation with 95% CI (mean ± 95% CI), for different MIL models compared to our proposed GNN models using single and multi-resolution graph respresentations. 80

5.3 Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for different MIL models compared to our proposed GNN models. 82

5.4 Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for GNNs when increasing the number of GATv2 layers in the network. 83
5.5 Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for GNNs when using an attention mechanism instead of global mean pooling operation. 84
5.6 Mean test AUC, BA and F1-score for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI). 88
5.7 The prognostic value of the 'high immune' and 'low immune' subtype classifications in multivariable analyses for the LMC dataset 89

6.1 The hidden layer size and batch size for graph inputs of different sizes. 99
6.2 Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models trained with different graph WSI representation inputs using global mean pooling mechanism. 106
6.3 Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models with an attention mechanism trained with different graph WSI representation inputs. 107
6.4 Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models using graph mean pooling, trained with different graph WSI representation inputs and reduced batch and hidden layer sizes. 107
6.5 Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the attention SurvGNN models, trained with different graph WSI representation inputs and reduced batch and hidden layer sizes. . . . 108
6.6 Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the SurvGNN models with global mean pooling mechanisms, that were trained with different graph WSI representation inputs. 111
6.7 Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the attention SurvGNN models, that were trained with different graph WSI representation inputs. 112
6.8 The affect of reducing batch and hidden layer size on SurvGNN models with global mean pooling mechanism that generate three risk groups. 113
6.9 The affect of reducing batch and hidden layer size on attention-based SurvGNN models that generate three risk groups. 114
6.10 The prognostic value of the 'high immune' and 'low immune' subtype classifications in multivariable analyses for the LMC dataset compared to the SurvGNN model risk group classifications. 116

6.11 Likelihood-ratio test statistics with p-values to assess whether the novel biomarkers from the models contain additional prognostic information compared to the standard clinical variables for predicting MSS. 118
6.12 Summary data of the clinical and clinicohistological features of the high and low risk LMC patients from the 10x SurvGNN model. 121
6.13 Summary data of the clinical and clinicohistological features of the high immune and low immune LMC patients (including the intermediate subset of patients) classified by the 40x attention GNN model. 122

B.1 Summary data of the clinical and clinicohistological features of the high and low risk LMC patients from the attention-based 40x SurvGNN model. 177

Chapter 1

Introduction

1.1 Background and Motivation

Melanoma is the most aggressive form of skin cancer [2] and is the fifth most common cancer in the UK [3]. Moreover, incidence rates of melanoma have been predicted to increase by 9% in the UK between 2023-2025 and 2038-2040 [2]. Immunotherapy has revolutionised the treatment of melanoma patients [4, 5, 6]. Despite the approval of the first immune checkpoint inhibitor for advanced melanoma over a decade ago [7], there are still patients who do not derive long-lasting benefits or respond to immunotherapeutic treatment. Moreover, the current melanoma staging system shows wide levels of variance, with recurrence of locoregional disease or distant metastases occuring in approximately 6% of stage I and 20% of stage II melanoma patients, with a significant number of melanoma-specific deaths occurring in patients who were initially diagnosed with early stage disease [8]. Consequently, determining new early-stage prognostic biomarkers to improve the current staging system and stratifying patients for effective treatment, remain increasingly important challenges, which we aim to address in this thesis.

Several studies have stratified patients into subgroups based on gene expression signatures, which have shown prognostic significance independent to American Joint Committee on Cancer (AJCC) staging [9, 10, 11, 12, 13]. In 2019, Pozniak *et al.* (2019) [1], used tumour transcriptomic data to stratify melanoma patients into immune-related subsets, based on the inferred abundance of 27 immune cells as described by Angelova *et al.* (2015) [14]. They carried out consensus clustering to delineate three resulting subsets with distinct immune phenotypes: a 'high', 'intermediate' and 'low immune' subgroup. The patients in the 'high' subgroup, had tumours enriched in pathways involved in immune signalling and had a significantly lower hazard of melanoma death compared to the 'intermediate' and 'low immune' subsets. Whereas the 'low immune' subgroup expressed genes enriched in pathways for the cell-cycle, metabolism and immune suppression. Furthermore, although the patients were enrolled before immunotherapeutic regimen became available, the three identified immune subsets had distinct molecular

profiles which could be targeted by current or future immunotherapies [15, 16, 17]. They were also able to replicate these findings, in an independent transcriptomic dataset of melanoma tumours, from the cancer genome atlas (TCGA).

However, the identification of prognostic molecular biomarkers for melanoma patients, using transcriptomic data analysis, is a costly and time-consuming process. Moreover due to the global shortage of staff in pathology departments, particularly in the UK, where only 3% of departments are fully staffed, there is a significant demand for automated screening and triaging tools to facilitate the detection of predictive biomarkers [18]. The integration of deep learning and image analysis presents a cost-effective opportunity to automate the detection of prognostic biomarkers within clinical workflows [19]. This approach utilises digitised haematoxylin and eosin (H&E) stained slides called whole slide images (WSIs), which are multi-resolution, gigapixel images that contain a wealth of diagnostic information. Furthermore by exploring the image-based histological biomarkers underpinning the molecular signatures, these models can be used to find new biomarkers and help to find new drug targets, thus playing a key role in precision medicine. Numerous studies have successfully demonstrated the use of these AI or deep learning techniques with digitised H&E-stained slides to predict genomic subtypes and the level of expression of specific genes. For example, deep learning pipelines that utilise WSIs were able to classify and predict mutations for lung [20, 21] and breast cancer subtypes [22, 23, 24, 25, 26], predict tumor mutational burden in bladder cancer patients [27] and detect microsatellite instability in colorectal, gastric, endometrial and ovarian cancers [28, 29, 30, 31, 32]. In this thesis, one of our main objectives is to classify melanoma WSIs into the immune subtypes derived by Pozniak *et al.* (2019) [1], which have distinct molecular profiles that could be targeted by current or future immunotherapies and differing survival outcomes.

As well as exploring molecular profile prediction, models can be developed for prognosis prediction, using survival data directly. For example Skrede *et al.* (2020) and Lee *et al.* (2022) have shown that deep learning models trained with WSIs, can stratify patients into more fine-grained and distinct risk groups, than the established prognostic biomarkers currently used [33, 34]. Moreover, multiple works [35, 34, 36] have shown how using graph-based representations of WSIs and graph neural networks (GNNs), where long-range dependencies are modelled within an image, are able to incorporate context from the local tumour microenvironment improving prognosis prediction. Currently, only certain histopathological features are included in cancer staging systems (e.g. tumour thickness), because of inter-patient heterogeneity and difficulty in carrying out reproducible measurements of certain biomarkers (e.g. tumour infiltrating lymphocytes, TILs)[37]. In this thesis we aim to harness the morphological features in WSIs, to investigate whether deep learning based stratification may improve melanoma patient separation beyond current staging and clinical factors associated with melanoma specific survival.

1.2 Thesis aims and contributions

The thesis aims are as follows:

1. To develop a deep learning pipeline for classifying digitised images of melanoma tumours into the immune subgroups identified by Pozniak *et al.* (2019) [1]. This will involve assessing current multiple instance learning (MIL) methods, which are a popular approach in computational pathology.

2. To investigate whether context-aware graph-based approaches improve on MIL methods for classifying digital pathology images into these immune subgroups.

3. To test the hypothesis that deep learning models can be used to stratify patients into risk groups with greater prognostic value than the immune subgroups [1] and 7^{th} edition AJCC staging system [38].

1.2.1 Contributions

The main contributions of this thesis are highlighted in the sections below.

Classifying digitised images of melanoma tumours into immune subgroups

In Chapter 4, we present our results for classifying melanoma WSIs from two independent datasets into the immune subgroups [1] based on gene expression signatures. To our knowledge, this is one of the first attempts to evaluate how resolution, feature extraction method and MIL method impact model performance when classifying melanoma WSIs. We demonstrate that 10x resolution input patches improve performance, through a balance of cellular detail and context, achieving a mean test AUC of 0.79, balanced accuracy of 0.74 and F1 score of 0.67 for classifying 'high' vs. 'low' immune cases. In addition we show that a self-supervised learning 'feature extractor', trained with histopathology images, improves performance over a supervised learning feature extractor trained with natural images. We also outline some unique challenges when classifying WSIs based on ground-truth labels which are localised to a small region (a 0.6mm needle-biopsy).

Graph representations of melanoma whole slide images for immune subtyping

In order to process multi-gigabyte, multi-resolution WSIs, methods often involve dividing images into smaller patches, which are then individually processed with a neural network. By processing patches independently, context and long-range relationships within the image are lost, meaning while we can capture local cell morphology, we are less able to model interactions within the immune microenvironment, which may be important for classifying patients into the immune subgroups [1]. In Chapter 5, we develop novel multi-resolution graph structures of melanoma WSIs, using encodings for each resolution, and process these representations using graph neural networks (GNNs) that can model interactions between different patches within

an image. We introduce novel 'multi-level' WSI representations and show show that GNNs are superior compared to the MIL models we evaluated at classifying the immune subgroups [1], achieving a mean AUC of 0.80. We again demonstrate the importance of attention, as the highest mean AUC performance for the immune subgroups (0.81), was achieved by a GNN model with an attention mechanism.

Development of image-based risk groups for melanoma patients

The AJCC staging system [39] for melanoma displays wide levels of variance [40, 41] and does not predict progression to metastasis in a subset of thinner stage I-II tumours. While certain prognostic features within H&E stained images, are used in the AJCC staging system, in Chapter 6, we argue that there is still a wealth of prognostic information within these images that is underutilised. We develop survival GNNs to stratify patients into risk groups that are predictive of melanoma-specific survival, even when adjusting for known prognostic factors (C-index = 0.73). Moreover we show that the predicted immune subgroups from the weakly supervised GNN models and the risk groups from the survival GNN models, provide complementary added prognostic value for predicting survival.

1.3 Thesis structure

In this section we provide a summary of the thesis structure and a brief description of what each chapter contains.

- **Chapter 2** introduces the histopathological features of melanoma in more detail and the current state of the art artificial neural networks used to classify melanoma WSIs.
- **Chapter 3** describes the two datasets used in our research, highlighting their strengths and weaknesses. We describe the pre-processing techniques applied to the WSIs for downstream analysis.
- **Chapter 4** reports our results when applying six state of the art multiple instance learning (MIL) frameworks to two independent datasets for classifying patients into immune subgroups [1] A shortened version of this work "Weakly-supervised learning for image-based classification of primary melanomas into genomic immune subgroups" was published as a fully peer-reviewed paper in the Proceedings of The 5th International Conference on Medical Imaging with Deep Learning [42] and a further extension of this work "Immune subtyping of melanoma whole slide images using multiple instance learning" was published in Medical Image Analysis [43].
- **Chapter 5** explores how novel multiple resolution graph-based representations of melanoma whole slide images can be used to classify patients into the immune subgroups [1]. The work in this chapter is an extension of work presented in the paper "Multi-level Graph

Representations of Melanoma Whole Slide Images for Identifying Immune Subgroups" for the Graphs in Biomedical Image analysis workshop [44].

- **Chapter 6** presents a methodology for classifying graph representations of melanoma whole slide images into *de novo* risk groups based on their survival. We compared the prognostic value and independence of our *de novo* WSI-derived risk groups, compared to the 7th edition AJCC staging system for melanoma [38], the immune subgroups [1] and other clinical factors.
- **Chapter 7** synthesizes all the findings reported in Chapters 4-6 into an overall discussion, highlighting our contributions to the fields of digital pathology and melanoma research. We also note the limitations of the work presented in this thesis and potential future work relating to our own research and more broadly digital pathology research applied to melanoma patient data.

1.4 Summary

In this chapter we introduced the main challenges in the enduring endeavour of classifying melanoma patients to inform stratified treatment. We then explained the motivation behind using deep learning to address these challenges and outlined the thesis aims, contributions and structure. In the next chapter we provide a more in depth review of the histopathology of melanoma and technical background of this thesis and explore the current work in our research area in more detail.

Chapter 2

Clinical and technical background

This chapter provides a brief introduction of the pathogenesis of melanoma and an overview of the histopathological and immune features of tumours that have prognostic value. We also provide technical background and discuss the limitations of the deep learning methods currently being used to classify melanomas based on their molecular and histological features.

2.1 Melanoma background

Cutaneous melanoma occurs when pigment producing cells, called melanocytes, mutate and begin to divide uncontrollably in the skin's epidermis [2]. While patient outcomes are generally favourable if the disease is diagnosed early, 20% of all melanoma deaths are from patients who present with early-stage disease [45]. Therefore, improved prognostication and improved prediction of how patients will respond to therapies is of great importance for clinical management. However, melanoma is a highly heterogenous disease, with both clinical and histopathological presentations varying widely both inter and intra-tumourally [46]. This can make it difficult to stratify patients and discover prognostic markers which will help in understanding patient outcomes, treatment decision making and clinical trial design.

2.1.1 Histological subtypes and variation in melanoma presentation

Variation in tumours includes phenotypic differences such as cellular size, shape and pigmentation and diverse genetic alterations [47]. Nevertheless, there are broadly four main histological subtypes of melanoma: superficial spreading, nodular, lentigo maligna and acral letiginous melanoma [48, 49]. Superficial spreading melanoma is the most common subtype comprising of around 70% of cases, followed by nodular melanoma which equates to around 20% of cases [48, 49]. Superficial spreading melanoma is characterised by slow horizontal growth, whereas nodular melanoma is characterised by rapid vertical growth. Lentigo maligna melanomas account for around 5-10% of all cutaneous melanomas and are often diagnosed in older patients

and who present with a brown macule (a flat, discoloured area of skin) on their head or neck [50, 51]. The fourth subtype, acral letiginous melanoma, presents as dark areas of pigmentation and occurs on the palms, soles of feet, or under toenails or fingernails [52]. It is one of the most rare subtypes of cutaneous melanomas in white populations, however, it is the most common form of melanoma diagnosed in non-white individuals [52]. As well as differences in clincial presentation, there is genetic diversity between these subtypes, with only 16% of acral letiginous melanoma presenting with BRAF mutations, compared to 66% of superficial spreading melanomas having this oncogenic (gene with the potential to cause cancer) mutation [53]. Although some studies [50, 52, 54] have suggested that histological subtypes are associated with patient outcomes or responses to treatment, unlike solid cancers such as breast or lung cancer, these subtypes are not currently incorporated into the prognostic staging system for primary melanomas and are often under-reported in pathology reports [55]. For instance, the TCGA, a public database for genomic and imaging data, includes histological grading for only a small fraction of melanoma cases (31 out of 478) [54]. The exclusion of histological subtypes from the melanoma staging system is primarily due to their confounding relationships with other histopathological factors that are included in the current staging system, such as ulceration and tumour depth (Breslow thickness), which are more frequently observed in the nodular subtype [54].

In addition to the differing histological subtypes, cutaneous melanomas can contain different cellular variants, which contribute to a wide range of histological appearances [56]. For example, there are rare balloon cell variants, which consist of foamy melanocytes lacking pigmentation and characterise 0.15% of malignant melanomas [57], which are extremely different in presentation to spindle cell melanoma variants, where cells are markedly elongated with long narrow nuclei and arranged in sheets [58]. The diagnosis of melanoma relies on a histopathologist's interpretation of the cellular morphology. The considerable heterogeneity and diverse cellular variants present within a tumor can, therefore, contribute to significant levels of interobserver variation among pathologists when diagnosing melanoma [59].

Furthermore, primary melanomas show great diversity in the immune cell infiltrate within the tumour microenvironment, with cells such as eosinophils, plasma cells, neutrophils, T-lymphocytes and phagocytic cells all being present [60]. The organisation and volume of tumour infiltrating lymphocytes (TILs) immune cells can be used as a prognostic marker. TILs are white blood cells, which leave the blood stream to surround and disrupt the vertical growth phase (VGP) of the tumour. TILs can include many different immune cells such as T cells, B cells and natural killer cells in different proportions. The different configurations of TILs can be graded into three different classifications: brisk TILs, non-brisk TILs and absent TILs [61]. Where brisk TILs are characterised by permeating the entire tumour or being diffuse within the entire base of the tumour (Fig. 2.1b). Non-brisk TILs are characterised as being 'multi-focal' (Fig. 2.1c), 'isolated' (Fig. 2.1d), or 'segmented', so restricted to certain areas within the tumour. Absent TILs indicate a tumour with no lymphocytes apposed to melanoma cells, or tumours

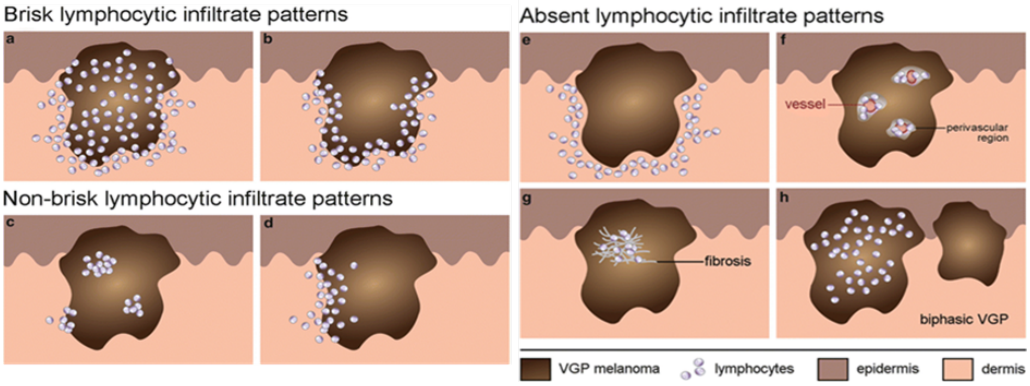


Figure 2.1: Figure adapted from Schatton *et al.* (2014) [61]. The three different gradings of tumour infiltrating lymphocytes (TILs) distributions: Brisk, non-brisk and absent lymphocytes. VGP stands for vertical growth phase. (a-b) Show diffuse brisk lymphocytes. (c) shows multifocal lymphocytes and (d) shows isolated lymphocytes. (e-h) Shows absent TILs.

where the lymphocytes are restricted to certain areas such as perivascular regions or areas of fibrosis (Fig. 2.1c and Fig. 2.1f). The grading of TILs is determined by the nodule with the smallest volume of TILs, if there are multiple nodules (Fig. 2.1h).

TIL grading can be done by visual observation of histopathology slides of a patient's tumour, however, a histopathologist would be unable to determine what type of cells the TILs were comprised of without using further tests. Some studies have suggested patients with brisk TILs are likely to have a better prognosis [61, 62]. Moreover, anti PD-L1 (programmed death-ligand 1) and PD-1 (programmed cell death protein 1) immunotherapies which are used to treat advanced melanoma are thought to act by modulating the activity of lymphocytes, by preventing T-cell activity being suppressed by melanoma cells [63]. Evidence also suggests that the presence of PD-L1 positive TILs, leads to improved patient response to immunotherapeutic

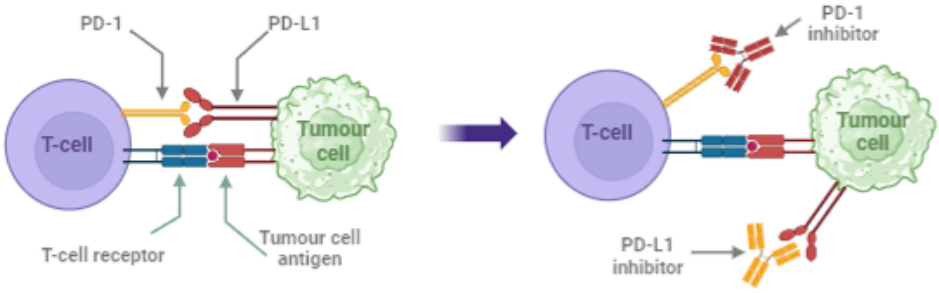


Figure 2.2: A simplified diagram of how PD-1 and PD-L1 are involved in the interaction of melanoma tumour cells and T-cells, leading to T-cell deactivation. PD-1 inhibitor and anti-PD-L1 inhibitor antibodies can then be used to block T-cell tumour cell interaction, preventing T-cell inhibition. Figure created with BioRender.com.

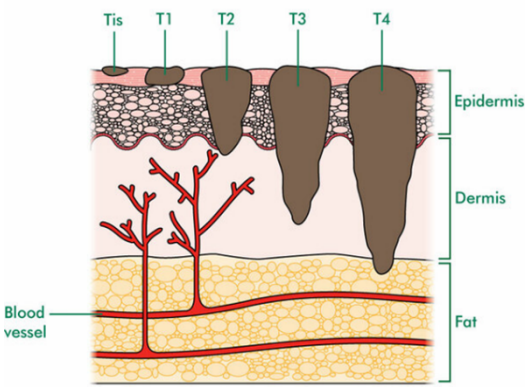


Figure 2.3: Figure from Macmillan Cancer Support [last accessed: Jan. 19, 2024] [66]. The different levels of Breslow thickness used in the American Joint committee on Cancer (AJCC) staging of melanomas.

treatments [64]. Nevertheless, while TIL grading has been established for many years, it has not been widely introduced into clinical decision-making due to conflicting evidence on the prognostic value of TILs, which could be due to tumour heterogeneity within and between patient populations and concerns for reproducibility between pathologists and institutions [62].

PD-1 is an immune checkpoint protein which is found on activated cytotoxic T-cells, a type of lymphocyte. PD-L1 is a ligand, which can be overexpressed by melanoma cells. If PD-1 receptors on the cytotoxic T-cells bind to the PD-1 on melanoma cells, the immune response of the T-cells is inhibited. Anti-PD-1 and anti-PD-L1 immunotherapies exert their efficacy by blocking PD-L1 from binding to PD-1 receptors, promoting cytotoxic T-cell activity so they can eliminate melanoma cells [65] (Fig. 2.2). Anti-PD-1 and anti-PD-L1 antibody therapies can be used in combination with other treatments that block other immune proteins, such as CTLA-4 (cytotoxic T lymphocyte-associated protein-4) and chemotherapies. Being able to determine, which patients will benefit from and respond to different treatments, or combination of treatments is an important area of research.

2.1.2 Histopathological factors associated with melanoma prognosis

As well as melanoma subtype and the presence of immune infiltrate, multiple other histopathological features such as Breslow thickness, microscopic ulceration, mitotic rate (a measure of how fast cancer cells are dividing and growing) [39], regression, lymphovascular invasion [67], neurotropism (perineural invasion and intraneural invasion [68]) and microsatellites (nests of metastatic cells adjacent to the primary tumour) have been associated with survival [39]. However, only Breslow thickness, microsatellites and microscopic ulceration are currently being used in the most recent 8th American Joint Committee on Cancer (AJCC) staging system [39]. This is an established risk stratification system for devising prognostic stage groups for patients with clinically localised primary melanoma. Breslow thickness describes how deep the tumour has

grown into of the skin (epidermis) and is measured in millimetres (mm). Melanomas can be categorised based on five levels of tumour thickness (this is known as T-staging): Tis where the melanoma cells are only at the top layer of the epidermis; T1 where the tumour thickness is $\leq$1.00mm; T2 where the tumour thickness is 1.01-2.00mm; T3 where the tumour thickness is 2.01-4.00mm and T4 where the tumour thickness is $>$ 4.0mm (Fig. 2.3) [66]. Depth has been shown to have a continuous association with patient prognosis [69, 70], however these categories are used in the staging of patients. Although Breslow thickness is considered to be the most reproducible measurement used for prognosis estimation [71], [72], studies have shown that histopathologist reported assessments can still suffer to interobserver variability [73, 74, 75]. With one study showing up to 0.86 mm difference in measurement between three different pathologists from the same institution [75].

Along with Breslow thickness, microscopic ulceration is also integral to the AJCC staging system and is an important prognostic marker [39]. Tumours with ulceration is the loss of full-thickness epithelium above the growing tumour and have reactive changes within the skin (with neutrophils and fibrin protein present) [76]. Studies have found that tumours with microscopic ulceration have a greater propensity to metastasise, [69, 77] and often coincide with higher mitotic rates [78, 79], suggesting a more aggressive tumour. Furthermore, the presence of ulceration be indicative of patient outcome after being administered with certain treatments such as adjuvant interferon therapies [80, 81]. Microscopic ulcerations are identified by pathologists viewing histopathology sections of a tumour; however, it can often be difficult to determine whether the markers have been induced by the tumour or are trauma-induced artifact, leading to interobserver variability [76].

Mitotic rate, has been shown to have adverse impact on melanoma patient survival [37, 39] and is also a histopathological feature that is used in the staging on melanoma patients. It is measured by a pathologist who evaluates the number of cells undergoing mitosis (mitotic figures) per mm^2 in the tumour tissue and reflects the proliferative behaviour of the tumour [82]. Mitotic rate was removed as staging parameter for characterising the 'T category' of the primary tumour, in part due to concerns for over-treatment of patients, as the presence of one mitotic figure was leading to patients being upstaged in the 7^{th} edition of the AJCC staging [37]. This demonstrates an important challenge with using cut-off points, such as one mitotic figure per mm, as although it can be useful for standardising prognostic parameter measurements, it can also lead to the loss of prognostic information across the dynamic range [37, 83].

While AJCC staging shows prognostic ability, it is limited by the prognostic parameters we are currently able to measure incorporate into it, with many additional prognostic features not utilised that could lead to more personalised and accurate prognostic estimates [37]. Moreover, challenges still persist regarding how to measure certain biomarkers, in a way that can avoid inter-observer variability and allow for reproducible biomarker cut points [37]. In addition, the reliance of the current AJCC staging system on only anatomical features, means that it does not capture the genetic variation and molecular heterogeneity that is displayed in the tumour

microenvironment. The current AJCC staging, while reflecting patient prognosis, does not indicate what treatment options may benefit patients, which is now a pertinent area of research due to advancements in immunotherapy and other targeted therapies. Therefore, being able to stratify patients into subgroups that reflect the genetic diversity and potential treatment options, while reducing potential subjectivity in decision making is an important step towards improving clinical decision making for patients.

2.1.3 Genetic subgroups for stratifying patients

Several studies have stratified patients into subgroups based on gene expression signatures, which have shown prognostic significance independent to AJCC staging [9, 10, 11, 12, 13]. Moreover, recent studies have also looked at the genetic make-up of tumours to gain a better understanding of the melanoma immune microenvironment by inferring the immune cell infiltrate through gene expression signatures of melanoma tumours [1, 84]. This is especially important, as studies have suggested that the immune composition of the tumour microenvironment might indicate which patients will respond to immunotherapy [85, 86], with a study by Larkin *et al.* (2015) [87] suggesting that checkpoint blockade immunotherapy is more likely to have efficacy in patients with PD-L1 positive tumours with T-cell infiltration. Furthermore, immune activation has been shown to have a prognostic role in metastatic disease [60], it is therefore an important area of research to understand the tumour-immune environment and find and measure features, which have prognostic value.

In 2013, Bindea *et al.* [88] published the first *immunome compendium*, profiling 28 different immune cell types throughout different stages of colorectal cancer. Following this, Nsenginmana *et al.* (2018) [84], in the Leeds Melanoma research group, adapted these methods to develop the *Melanoma Immunome*, by filtering out genes that they found were highly expressed in two melanoma cell lines from their lab. Based on the bulk expression of these 380 genes specific to 24 immune cell types found in melanoma tumours, Nsenginmana *et al.* (2018) [84] performed consensus clustering analysis on the primary tumours of 703 melanoma patients, finding six genetically distinct immune subgroups of patients. With the largest subgroup having the least inferred immune cell infiltration in the tumour and significantly worsened melanoma-specific survival. Whilst histopathological features, such as Breslow thickness, mitotic rate and ulceration, were highly associated with patient outcome, the subgroups found from clustering the primary tumours also showed independent prognostic value. In 2019, Pozniak *et al.* (2019) [1], in the same Leeds research group, then further refined the subgroups found by Nsengimana *et al.* (2018) [84], using an updated and more comprehensive *immunome compendium* established by Angelova *et al.* (2015) [14], applied more stringent filters for an updated melanoma immunome compendium that included 376 genes representing 27 immune cell subtypes. Immune cell scores were then generated using the mean expression values of all genes attributed to the different 27 cell types for the 703 primary melanoma transcriptomes, to infer the immune cell composition. These immune cell scores were then used to cluster the patient tumours and three prognos-

tic subgroups were found, with distinct low, intermediate, and high immune cell signatures. Patients in the high immune subgroup had a significantly lower hazard of melanoma related death compared to patients in the low and intermediate subgroups. Moreover, tumours within the high immune subgroup more frequently presented histopathological features that were associated with better outcomes compared to tumours from the intermediate and low immune subgroups, such as thinner tumours, significantly lower mitotic number, and greater numbers of reported TILs. Pozniak *et al.* (2019) [1] were also able to relate the gene expression of the different subgroups to different biological pathways, with pathways in the low immune subgroup being enriched in cell-cycle and metabolic pathways related to high proliferation, whereas the genes upregulated in the high immune subgroup related to immune pathways. The studies carried out by Nsengimana *et al.* (2018) [84] and Pozniak *et al.* (2019) [1] showed how gene expression signature analysis could be used for finding prognostic immune-related signatures and further understanding the biological mechanisms underpinning patient immunity, which could be used for improved treatment stratification. However, acquiring transcriptome data for this kind of analysis can be a time-consuming and costly process, moreover, these complex analysis methods are not typically in use in clinical laboratories for melanoma patients. Yet, the application of artificial intelligence and machine learning to pathology images, presents a cost-effective opportunity to automate the detection of prognostic biomarkers within clinical workflows [19], this is the approach I have developed in this thesis.

2.2 Technical background

In this section we introduce the methods that will be implemented in the following chapters (4, 5, 6).

2.2.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a machine learning method, developed to mathematically mimic the actions of neurones within a biological system [89, 90, 91, 92]. Bishop (2006) [93] describes a basic ANN as 'a series of functional transformations', which allow us to model nonlinear relationships. With the increase in availability of large datasets and improved computing hardware performance, neural networks have become one of the most popular methods for large-scale prediction and classification problems, where other methods may be unable to deal with high data dimensionality. The simplest ANNs, also known as feed forward neural networks or multi-layer perceptrons (MLPs) consist of an input, which is fed through layers of 'neurons', to give an output [93]. Within each layer there are nodes, each node is an input quantity from the previous layer and is connected to every node in the next layer by a weight, hence the layers are called 'fully-connected layers'. The sum of the inputs multiplied weights added to the bias of the layer are subsequently transformed using a differentiable non-linear activation function. The bias allows the activation function to be shifted by a constant value.

Linear combinations of the input variable, with the weights and biases of the first layer:

$$a_j = \sum_{i=1}^{D} w_{ji} x_i + w_{j0}$$

Where a_j is the activation of the j^{th} neuron in the first layer of the neural network, and $j = 1, ..., M$ with M neurons in the layer. w_{ji} corresponds to the weight associated between the j^{th} and i^{th} neuron. x_i corresponds to the network i^{th} input variable, where $x = 1, ..., D$ and D corresponds to the number on input nodes or features in the input layer of the network. w_{j0} corresponds to the bias term associated with the j^{th} neuron.

The quantities are then fed into a differentiable non-linear function $h(\cdot)$:

$$z_j = h(a_j)$$

for example, the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) [94]. Non-linear functions help to improve the expressiveness of the network, so it can approximate more complex data, with each layer capturing different levels of abstraction [94]. The z_j quantities can form inputs for the next layer. These are called the hidden units or layers of the neural network.

The final outputs of the network are then transformed using an appropriate activation function. The choice of this function depends on the distribution of the target variables. For example, if the problem is a multi-class problem e.g. developing a model to predict a type of flower, the softmax activation [94] function could be used to give a vector of probabilities for each flower type.

Furthermore, by having differentiable non-linear hidden units, error backpropagation can be implemented. Error backpropagation or backprop, first suggested by Werbos (1994), is a way of optimising the weights within a network, by minimising an error function, to train the model to generate the correct output values for the corresponding input values [95].

2.2.2 Convolutional neural networks

Convolutional neural networks (CNNs) are an adaptation of neural networks, which were inspired by Hubel and Wiesel's experiments on the visual cortex of cats and the idea of 'receptive fields' [96]. These experiments showed that neurons work in a hierarchical way, with simple cells sensing local visual features, such as the edges of shapes, which more advanced cells then process into more complex high-level features. The networks are named after the convolving operation, which occurs when two functions are multiplied together and integrated. In a CNN this happens when a 'filter', which is a small matrix of weights (e.g. a 3x3 or 5x5 matrix), is slid over an image input to perform element-wise multiplication between the filter and image pixel values. These values are then summed together to produce a single value and the filter continues to slide over the image, by a certain number of pixels (the stride) [93]. The result of

this colvolution over the input image is a feature map, which represents the learned features from the network. CNNs often consist of alternate convolutional and sub-sampling or pooling layers, before a few fully connected layers and activation function, like a MLP. CNNs improve upon the limitations of the traditional MLP neural network through a combination of local receptive fields, weight sharing and subsampling. In this way a CNN can retain the spatial relationships of the pixel inputs of an image, unlike an MLP and also have the property of shift invariance.

CNNs have revolutionised the field of computer vision, with Krizhevsky *et al.* in 2012 [97], outperforming the other machine learning methods with a CNN, AlexNet, that was able to classify the 1.2 million images into their respective 1000 classes in the ImageNet dataset challenge [98]. Since the AlexNet victory, CNNs have been deployed in tasks such as image classification, semantic segmentation and image and video recognition with diverse applications in facial recognition [99], autonomous vehicles [100] and medical imaging [101].

2.2.3 Deep learning for computational pathology

Histopathological images are generated when a formalin-fixed, paraffin wax-embedded (FFPE) or frozen tumour block is sliced to give around a 5µm thick section of tissue, which can be mounted on a slide. The slides then undergo staining, the principal tissue stain for diagnostic pathology is a combination of two stains hematoxylin and eosin, known as H&E. Hematoxylin imparts a purplish-blue hue to the nuclei, while eosin gives a pink coloration to the cytoplasm and extracellular matrix it stains [102] (refer to Figure 2.4). The slides can then be viewed using a light microscope, which allows a highly trained pathologist to assess the rich phenotypic information in the slides [103]. Changes in colour, shape and organisation of the stained cells can then be used as an indicator for the diagnosis and prognosis of disease [102]. H&E is also a relatively cost-effective method for most laboratories, as it is simple to perform and does not require specialised equipment or reagents [104]. In the 1990s, the advent of whole slide scanners led to the digitisation of slides, providing whole slide images (WSIs), which could be analysed computationally in a quantitative manner at different resolutions [101]. This can be challenging, as at their highest resolution, each image can be around 100,000 pixels by 100,000 pixels, demanding large amounts of processing power and storage. However, the digitisation of large datasets of images has also offered the opportunity for streamlining many clinical processes through the implementation of algorithms and deep learning pipelines [101].

In 2012, at around the same time that Krizhevsky *et al.* won the ImageNet challenge with AlexNet [97], Cireşan *et al.* showed that CNNs were able to outperform other machine learning methods at detecting mitotic cells in digitised histopathology images [105]. Subsequently CNNs have become one of the most popular methods for analysing WSIs, through their ability to learn features - implicitly quantifying cellular content - without needing annotations from a pathologist.

2.2.4 Multiple instance learning

One of the biggest challenges when analysing WSIs, is finding a computationally feasible way to analyse images comprising of billions of pixels. Moreover, this challenge is further complicated by the fact that the diagnostically important regions of interest (ROIs) can take up only a small fraction of the image. Furthermore, pathologists do not routinely perform pixel-level annotations on these ROIs, so many datasets only have a slide-level label for the WSI, meaning it is often not possible to carry out fully supervised methods where the ROI is labelled, or patch based methods where the goal is to classify the labels of individual patches within the image. One weakly supervised methodology that has been implemented in an attempt to resolve these challenges, is multiple instance learning (MIL) [106, 107]. MIL treats the image as a 'bag' made up of 'instances' which are patches that an image can be divided into:

$$X_i = \{x_1, \ldots, x_k\} \subset \mathbb{R}^m$$

Where X_i is the bag and $x_1, \ldots, x_k$ are the k instances, which are m-dimensional feature vectors with no explicit labels (Fig. 2.4). This means there are two main tasks that can be solved when applying MIL: *global detection* – determining the correct Y_i label for N bags (X_i), where $\{(X_i, Y_i) \mid i = 1, \ldots, N\}$, and *local detection* – classifying individual instances x_i, with the correct hidden patch-level labels y_i, where $\{(x_i, y_i) \mid i = 1, \ldots, k\}$. This allows us to solve a diagnostic task of predicting the slide-level label and can also be used to flag patches within the WSI that have triggered the slide-level label prediction. In this way, we can process a WSI in a more computationally efficient way by using smaller patches, while only requiring the slide-level label. In more recent applications of MIL, certain assumptions are relaxed, where traditionally MIL has been used to solve binary classification tasks [106, 107], more recent application have adapted MIL for multi-class, regression, or more complex tasks [20, 104, 108, 109, 110]. In addition, traditional MIL approaches often had a strict label assumption that there is at least one positive instance in a positive bag, however, more recent approaches have relaxed this assumption, so unlike the bags shown in Fig. 2.4 positive and negative bags can have overlapping distributions of instance-level features. These newer methods do not adhere strictly to identifying key instances, but instead the patterns of instances, integrating information across these instances, without assuming that every single patch-instance reflects the target of interest.

In practice, MIL is often applied in three steps. The first step involves using a feature extractor, for example a CNN, to reduce the patches from the WSI into compressed numerical vectors (called embeddings or features) e.g. from a 256×256 pixel patch to a 1024-dimensional feature vector [109]. Although the optimal way to carry out this feature extraction might be in an end-to-end manner, to learn the most salient features for a specific classification task, the extensive size of WSIs and consequently the patches extracted from these images, poses challenges for storing the entire network within the memory of a single graphics processing unit

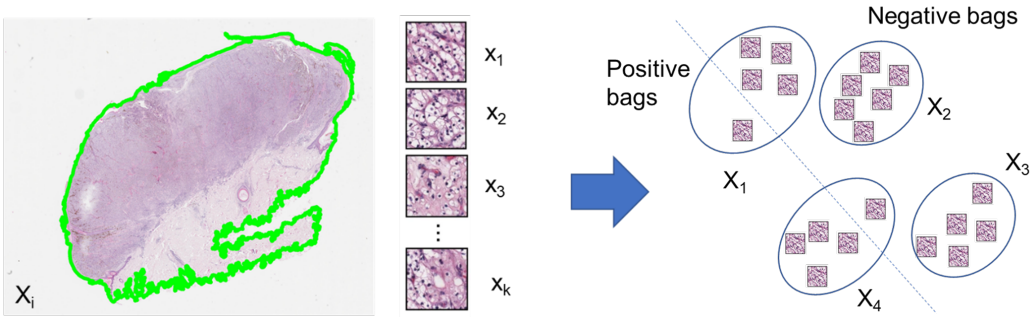


Figure 2.4: MIL used to solve a binary diagnostic problem. X_i represents a WSI, which is divided into $x_1, \ldots, x_k$ patches. X_1 and X_4 are positive bags or images containing at least one positive instance, whereas X_2 and X_3 are negative bags. MIL, multiple instance learning; WSI, whole slide image.

(GPU). Accordingly, the standard method to generate patch-level features has been to use CNNs pretrained with the benchmark dataset ImageNet [98]. This method of feature extraction has been shown to be effective in a wide variety of cancer subtyping tasks [108, 109, 111, 112, 113], yet, this pathology-agnostic feature extraction method has also also been shown to perform poorly in other classification tasks where diagnosis is based on subtle cellular patterns [21]. This arises because pathology images possess visual features that are distinct from those of natural images. For instance, they exhibit low color diversity due to the predominant presence of pinks and purples resulting from H&E stains. Moreover, they lack canonical orientation (a "right way up"), compared to an image of a cat for example. Additionally, interpretation can vary depending on the field of view; for instance, when examining tissue at higher resolutions, attention shifts toward individual cells rather than the tissue microenvironment [114].

In recent years, self-supervised learning (SSL) pretraining methods which ultilise unlabelled patches from pathology images have been leveraged to develop pathology-specific feature extraction methods. Broadly there are four major paradigms for SSL[114]: contrastive learning [115, 116, 117], non-contrastive learning [118, 119, 120], clustering [121, 122] and SSL with vision transformers [123, 124]. In Chapter 4, we compare the performance of two ImageNet pretrained CNNs against a CNN pretrained using SSL with histopathology images [125] using the constrastive SSL technique Simple Contrastive Learning of Representations (SimCLR) [126]. The objective of this method is to train a model to discriminate between feature embeddings, by maximising agreement between similar features that come from the same image but have undergone transformations and minimising agreement with features from different images. To do this SimCLR involves performing data augmentation, applying different transformations, on the input data. Ciga *et al.* (2022) employed randomly resized crops, 90° rotations, horizontal and vertical flips, photometric manipulation (colour jittering) and Gaussian blurring as augmentations [125] on patches from multiple pathology datasets. A base encoder or feature extraction model is then used to extract vector representations from the augmented images, for

this Ciga *et al.* (2022) used a ResNet [127], as did Chen *et al.* (2020) in the original SimCLR paper [126]. A projection head is then used to map the vector representations to smaller feature embeddings, for this Ciga *et al.* (2022) used a single hidden layer MLP to project the representation vectors into a lower embedding space [125]. Finally a contrastive loss function is applied, to maximise agreement between the representations from augmented views of the same image and minimise agreement between representations from different data instances [126]. Here Ciga *et al.* (2022) [125] applied Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) as a contrastive loss function, which utilises cosine similarity to evaluate the similarity between the vector representations. Through this method Ciga *et al.* (2022) were able to demonstrate that using features from a SLL ResNet18 improves performance on downstream pathology tasks compared to ImageNet pretrained ResNets [125].

The second step in a MIL pipeline is using an aggregation function to pool the features in the WSI to give a slide level feature. These function can be non-parameterised, for example Max-MIL pooling, where the model selects the patch with the highest classification probability and uses this to determine the overall slide-level label. Moreover, in recent years, it has become more popular to use parameterised aggregation methods, for example attention-MIL [128], where a neural network learns attention scores for each patch feature in the image, learning to assign higher attention scores or weighting to patches which should contribute more to the correct slide-level prediction. These methods leverage saturating activation functions such as the softmax function, which converts raw attention scores into a normalised probability distribution, with an output range of [0,1] and all output values summing to 1 [128]. The saturating behaviour of these functions ensures numerical stability and prevents vanishing or exploding gradients, which could lead to excessively large or small values for attention scores. The attention scores or probabilities can then be used to weight the instances, with instances with higher corresponding attention scores contributing more to the aggregated bag representation. The scores can then be visualised for model explainability, as we can determine 'high attention' patches or regions within the WSI and see if these correspond to ROIs a pathologist would find diagnostically relevant. In the field of computational pathology, MIL aggregation methods are continuously being developed and published, making it difficult to determine which aggregator is the best fit for a specific task. Consequently, in Chapter 4, we apply six different MIL methods to the task of subtyping melanomas to determine, which are the most suitable for this task.

The final step is then to map the aggregated WSI-level representation to a predicted class. While MIL has become established as one of the highest performing methods for WSI subtyping and future improvements in hardware may resolve end-to-end training memory constraints [129], one major issue with this method is that the patches are assumed to be independent to each other. This means that classifications do not take into account the spatial aspects of the tissue, beyond a patch-level. For example, as described previously in this chapter, the configuration of TILs have been found to be associated with patient prognosis [130], so when a WSI is divided into patches which are independently processed we lose these spatial histological features.

2.2.5 Graph representations and graph-based methods

One way to avoid contextual information loss when using patches from a WSI, is to represent a WSI as a graph, explicitly encoding the spatial relationships between patch entities within the WSI [131]. A graph is defined by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{G}$ is the graph, $|\mathcal{V}| = n$ corresponds to a set of n vertices or nodes and $\mathcal{E}$ is the set of edges or connections between each of the vertices. Edges usually connect two nodes, with the exception of edges in hypergraphs where an edge can connect more than two nodes [131]. Graphs provide a way of capturing spatial relationships in WSIs, as local information can be represented through edges between neighbouring patches or cells, while global patterns can be inferred from the overall graph structure [132]. Moreover, they provide an intuitive way of incorporating multi-scale information, allowing modelling of both high resolution detail (cellular information) and low resolution details (e.g. tissue architecture) [133]. Several studies have shown WSI graph representations can be used to model tissue regions [132, 134, 135, 136], cells [137, 138], or patches [36, 133, 139] within an image, with edges representing the interactions between these nodes. Graph-based deep learning methods can then be applied to these WSI representations to classify the images at a node or graph-level.

Spatial-based and spectral-based graph neural networks (GNNs) [131, 140] have become popular methods for processing graph representations, with software packages like PyTorch Geometric [141], allowing for the quick implementation of GNNs. GNNs differ from CNNs, because graphs are irregular structures that have no fixed notion of locality unlike image inputs, so while CNNs can utilise a sliding window or filter to capture local patterns within the grid-like structure of an image, GNNs use *message passing* mechanisms to learn abstract feature representations at each node. In order to capture spatial information, message passing layers operate using two main operations: an AGGREGATE operation and a COMBINE operation. For the AGGREGATE operation [131], information from the neighbouring nodes of a target node are collected:

$$m_v^{(l)} = \text{AGGREGATE}\left(\left\{\mathbf{h}_u^{(l-1)} : u \in \mathcal{N}(v)\right\}\right)$$

Where $m_v^{(l)}$, represents the aggregated neighbouring nodes for target node v at layer l, AGGREGATE represents an aggregation operation used to aggregate the information from the neighbouring nodes, $u \in \mathcal{N}(v)$ indicates that u is a node from the set of neighbouring nodes $\mathcal{N}(v)$ of target node v, and $\mathbf{h}_u^{(l-1)} : u$ represents the feature vector associated with the neighbouring node u at layer $l-1$. Common *AGGREGATE* operations include mean aggregation, max aggregation, or attention-based aggregation, for example, summation of the neighbouring feature representations of a target node. The weights from these operations are learned to determine the contribution of each neighbour's representation.

In the COMBINE operation, the aggregated neighbouring node representations are then combined with the target node representation:

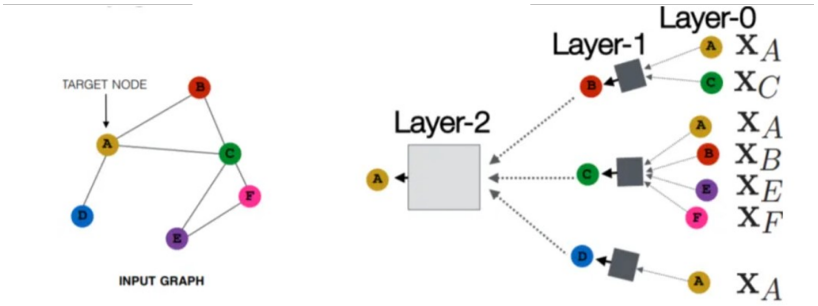


Figure 2.5: Figure from Representation Learning on Networks [last accessed: Feb. 19, 2024] [142]. A GNN with two message passing layers, showing how information from the neighbouring nodes of target node A from an input graph, are incorporated into the graph-representation. GNN, graph neural network.

$$\mathbf{h}_v^{(l)} = \text{COMBINE}\left(\mathbf{h}_v^{(l-1)}, \mathbf{m}_v^{(l)}\right)$$

Here $\mathbf{h}_v^{(l)}$ denotes the feature representation of the updated target node v, COMBINE represents the function used to combine the target node representation ($\mathbf{h}_v^{(l-1)}$) with the aggregated surrounding node representation ($\mathbf{m}_v^{(l)}$), with $\mathbf{h}_v^{(l-1)}$ representing the feature representation of target node v at layer $l-1$ (before the node update). For the *COMBINE* function, common operations include concatenation, weighted sum or element-wise addition to combine the target node representation with the representation of the aggregated neighbouring nodes. The weights in these operations are learned by the GNN to determine how the combined representations are aggregated to update the embedding of the target node.

The message passing layers are shown in more detail in Fig. 2.5, where it shows the updating of target node A, from the input graph shown. Layer-0 shows the input feature nodes, these could be features extracted from WSI e.g. patch or cell features. These nodes are the neighbours of the direct node neighbours (B, C and D) of the target node A. These features are passed through the aggregate and combine steps (layer-1), aggregating with the neighbouring nodes of the target node A. The direct neighbour representations (which have already been combined with their neighbours) are then aggregated and combined with the target node A. The final representation of node A, now has contextual information from it's direct neighbouring nodes and the neighbouring nodes of its neighbouring nodes. This would happen for every node in the graph, with each node defining its own computation graph, based on its surrounding neighbourhood.

Increasing the number of message passing layers in the network, leads to increased incorporation of higher order information, as adding a layer means target node is aggregated not only with immediate neighbours, but also the neighbours of the immediate neighbouring nodes (Fig. 2.5). With fewer layers, a GNN will capture information from a smaller neighbourhood, for example the layer-1 network aggregates information from 1-hop neighbours. These shallower

networks are effective when local relationships dominate, but to learn global patterns or long-range dependencies deeper GNNs are required with more message passing layers. In this way a GNN model can incorporate more contextual and long-range spatial information to the final prediction than CNN models which are restricted by the fixed size of the local receptive fields. In Chapter 5 we compare the performance of patch based MIL methods against GNN methods which use patch-graph representations.

2.2.6 Application of deep learning to survival prediction

So far we have described deep learning methods in the context of diagnostic tasks such as tumour subtype classification, however, when attempting to understand patient prognosis, often time to event prediction can be a more flexible and informative way to understand the predictors of survival [143]. Unlike binary classification based on whether a patient has died or not, time-to-event methods model the occurrence of events, such as disease-specific death, as probabilities dependent on time [143]. Moreover, time-to-event prediction is able to handle 'censored' patients, which include patients who are not followed up all the way to their event time, for example, patients who drop out of a study before it ends meaning it is unknown when or if they have experienced the event of interest.

One of the most popular types of models used for survival analysis, is the Cox proportional Hazards (CPH) model also known as Cox regression [144]. This is a semi-parametric method, that models the relationship between the time until the event of interest occurs and the predictor variables of the event of interest. When carrying out survival analysis, predictor variables (also known as covariates) can be features that are associated with the event of interest based on prior knowledge, for example age is strongly associated with melanoma death [145], or exploratory predictors which are included to measure potential association with the event of interest. The key assumption of the CPH model, is that the Hazard function (which is the probability of the event of interest occurring at any time) for any two patients in the study is proportional over time. This means that the ratio of the hazards of the two patients with different covariates are constant and don't vary over time. This allows us to obtain hazard ratios (HRs), which examine the relative risk of the event of interest occurring in two different patient groups. This proportional hazards assumption also indicates that the patients risk of the event is based on a linear combination of the covariates, however, that does not allow for modelling of more complex real-world associations, that a non-linear risk function may be able to capture [146].

As early as 1995 Farragi and Simon [147] integrated Cox regression with a neural network, replacing the linear predictor the Cox regression model with an MLP. Nevertheless, this method did not improve upon the standard CPH model. Then in 2016 Katzman *et al.* [146] proposed the DeepSurv model, modeling nonlinear interactions between the covariates and risk function. Based on increased concordance index (C-index) metrics, DeepSurv outperformed traditional linear CPH and random survival forest methods for predicting survival using real and simulated data. Moreover, by replacing the MLP of DeepSurv with a CNN, Zhu *et al.* (2016) [148] and

Zhu *et al.* (2017) [149] performed end-to-end survival analysis using unstructured pathology WSIs. In more recent years, several studies have shown how neural networks integrated with Cox regression can be a powerful tool for determining at risk groups of patients [34, 150], from WSIs, which we will explore in Chapter 6.

2.3 Review of deep learning methods applied to melanoma whole slide images

In this section we examine the deep learning methods that have been applied to melanoma WSIs to predict molecular signatures and immune phenotypes. We also highlight some of the limitations of these works and current challenges within melanoma digital pathology research.

2.3.1 Predicting genetic subtypes of melanoma

Image-based tools, which use WSIs as inputs and deep learning pipelines for analysis, offer a cheaper and potentially quicker alternative to the genetic analysis currently used to subtype patients into risk groups [19]. Multiple works have show that deep learning methods can be applied to the classification of melanoma WSIs based on molecular subtypes, but have mainly focused on single gene alterations. For example, Sun *et al.* (2019) [151] developed a model to classify BAP1, a genetic mutation found in uveal melanoma. The authors extracted patches from WSIs and had an opthalmic pathologist label each of the 8176 patches, based on a 4-point scoring system reflecting proportion of BAP1 expressed in the patch. The authors then used a modified DenseNet-121 [152], a type of CNN, which was pretrained with ImageNet [98], by fine-tuning the final layers of the network on their dataset, to classify the patches. The methodology achieved an F1-score of 97.8%, but is limited by the small dataset size (n=47 patients) and reliance on a single pathologist's annotations. There was also a potential risk of data leakage within this study due to patches from the same patient WSI being included in both the test and training set. Then in 2021, Kim *et al.* [153] demonstrated that a two step approach could be implemented to detect BRAF mutational status, using melanoma WSIs as inputs. This involved training an Inception-v3 CNN [154] from scratch on the WSI patches using four GPUs to detect tumour rich regions which were annotated by a pathologist. Then to classify the BRAF-mutational status of the images, the authors again used an Inception-v3 CNN architecture [154] and found that when using transfer learning, the model achieved an AUC of 0.77 for classifying BRAF mutations in an independent cohort of patients and an AUC of 0.67 for the external TCGA cohorts. The authors cite the size of the datasets used for training and evaluating the models as a weakness of their study, but concede that there is a lack of available data with detailed annotations from a pathologist. Moreover in order to train the Inception-v3 [154] from scratch the Kim *et al.* (2021) [153] required four GPUs, which is a large amount of compute resource that is not accessible to all researchers.

As well as classifying driver mutations at a patch-level, several studies have looked at classification of genetic alterations at the WSI-level. Zhang *et al.* (2020) [155], like Sun *et al.* (2019) [151], developed a method for predicting BAP1 expression in uveal melanoma WSIs, however, their approach was informed by the assumption that if the patches surrounding a tumour patch of interest are mutation positive, it is more likely that the patch of interest is too, highlighting the importance of context in predicting BAP1 mutational status. Zhang *et al.* (2020) [155], initially extracted features from patches using a ResNet18 CNN [127] pretrained on ImageNet [98], these features were then arranged into a set of global feature maps based on their original patch locations. These feature maps were then used as an input for a U-Net [156], an architecture which uses a "encoder-decoder" architecture, to output a probability map for BAP1 mutational status across the WSI. The final binary classification was an average of the BAP1 status probabilities of all patches in the map. They achieved an AUC of 0.93 on the test set, but were limited by a small cohort of n=44 WSIs in the test set and n=144 WSIs in the training set, meaning they were unable to test how their model would perform on a more heterogeneous dataset.

Loeffler *et al.* (2022) [157], Komura *et al.* (2022) [158] and Fu *et al.* (2020) [159] all carried out pan-cancer studies utilising multiple cancer datasets from the TCGA repository to study the correlations between genomic features and image-based features from WSIs. Loeffler *et al.* (2022) [157] explored whether the expression of nine driver and suppressor genes involved in the pathognesis of cancer could be predicted using 136 melanoma WSIs. The tumour regions were manually annotated using QuPath [160] software and patches were extracted from the segmented tumour tissues. A ShuffleNet [161], an efficient CNN that was optimised for mobile devices, pretrained using ImageNet [98] was then fine-tuned to predict the mutational status of the genes at a patch-level. Then, to generate a slide-level prediction, the patch-level predictions were aggregated using a majority vote. The AUC performance yielded for different mutations varied with CDKN2 gene mutation classification yielding an AUC of 0.79, whereas performance for other genes, such as ARID2 (AUC: 0.51) or MTOR (AUC: 0.56), yielded results that were just slightly better than a random guess. Komura *et al.* (2022) [158] leveraged a slightly different approach, utilising a NASNet-A Large [162] CNN architecture pretrained with ImageNet [98] to generate patch features, through bilinear pooling of CNN layer filter responses. The authors then generated patch-level probabilities through logistic regression, with the average of all the patch probabilities being used to generate a slide-level the probability. The authors also examined how resolution affected performance, selecting the optimal model based on the mean AUC. The mean AUC performance of gene prediction varied from 0.63 for MAP2K1 to 0.90 for SF3B1. This work required pathologist annotations, with regions of interest hand-chosen by pathologists, moreover, pathologists manually assessed all 9,662 diagnostic slides removing n=926 due to staining and processing artifacts meaning the dataset is less likely to reflect a real-world dataset. Meanwhile Fu *et al.* (2020) [159] carried out fine-tuning of Inception-v4 [154], a CNN, to extract feature vectors, then using regression techniques to test the prognostic

associations of genetic aberrations, such as whole genome duplications, with survival achieving a mean Concordance score of C = 0.599 +/- 0.04.

2.3.2 Predicting immune subtype signatures of melanoma

Understanding the immune landscape of melanoma, has become a vital area of research, due to the advent and wide-spread administration of immunotherapy to treat patients with advanced disease [163]. Although immunotherapy has improved survival from 9 months to > 3 years for patients with metastatic disease [6, 87], there are still patients who don't respond to treatments. In addition, from the patients who do respond to treatments around 40% of those patients become secondary resistant to therapy [164, 165]. This is a problem as immunotherapy is both costly and toxic, so it is essential to identify patients who will respond best to these treatments, to prevent patients from unnecessarily suffering from side-effects. Pre-existing CD8+ T-cells, a type of immune cell, within the tumour region, have been found to be a positive indicator of how patients will respond to immune blockade treatments [64] and, as described in the previous section, TILs have been found to be associated with differential survival outcomes based on their spatial configurations (see Fig. 2.1). The development of automated image-based tools to quantify the presence of these immune cells, therefore, could offer a cost-effective way to stratify patients into immune-related treatment or risk groups.

Both Kather *et al.* (2020) [166] and Diao *et al.* (2021) [167] utilised the ShuffleNet efficient CNN architecture [161] with finetuning for feature extraction to classify melanoma WSIs into immune-related subsets. Diao *et al.* (2021) [167] attempted to predict the expression of immunomodulatory genes, specifically T cell immunoreceptor with Ig and ITIM domains (TIGIT) in melanoma WSIs from the TCGA archive. This method required both region annotations based on tissue type and point annotations based on cell type and models were trained at both pixel-wise cell and tissue-type prediction-level, with pathologists reviewing heatmap outputs and re-training until the output was deemed 'sufficiently trustworthy'. These features were then clustered into groups of human-interpretable image features, which were used as an input for a logistic sparse group lasso regression model to output a binary prediction of high or low TIGIT expression, with an AUC of 0.86. While human-in-the-loop systems offer a rigorous solution to imbuing explainability into AI systems, this level of supervision and annotation is not always possible and can be costly, as well as time consuming for pathologists [18].

Kather *et al.* (2020) [166] experimented with multiple CNNs pretrained with ImageNet [98] to infer the genotype of tumours using WSIs as inputs from the TCGA archive (N=430 patients, [with N=230 primary tumours and N=141 metastatic tumours]). Here they found that ShuffleNet [161], was the best performing CNN for genotype classification. Kather *et al.* (2020) [166] divided WSIs into patches and trained the CNNs to output a categorical prediction for the molecular status for each patch, they then used the percentage of positive patches for the molecular status as the slide-level probability score (this methodology was the basis of the work described early by Loeffler *et al.* (2022) [157]). As well as achieving an AUC of 0.73 for

classifying TP53 expression in melanoma WSIs, the group were able to classify multiple gene expression signatures, including specific immune signature expressions like Immune inflammatory Subtype C3 [168] (AUC: 0.738) , Macrophage Regulation [168] (AUC: 0.637) and CD8 positive lymphocyte expression [168] (AUC: 0.623) . While the AUC performance, is inferior to wet-lab testing [104, 169], the work still highlights the potential that image-based CNN tools for classifying melanoma WSIs based on the expression of immune cell signatures.

Sobottka *et al.* [170] explored immune phenotyping of melanoma whole slide images based on quantifying the expression of TILs in melanoma WSIs. Sobottka *et al.* (2021) [170], using 113 digitised IHC slides of metastatic melanoma cases, stained for CD8 protein to detect the presence of CD8+ T-cells and aimed to classify slides into three subgroups based on the presence and configuration of these cells. The three classes were the 'immune desert', which indicated tumours with no CD8+ T-cell infiltrate; 'immune exclusion', which indicated tumours where there were no CD8+ T-cells in the tumour region, but some at the invasive margin; and 'immune inflamed' if CD8+ cells were found with tumor parenchyma (directly interacting with tumour cells). Sobottka *et al.* (2021), following patch-based segmentation to detect tissue types using a DenseNet [152], utilised watershed algorithms and boundary detection to extract cellular features, then used different cut-off thresholds based on the calculated CD8+ T-cell densities [170]. The study was limited by difficulties in distinguishing preexisting lymphoid stroma and true tumor-associated stroma, which is also challenging for experienced pathologists. Moreover, rare subtypes such as spindle cell melanoma, were underrepresented in the small dataset and could limit the generalisability of the algorithm.

Studies by Saltz *et al.* (2018) [171] and Chou *et al.* (2020) [172], explored how TILs specifically are associated with patient survival. Saltz *et al.* (2018) ultilised a model to segment lymphocytes at a patch level and a model to detect necrosis at a pixel level from melanoma WSIs, using local high resolution 40x patches for TIL detection and larger lower resolution input patches to detect contextual regions for necrosis [171]. They then performed clustering on these TIL features and using multivariate analysis, found that an increased Banfield Raftery-index (cluster count) for TILs, was associated with a significant ($P < 0.001$) superior overall survival in melanoma patients. In addition, Chou *et al.* (2020), found that using that the percentage of TILs, measured through automated TIL scoring, was associated with recurrence free survival (adjusted HR=0.92 [0.84–1.00] per 10% increase in percentage of TILs) and overall survival (adjusted HR=0.90 [0.83–0.99] per 10% increase in percentage of TILs) [172]. In addition, Zormpas-Petridis *et al.* (2019) [173], showed the importance of the tumour microenvironment and spatial relationships of cells within melanoma H&E-stained images, through combining cellular neighbourhood information from a spatially-constrained CNN and regional tumour classifications from a multi-resolution super-pixel representation. Through this method Zormpas-Petridis *et al.* (2019) showed that a high ratio of lymphocytes in the tumor area compared to a low ratio of lymphocytes in the stromal compartment was associated with better overall survival in 94 melanoma patients [173].

Harder *et al.* (2019) [174] explored several different machine and deep learning methods, including classification and regression trees (CART), support vector machines (SVM), perceptrons and logistic regression, to try to uncover image-based features which could predict melanoma patient response to immunotherapy. The authors used a finetuned Inception-v1 CNN [154] to extract features from digitised histopathology images stained for the immune proteins CD3, CD8, and FoxP3. Then, by using these features as an input for a random forests classifier, they determined that CD8+ lymphocyte infiltration was associated with longer median overall survival and potential clinical therapy response. The study was limited by a small sample size of 31 patients, but showed how deep learning and machine learning methods can be used to quantify and characterise complex cellular interactions that may influence therapeutic response. Moreover, in 2020, Failmezger *et al.* developed a graph-based approach, which integrated genomic, transcriptomic data and histology image data to study the spatial interactions of cells in the stromal-immune interface [175]. Topological network analysis allowed the group to define two stromal features, which were related to lymphocyte distribution and overall survival, suggesting they have prognostic significance and could have implications for patients undergoing immunotherapy.

2.3.3 Current challenges and opportunities

Scarcity of melanoma datasets and annotated datasets

The majority of papers acknowledge the limited size of the datasets used, as limitation for their work. Small datasets can negatively impact model performance, as there can be a lack of variability in images for model learning, leading to less robust models that may not adequately represent the patient population. As discussed in the *Melanoma background* (section 2.1), melanoma presents as a highly heterogeneous disease, which is reflected in cellular, molecular and structural variation [47, 56]. Consequently large datasets are required to capture this and also represent rare variants, such as spindle cell variants [58]. Otherwise models may overfit to noise present within a limited dataset, further underscoring the need for larger melanoma datasets. Often papers cite using the TCGA dataset, which is a valuable opensource resource, but the melanoma tumors within the TCGA are highly selected at an advanced stage and have limited clinical metadata [1], so less likely to be reflective of the patient population. In Chapter 3, we describe the LMC dataset which includes 667 WSIs which we use in our subsequent works. The LMC is a large dataset which includes patients with early stage disease, with clinical metadata to allow for comprehensive analysis.

A potential contributing factor for papers citing a scarcity of datasets, is that there is a lack of annotated datasets, which are not always readily available. Across the papers referenced, many model pipelines required pathologist supervision for annotating tissues, performing quality control and checking the outputs of models. Currently, the human-in-the-loop approach is essential for validating models and adhering to medical regulations. However, this approach

also imposes limitations on the amount of data that can be effectively utilised. This is due to the laborious and time-consuming nature of pathologists annotating images, especially when pathology departments are already facing understaffing issues [18], making it challenging for them to balance this task with their primary responsibilities. Automated annotation tools such as QuPath [160] or the $HALO^{AI}$ platform [170] can help in automatically detecting cellular regions, however, these tools also have limitations, especially when dealing with melanoma WSIs, as is evidenced by their inability to process slides which contain high melanin content. This exclusion of certain skin types has been documented by Kulkarni *et al.* (2020) [176] and Sobottka *et al.* (2021) [170]. On the other hand, weakly-supervised approaches, as employed by Kather *et al.* (2020) [166], which only require a slide-level label, offer another potential solution for avoiding the need for pixel-wise pathologist annotations. Weakly-supervised approaches, such as MIL, have been shown to generate state-of-the art (SOTA) performance for classifying molecular signatures, such as microsatellite instability (a type of genetic alteration) in coloretal cancer WSIs [28, 30], therefore, we explore a MIL based approach in Chapter 4 to avoid the need for extensive annotations and avoid exclusion of certain skin types.

Attention and context-based approaches

Due to the large size of WSIs, most of the studies we noted relied on patch-level classification approaches, where an image is divided into its constituent patches, with each patch label being generated through pathologist annotation or inheriting the slide-level label. A CNN can then be used to generate features from each patch that can be passed through a prediction model to get a patch label prediction. The final slide-level prediction is then determined by aggregating the predictions of all individual patches, using a non-parameterised method like averaging or majority voting or a neural network. This method has several limitations, for example, as described in the previous sub-section 2.3.3, obtaining patch-level labels through pathologists manually annotating ROIs is time-consuming so cannot be scaled to large volumes of WSIs [129]. On the other hand, by assigning all patches with the same label as the slide-level label, there is a risk of introducing noise to the patch labels, as often the ROI which generates the label only occupies a small area of the image, encompassing a small subset of the patches. Assigning all of the patches within an image the same label also ignores the intra-tumour heterogeneity, which is common in melanoma histology [56], creating noise that can affect model learning and performance.

MIL frameworks are able to avoid these limitations, as they use a slide-level label for weak supervision during model training and only assume that a subset of patches or features within the image will inherit the slide-label. Moreover, in recent years different trainable aggregators to generate slide-level predictions have been developed [109, 128, 177, 178], such as attention-based MIL which can prevent the need for ROI annotations from pathologists, by learning high attention scores for tumour regions. Multiple works [179, 180, 181] and a 2023 study of two large multi-centre colorectal cancer datasets by [182] have now indicated that combination

of Attention-based MIL with a SSL feature extractor can improve molecular biomarker prediction, beyond using an ImageNet pre-trained approach with fine-tuning [20]. In Chapter 4, we therefore chose to examine the performance of six different MIL frameworks (including five methods with trainable aggregators that utilise attention mechanisms) for the task of classifying melanoma WSIs into the immune subgroups determined by [1],

Furthermore, works by Zormpas-Petridis *et al.* (2019) [173], Failmezger *et al.* [175] and Zhang *et al.* (2020) [155], indicate that spatial information and contextual features can improve the classification molecular subtypes and detection of cellular features in melanoma WSIs. Yet, graph-based methods are under-explored in the context of melanoma research. Studies examining graph-based approaches applied to other tumour types [24, 34, 183] have shown that context can give SOTA performance and also be applied in survival analysis and deriving *de-novo* biomarkers. In Chapter 5, we develop graph representations of melanoma WSIs and also examine whether multi-resolution representations improve the accuracy of the immune subgroup classifications.

Depth and aggregation of graph neural networks

As shown in Fig. 2.5, increasing the number of message passing layers and depth of GNNs can aid in modeling long range dependencies, by aggregating nodes that are further apart in a graph (multi-hop neighbours). However, as depth is increased, nodes begin to aggregate similar information, becoming more similar in feature space and indistinguishable from their neighbours. This problem is termed 'over-smoothing' and reduces the discriminative value of GNNs [184]. In addition, increasing the depth of GNNs can increase computational burden and the risk of overfitting. In Chapter 5, we explore how increasing the depth of GNNs impacts the performance of models and can contribute to over-smoothing, with different graph representations of WSIs. We also examine how reducing the size of hidden layers within GNNs can reduce memory constraints and overfitting of GNNs adapted for predicting patient survival in Chapter 6.

As described in the previous section, which aggregation method to use remains an open question for MIL methods. This is also the case for GNNs, therefore, in Chapter 5 and Chapter 6, we explore how attention-based aggregators can be used in GNNs and experiment with different WSI graph representations. Here we also aimed to test whether attention-based aggregation strategies improved model performance.

Improving feature representations

A common theme in methodologies applied to melanoma WSIs, is the application of different CNN architectures to derive image-based features for downstream analysis. Often this will mean utilising CNNs pretrained with the ImageNet dataset, or performing fine-tuning or transfer learning [151, 159, 166, 167, 174, 185] to derive more salient features for the task at hand, as was popularised by Coudray *et al.* (2018) [20]. However, in recent years, SSL has been proven

to be a promising method for representation learning [186] and potentially a new way to learn melanoma WSI representations. SSL does not require labelled training data, as it learns from signals within the data itself, avoiding the problem of lack of annotated melanoma datasets [186]. Moreover, multiple training strategies have been proposed for training models with pathology specific images [125, 177, 187], suggesting it could be a way to train feature extractor models to generate more salient pathology specific embeddings. In Chapter 4, we compare the downstream model performance for classifying melanoma WSIs into immune subtypes [1], when using three different ResNet CNN architectures [127], two pretrained with ImageNet [97] and one pretrained using contrastive SSL [126] with pathology-specific images [125].

2.4 Conclusion

There are multiple image-based histopathological and molecular features that are not included in the current melanoma staging due to subjectivity and variability in measurements. Automated deep learning pipelines could potentially capture these features and discover novel biomarkers. Yet, current works are limited by the lack of melanoma datasets and the need for pathologist annotations. In the following chapters we aim to address some of the challenges mentioned above, introducing the Leeds melanoma cohort dataset in the next chapter and exploring weakly-supervised deep learning methodologies for image analysis in the subsequent chapters.

Chapter 3

Datasets and image pre-processing

In this chapter we provide an overview of the two datasets which will be utilised in the next chapters. This chapter does not describe my direct contribution to data collection, but it gives an overview of those data and an examination of the strengths and limitations of both datasets, as they were essential to my work. Moreover, we discuss the generation of the ground truth labels and the standard pre-processing pipeline for the WSIs we use in the ensuing chapters.

3.1 The Leeds Melanoma Cohort

The primary data we used for this project is from the Leeds Melanoma Cohort (LMC) study [188, 189]. The cohort recruitment received ethical approval MREC 1/03/57 and PIAG3-09(d)/2003. This prospective cohort was assembled at the University of Leeds by a research team led by Prof. Julia Newton-Bishop as the clinician and principal investigator, with 2183 participants being recruited with a long follow-up. The study began recruiting participants from the North of England who had recently been diagnosed with melanoma of the skin [1] during the period of 2000-2012. The study aimed to recruit patients of all ethnicities, irrespective of skin, colour and had a 63% participation rate. 99% of participants were Caucasian and 1% other ethnicities. While this does reflect the incidence rates seen in England, where a study of incidence rates from 2013-2017, found non-White minority ethnic groups had lower incidence rates of melanoma skin cancer (less than 0.5%), it does not reflect the demographics of the wider population [190].

[1] In addition to cutaneous melanoma, which primarily affects the skin, melanoma can also be found in other sites of the body, for example mucosal and acral letiginous melanoma which can be found in mucous membranes (e.g. mouth) and soles of the feet, respectively. Moreover, aside from melanoma there are different types of skin cancer, such as squamous cell skin cancer, however, in this study the focus was cutaneous melanoma.

This extensive study encompasses a wide range of data types, including imaging, genetic, environmental exposures and clinical data. Information about participants' lifestyle, co-morbidities, and medication history was gathered through questionnaires, while clinical data, was extracted from clinical records. In our project, we specifically accessed the H&E slides, some label and expression data derived from the genetics and corresponding clinical data, such as AJCC staging, which we utilised in subsequent chapters for conducting multivariate survival analysis (Table 3.1). Additionally, we obtained survival data for the participants in the LMC study, which had been collected through the national cancer registry. The time of melanoma-related deaths was determined through various sources, including death certificates, records from general practitioners, hospital records, summary care records, or obituaries.

Patient enrollment began in the year 2000 and recruitment stopped in 2012, with the last follow-up being carried out in 2019. Prior to the advent of immunotherapy, advanced stage patients were offered a range of treatments from wider surgical resection, lymph node dissection, chemotherapy and radiotherapy [191]. This was patient specific, for example based on the age of the patient or site of the tumour. Despite the application of these different treatments, patient outcomes remained relatively poor with median survival of unresectable metastatic disease being 6–9 months [192, 193]. Fortunately, in 2012 the National Institute for Health and Clinical Excellence (NICE) published guidance recommending the first immune checkpoint inhibitor, ipilimumab, as an option for treatment of advanced melanoma in the UK [163]. The introduction of immunotherapy for the treatment of advanced melanoma, subsequently led to a striking improvement of patient outcomes, with median survival for patients with advanced melanoma increasing to almost 6 years when treated with a combination of CTLA-4 and PD-1 blockade [194]. Yet, because the introduction of the first immunotherapeutic regimens began in 2012, only a small minority of LMC participants will have undergone immunotherapy (moreover only around half of these patients will have responded to these regimens), so it should not significantly affect survival outcomes. This means in our analyses in Chapter 4 and Chapter 5, which attempt to stratify patients into the immune subgroups [1], which have potential treatment targets, will be relatively unaffected by treatment. This is a major strength of this study, as we can observe a consistent reflection of disease-based survival, without the effects of immunotherapy.

3.1.1 H&E slide and whole slide image generation

To generate H&E slides, 5µm histology sections were sampled by Dr. Jonathan Laye, from formalin-fixed, paraffin embedded primary tumour blocks using a microtome. The sections were then mounted onto Superfrost™ glass slides, dried overnight and then subjected to a hotplate at 70°C for a minimum of 20 minutes. Subsequently, a standard H&E staining protocol was performed by the staff in the Section of Epidemiology and Biostatistics at the University of Leeds. The slides were then digitised, by scanning them in batches using a Leica Biosystems Aperio AT2® Digital Pathology Slide Scanner, at 0.25 micrometers-per-pixel (m.p.p.).

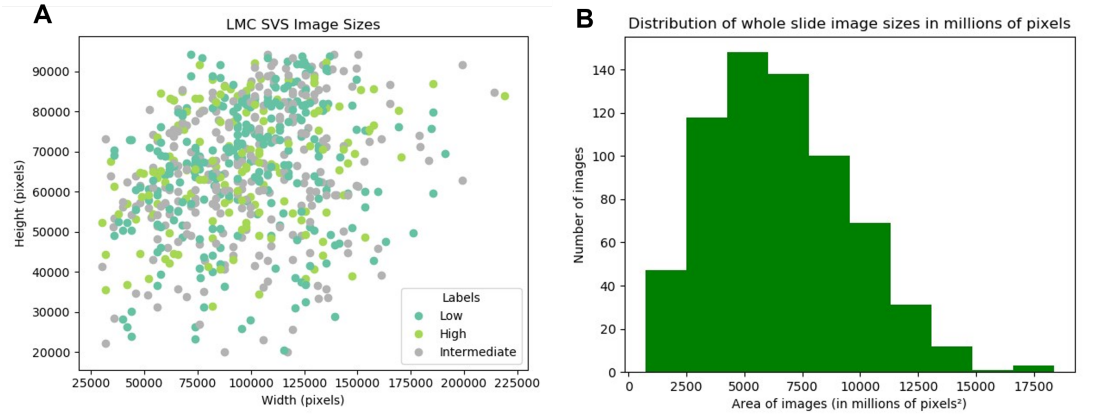


Figure 3.1: (A) The width and height of the 667 images with corresponding immune subgroup labels. (B) A histogram showing the distribution of the areas of the images in millions of pixels^2.

Out of the 2389 digitised tumour images, we worked with 667 of these WSIs (667 slides, 667 patients), as these cases had corresponding transcriptomic data available. These images range from 53MB-3.5GB in size (an average of 1GB/image and 668GB in total), with the smallest comprising of over 709 million million pixels^2 and the largest comprising of 18,392 million pixels^2 (Fig. 3.1B).

3.2 The Cancer Genome Atlas

We also utilised a second, independent, publicly available dataset from the Cancer Genome Atlas (TCGA), which contains tissue specimens from multiple hospitals across the world [195]. The WSIs (285 slides, 257 patients) are from both primary (n=177) and metastatic melanomas (n=107) from regional cutaneous or subcutaneous tissue (including satellite and in-transit metastases), regional lymph nodes and distant metastases. Metastatic lesions were included in our comparison against, as metadata from the primary melanoma cases suggested they had worse survival compared to metastatic patients, with all patients being biased towards late diagnosis. This bias towards larger tumours, may also account for the larger volume of extracted patches from the images (Table 3.2) and larger scanned image size (Fig. 3.2) compared to the LMC primary melanoma scanned WSIs.

We selected H&E stained FFPE diagnostic slides (eliminating frozen sections) and images which had been scanned in at 0.25 micrometers-per-pixel (m.p.p.), to allow for consistent results and observe whether our results were replicable in a second dataset. These images range from 89.96MB-3.76GB in size (and average of 1.38GB/image and 380GB in total), with the smallest comprising of over 365 million pixels^2 and the largest comprising of 20,196 million pixels^2 (Fig. 3.2B. The WSIs and corresponding RNA-seq data are available from the NIH genomic data commons (`https://portal.gdc.cancer.gov` [date last accessed: November 2022]).

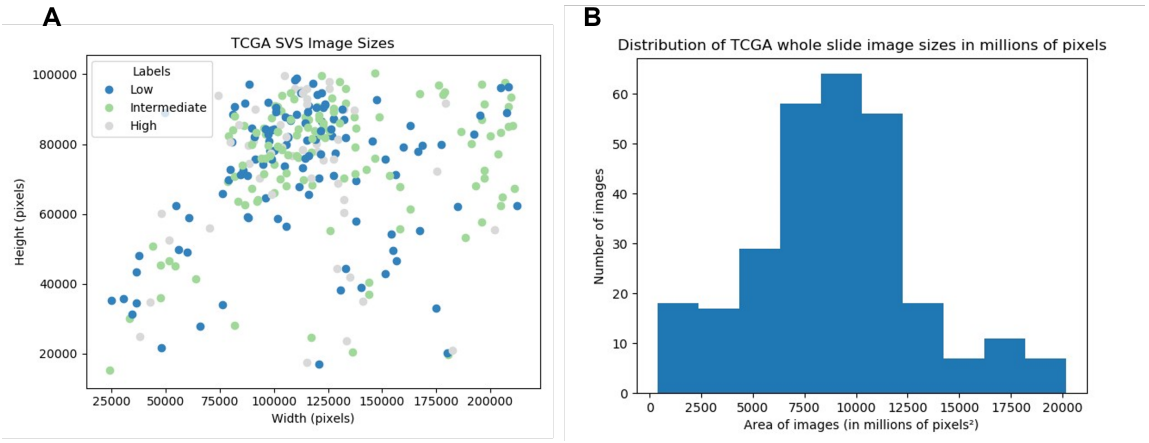


Figure 3.2: (A) The width and height of the 285 images with corresponding immune subgroup labels. (B) A histogram showing the distribution of the areas of the images in millions of pixels^2.

3.3 Limitations of the datasets

As previously mentioned, the LMC study was conducted as a population-based cohort study, drawing its participants from the North of England. Consequently, it is important to recognize that the findings we present may not be readily applicable to populations beyond the scope of this study. Additionally, it is noteworthy that issues of bias within AI have gained significant attention [196, 197]. While this study aimed to include patients from diverse ethnic backgrounds, it is worth noting that an overwhelming 99% of the participants were of Caucasian ethnicity. Although melanoma is typically more prevalent among individuals with lighter skin tones, it is imperative to emphasise that melanoma, as a type of skin cancer, can affect individuals of all racial and ethnic backgrounds [3].

Consequently, it is essential to acknowledge that the predictive models developed in this study may exhibit a bias towards classifying patients with lighter skin and may not be easily generalised to other ethnicities. Moreover as this is a retrospective study any associations we show would need to be further validated using a randomised controlled trial, in addition because the study was carried out before the new AJCC staging 8$^{\text{th}}$ edition was established, when carrying out survival analysis we use the 7$^{\text{th}}$ edition.

Moreover, as outlined by Nsengimana *et al.*, the cores taken for transcriptomic data were more likely to be sampled from thicker tumours, creating a sampling bias [84]. This was due to the protocol requiring no block to be destroyed in case patients required further testing using the tumour blocks during their treatment. This meant only 703 unique patients had transcriptomic data analysed (around a third of the total LMC participants), so while the mean thickness for the sampled patients was 3.0mm, the mean thickness of tumours was 2.2mm for the whole cohort. Furthermore we were only able to access 667 digitised images corresponding to the patients.

Unlike the LMC dataset, where patients are recruited from a geographically defined region, the patient data from the TCGA are collected from different data centres across the world. While this avoids geographical bias, it can lead to other problems, for example Dehkharghanian *et al.* (2023) [198], showed that networks can be trained to high accuracy to predict the centre a WSI comes from. This suggests images from different centres have different medically irrelevant visual features, potentially from different tissue processing procedures and unique imaging protocols, which could interfere with downstream applications. In addition, there may be domain shift in the data caused by other local demographics of patients from different centres which could contribute to batch effects in the TCGA dataset. Moreover, there is not detailed metadata for every patient, as for some cases clinical or pathological information can be missing. For example, the cases we examine in Chapter 4, only have overall survival for patients not disease-specific survival and we are missing the clinical data we show in Table 3.1 for the LMC dataset. This means we were limited in the data analysis that we could carry out for this dataset.

For the work I carried out in this thesis, I did not perform manual or automated quality control (QC). Whilst this may be seen as a limitation, as models can learn spurious correlations due to batch effects and artifacts [198], the reason I did not perform QC is that I wanted to maximise the data used for our studies and utilise data that is representative of the 'messy' data that is currently being used in clinical research. Moreover, in Chapter 4 and Chapter 5, we do show where overstaining can affect model performance [102], however this is a helpful finding and highlights a difficulty inherent to melanoma WSIs, which is the challenge of discriminating between melanocytes and TILs [199]. In future studies, it may be helpful to compare our results against models trained with datasets generated through QC, as this could improve the performance of the models presented in my work.

3.4 Immune subgroup ground truth labels

As we described in Chapter 2, one way to stratify patients and elucidate potential treatment targets is through the analysis of genetic data. In Chapters 4 and 5 we will explore how deep learning models can be used to classify WSIs into these immune subgroups, without additional genetic sequencing and analysis. The ground truth labels that we attempt to subtype the patients into were delineated by Pozniak *et al.* (2019), through clustering the transcriptomes from the 667 LMC patients, based on the inferred abundance of 27 immune cell types [1, 14]. The three subgroups are the 'high immune' class (22.5%), which corresponds to patients with a greater inferred immune cell infiltration in the primary tumour and better associated patient survival outcomes, the 'intermediate immune' class (39.1%), which corresponds to less inferred immune cell infiltrate in the primary tumour and the 'low immune' class (38.4%), which had the least inferred immune cell infiltrate in the tumour and worst survival response of patients. Pozniak *et al.* (2019), were also able to replicate these subgroups in the TCGA tumours. For

Table 3.1: Summary data of the clinical and clinicohistological features of the three LMC immune subgroups found in the paper published by Pozniak *et al.* (2019) [1]. LMC, Leeds Melanoma Cohort; WSI, Whole slide image.

	Low immune	Intermediate immune	High immune
Number of tumours	256	261	150
Melanoma death (%)	36	27	21
Age at diagnosis (median, years)	58	58	60
Sex (% males)	43	45	49
Breslow thickness (median, mm)	2.45	2.29	2.00
Ulcerated (%)	30	34	35
AJCC stage (%)	Low immune	Intermediate immune	High immune
I	30	36	39
II	50	50	47
III	18	13	14
Unknown	2	2	1
TIL classification (%)	Low immune	Intermediate immune	High immune
Brisk	12	8	13
Non-brisk	44	50	49
Absent	14	11	10
Unclassified	30	31	28
Histological subtype (%)	Low immune	Intermediate immune	High immune
Superficial spreading	49	49	49
Nodular	31	30	33
Lentigo maligna	1	2	1
Acral lentiginous	5	2	3
Other	5	10	7
Unclassified	8	7	7

our experiments, we worked under the assumption that each group with a distinct immune genetic signature would have a distinct histological pattern. In Chapter 4 and Chapter 5, we carried out initial experiments by training models using the three subgroups, but we also examined model performance when training and testing the models using only the 'high' and 'low immune' as these groups are more well defined compared to the 'intermediate' subgroup which is highly heterogeneous. Furthermore the 'high' and 'low immune' are more informative for immunotherapy options, so it is valuable to predict these classes.

In generating the transcriptomic data, later used to stratify patients into the immune subgroups, the slides were reviewed by Prof. Julia Newton-Bishop and Dr. Jonathan Laye. Prof. Julia Newton Bishop and her team selected areas of the H&E slide corresponding to the invasive tumour, which was marked for sampling, ensuring that the selected region had the least stromal content and the least infiltrating lymphocytes. The intent was to ensure comparability between tumours. The marking on the slide, was then used to orientate the corresponding FFPE block, so a 0.6-mm diameter microarray needle could be used to extract a tumour core from the FFPE block.

The mRNA transcriptomic data was then isolated from the cores and assayed using Whole-Genome DASL Illumina Array HT12.4 and normalised using standard methods as described in the study by [84]. The transcriptomic data used to generate the LMC immune subgroup labels was deposited into the European Genome Phenome Archive public repository with accession number EGAS00001002922. For the TCGA cases, which we used to replicate our findings in Chapter 4, the immune subgroup labels were formulated using RNA-seq expression data and were assigned to a subtype cluster based on which centroid had the strongest Spearman's correlation with the case [1].

3.5 Image processing

Due to the size of WSIs within the LMC and TCGA datasets, we pre-processed the images using a protocol designed by [200] to remove the white background from the images, as often this does not contain informative regions and helps to reduce the data size. Following this, we generated patches from the images, as this can be more computationally efficient when running models using GPU hardware. The patches were generated at different resolutions as this can capture different diagnostic information from the different levels of cellular and structural detail (Fig. 3.3 & Fig. 3.4). The methodologies for these preprocessing steps are discussed in further detail in the following subsection and we also describe how we split the data in order to carry out 10-fold cross validation for our experiments.

3.5.1 Segmentation and patch generation

The H&E tissue in all WSIs was segmented from the background by using downsampled versions of the images, then converting them from a red, green, blue (RGB) to hue, saturation, value (HSV) colour space in order to threshold regions containing tissue from background using the saturation channel. Morphological operations were then performed to close gaps and holes within the segmented tissue portion of the image and a threshold filter was then applied to remove foreground objects that did not meet the area threshold requirements. The tissue from the segmented images was then split into 256×256 pixel non-overlapping patches at five different resolutions (2.5x, 5x, 10x, 20x and 40x) (Fig. 3.3 & Fig. 3.4). To do this we used the OpenSlide library to access the levels in the pyramidal Aperio image format, then used downsampling and resizing operations for the resolutions in between these levels. The 40x magnification refers to the highest resolution version of the image, signifying the highest level of detail, resulting in the generation of the largest average number of patches per image. In contrast, 2.5x refers to the lowest resolution version of the image we generate patches from, creating the smallest number of patches per image (Table 3.2).

Table 3.2: The average number of patches extracted from the segmented tissue from LMC (N = 667) and TCGA (N = 285) WSIs at multiple resolutions. LMC, Leeds Melanoma Cohort; TCGA, the cancer genome atlas; WSI, Whole slide image.

Resolutions:	2.5x	5x	10x	20x	40x
LMC	113	406	1519	5824	20246
TCGA	239	902	3498	13647	54068

3.5.2 Dataset splits

For the following experimental chapters (Chapter 4, Chapter 5 and Chapter 6) the datasets were split with identical folds to allow for comparison. We implemented 10-fold Monte Carlo cross validation, where each fold was split with 80% of data being used for training data and 10% being kept for both the test and validation datasets. Datasets were split at a patient-level, to prevent different slides from the same patient being in the train, test and validation sets. The splits for the LMC dataset are shown in Table 3.3 and for the TCGA dataset in Table 3.4.

3.6 Conclusion

In this chapter we describe the two independent datasets, which were utilised in the experiments in the following chapters. We also discuss the strengths and potential limitations of the datasets in our research and how the immune subgroup ground truth labels used in Chapter 4 and Chapter 5 were generated. We also describe the preprocessing techniques we used to generate the input features for our experiments. In the following chapter, we will demonstrate how MIL

Table 3.3: LMC dataset splits, showing the number of WSIs labelled with the three different immune subgroups found by [1], within the training, validation and test sets, when carrying out Monte Carlo 10-fold cross validation. LMC, Leeds Melanoma Cohort; WSI, Whole slide image.

Immune subtype	Train	Validation	Test
Low	204	26	26
Intermediate	209	26	26
High	120	15	15

Table 3.4: TCGA dataset splits, showing the number of digitised WSIs labelled with the three different immune subgroups found by [1], within the training, validation and test sets, when carrying out Monte Carlo 10-fold cross validation. TCGA, the cancer genome atlas; WSI, Whole slide image.

Immune subtype	Train	Validation	Test
Low	92	11	11
Intermediate	98	12	12
High	39	5	5

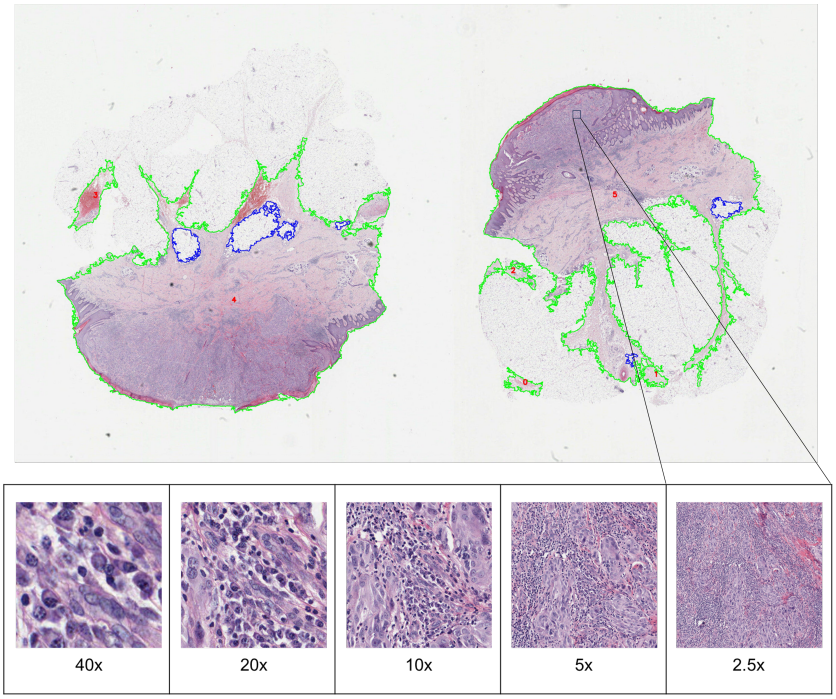


Figure 3.3: The top image shows segmented tissue from background from a LMC WSI. The bottom panel shows the different cellular detail and composition in 256×256 pixel patches from the image extracted at five different resolutions from the same region. LMC, Leeds Melanoma Cohort; WSI, Whole slide image.

models can be used to classify WSIs into the immune subgroups described above.

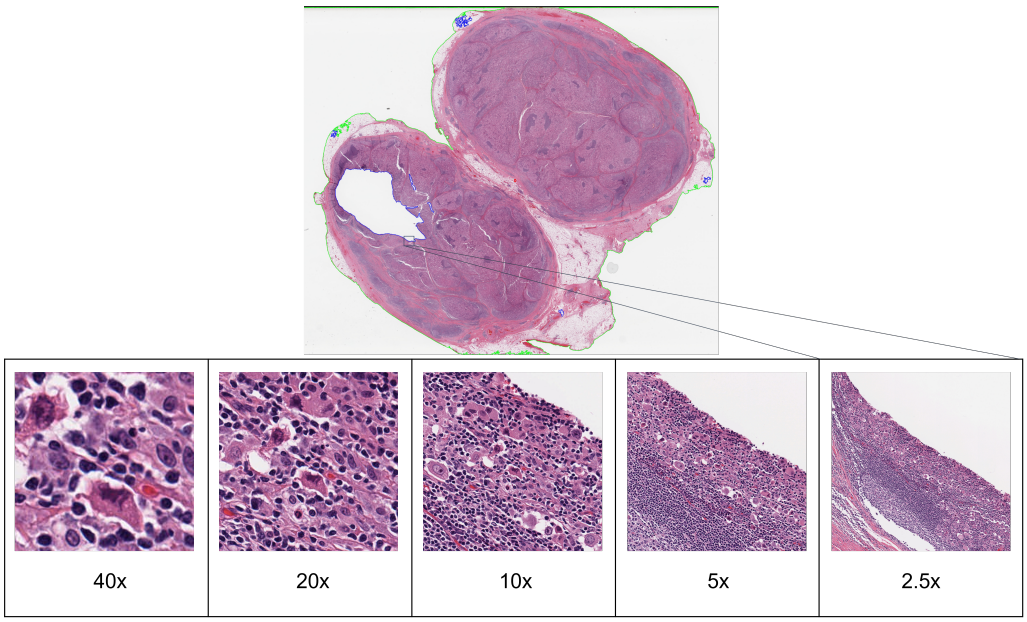


Figure 3.4: The top image shows segmented tissue from background from a TCGA WSI. The bottom panel shows the different cellular detail and composition in 256×256 pixel patches from the image extracted at five different resolutions from the same region. TCGA, the cancer genome atlas; WSI, Whole slide image.

Chapter 4

Immune subtyping of melanoma whole slide images using multiple instance learning

In this chapter we explore how multiple instance learning (MIL) methods can be applied to the task of classifying melanoma whole slide images (WSIs) into immune subgroups based on tumour transcriptomic data. We experiment with six different MIL methods, including dual-stream MIL [177], Transformer based MIL [178] and Clustering constrained attention MIL [109] the current state of the art methods for subtyping WSIs. We also modify the commonly used experimental pipeline by using a convolutional neural network (CNN) pretrained with pathology specific images for feature extraction and demonstrate how input patch resolution can affect performance. Overall in this chapter we carry out a systematic evaluation of MIL methods (carrying out 360 experiments) for immune subtyping of whole slide images, a task that could help clinicians determine which patients should receive different treatment regimes for better outcomes.

4.1 Introduction

Through the utilisation of digitised WSIs, it has become possible to apply computer vision algorithms with integrated computational pipelines for the analysis of H&E stained slides. Yet, the considerable size of these multi-gigabyte images and absence of annotation or pixel-wise labelling presents a challenge for processing WSIs directly in an end-to-end manner. To address this, multiple instance learning (MIL) frameworks, where an image is treated as a bag containing many instances that inherit the bag or slide-level label [106], have been implemented to effectively analyse WSIs [109, 112, 200, 201, 202, 203, 204]. In MIL frameworks, a histopathology

image can be subdivided into smaller patches, which can be further processed, using convolution neural networks (CNNs), to create feature representations. Following this, different ML algorithms or mathematical aggregators, such as maximum or mean operators [128], can be applied to the features to generate a slide-level label classification.

In recent years, there has been continued developments of new MIL frameworks based on the MIL paradigm. For example, in 2018, Ilse *et al.* [128] added a trainable attention-based pooling operator and in 2020 Lu *et al.* [200] further added to this work by implementing a clustering mechanism to further separate instances for cancer subtype classification. Li *et al.* (2021) [177] then proposed Dual-Stream MIL (DSMIL), applying an attention mechanism to both the instance-level stream and the bag-level stream and using the instance with the highest attention score to re-calibrate other instances within the bag. Moreover, Shao *et al.* (2021) [178], developed a transformer-based MIL (TransMIL) methodology, which processes images as sequences of instances to capture the relationships between these instances using self-attention mechanisms. This instance level information is then aggregated together to make a bag-level prediction. In this study we focus on implementing these aforementioned MIL models, because they have open source code and are well-established baselines, but many other notable variations on MIL models continue to be developed for computational pathology tasks with high classification accuracy [205, 206, 207, 208, 209].

When encoding the patches from the WSIs as inputs for MIL models, it has become common practice to use feature extractors pretrained on the benchmark visual recognition dataset ImageNet [111, 112, 113, 109, 108], which is comprised of millions of annotated natural images, from thousands of categories, such as food, locations, animals and people [98]. These CNN feature extractors, pretrained using ImageNet [98] are widely available, avoiding the need for researchers to have to train a feature extraction network from scratch, which can require large computational resources and the challenge of acquiring a large, diverse and high-quality annotated WSI dataset. Nevertheless, while using pathology-agnostic features from networks pretrained with ImageNet has been shown to generate effective results [111, 112, 113, 109, 108], the differences between the image domains, can lead to a reduction in accuracy for certain pathology specific tasks. For example, Yu *et al.* (2020) [21], reported that when detecting TP53 mutational status from H&E stained images, the classification is based on pixel intensity within the cytoplasm, which is not represented in non-histological image datasets like ImageNet. Furthermore Noorbakhsh *et al.* (2020) [210] found that network AUC performance improved when training CNN parameters on cancer images and noted that pre-training with cancer histology images improved distinction between classes where differences are more subtle. However pretraining a network can require large labelled datasets [20, 159, 202], which are not always available.

One solution that addresses both domain specificity and lack of annotated pathology images is self-supervised learning (SSL). SSL is a method which learns features from the signals within the data itself and does not rely on manual labels, hence SSL models can be trained using un-

labelled patches from WSIs. Abbasi-Sureshjani *et al.* (2021) [211] compared the performance of models using features from a CNN pretrained using a Bootstrap your own latent (BYOL) [118] SSL approach with H&E stained images, against features from a CNN pretrained with ImageNet. While the performance of the SSL approach and standard ImageNet approach were similar when classifying test set images from the same scanner, the SSL feature model was able to generalise better to images from an independent unseen dataset from new scanners. In addition, Saillard *et al.* (2021) [212] implemented SSL to pretrain a MoCo V2 model with TCGA images to generate feature representations for downstream MIL aggregators. They found that using SSL feature representations, consistently improved performance for classifying microsatellite instability (MSI) in colorectal and gastric cancer images, compared to using features from a 50 layer residual network (ResNet50) pretrained using supervised learning on the ImageNet dataset. Moreover Schirris *et al.* (2021) [203], demonstrated how using a histopathology-specific feature extractor, pretrained using simple contrastive learning for learning visual features (SimCLR [126] [an SSL approach]), improved classification of MSI and homologous recombination deficiency in breast and colorectal cancer datasets.

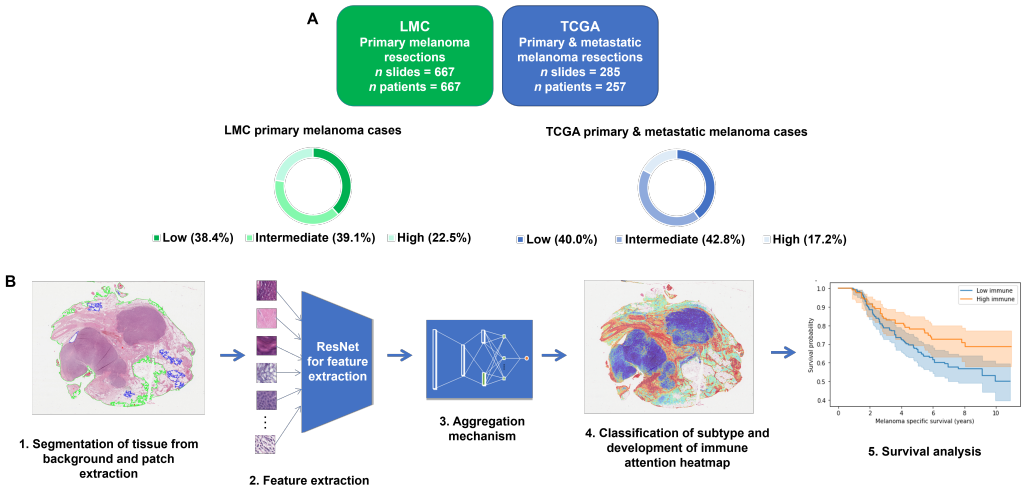


Figure 4.1: Datasets and experimental framework. A. Patient cohorts: two independent datasets that were used for training and testing of models and the distribution of immune subtype labels within these datasets. B. The images were processed, then three different feature extraction methods were used to generate feature embeddings. Six different MIL aggregation mechanisms were used to get final slide-level classifications. Immune attention heatmaps were used to interpret correctly classified and misclassified cases. Survival analysis was carried out to determine the prognostic ability of the best performing models. LMC, Leeds Melanoma Cohort; MIL, Multiple Instance Learning; TCGA, The Cancer Genome Atlas.

4.1.1 Contributions

In this study, we comprehensively assess the performance of six different MIL models, using three different feature extraction methods and five different patch resolutions on classifying melanomas into the subgroups delineated by Pozniak *et al.* (2019) [1] (see Fig. 4.1).

Our contributions are as follows: (1) We show that using a SSL pathology-specific ResNet18 to extract features can improve model performance for immune subtyping melanomas, compared to a ResNet18 and ResNet50 pretrained with ImageNet using supervised learning. (2) We show that MIL models which apply an attention mechanism are superior to standard max pooling MIL, but there is little difference between these attention-based approaches. (3) We demonstrate how 10x resolution input patches are superior for classifying 'high' and 'low immune' subtypes by providing a balance of cellular and contextual detail, illustrating this through immune attention heatmaps. (4) We implement survival analysis, to show the best performing MIL models are able to stratify LMC patients into prognostic subgroups.

4.2 Methods

4.2.1 Segmentation and feature extraction

The H&E tissue in the WSIs was segmented from the background using the protocol designed by Lu *et al.* (2021) [200] and described in more detail in Chapter 3. The tissue from the segmented images was then split into 256 pixel x 256 pixel non-overlapping patches at five different resolutions (2.5x, 5x, 10x, 20x and 40x). We utilised three different feature extraction methods, a modified ResNet50 CNN architecture [127], pretrained using supervised learning on the ImageNet dataset [98], a ResNet18 CNN, which was also pretrained using supervised learning on ImageNet and a ResNet18 pretrained using SSL on histopathology images [125]. The ResNet50 is often used as an upstream feature extractor in MIL tasks, so was chosen as a baseline and the ResNet18 pretrained with ImageNet was implemented as a method of comparison to the SSL ResNet18 as they have the same 18-layer architecture. Adaptive mean-spatial pooling [213] was utilised after the third residual block to modify the ResNet50, to extract 1024-dimensional feature embeddings from each patch. Whereas the ResNet18 architectures extracted 512-dimensional feature embeddings from the patches. The SSL ResNet18 is a publicly available SSL model [125] from `https://github.com/ozanciga/self-supervised-histopathology` (last accessed: November 2022), which has been trained using 57 multi-organ, multi-resolution (10x, 20x, 40× and 100x) histopathology datasets.

4.2.2 Model architectures and training

ResNet50 CNN representations were then further compressed by a fully connected (FC) layer, to K 512-dimensional vectors ($\mathbf{h}_{km}$), where K is the number of patch instances in the slide. The ResNet18 features were already this size, but were also passed through this FC layer. We

then tested different MIL pooling aggregation functions for classifying the immune subtypes from the patch embeddings.

Max pooling MIL

The baseline methodology we used was a variation on the max pooling MIL method, which was adjusted to predict the class probability distribution of every patch in the slide. It is referred to as MIL model in this Chapter. For this MIL model each patch feature (h_i) is transformed by a FC layer ($f(\cdot)$), and then passed through a linear classifier $\mathbf{W_c} \in \mathbb{R}^{512\times K}$, where $\mathbf{K}$ is the number of classes, to generate logits (unnormalized scores or outputs) for each class (K). Subsequently these logits are then converted to instance-level probabilities using a softmax normalisation function. An argmax operator is then used to select the highest probability instance for the positive class if it is a binary task or highest probability instance across all three classes when classifying the three immune subgroups. Once this top instance is identified $\mathbf{h}_{i*}$, it is passed through the neural network layers and softmax and argmax operators again to predict the class of the slide-level label ($\hat{y}$):

$$\hat{y} = \arg\max_j \text{softmax}(\mathbf{W}_c f(h_{i^*}))$$

where $\mathbf{h}_{i*}$ is the patch instance embedding that has the highest probability for a class j and provides the overall WSI label ($\hat{y}$). softmax($\mathbf{W}_c f\mathbf{h}_{i^*}$) represents the probability distribution for the K classes $j \in \{0, 1, \ldots, K-1\}$ for the $\mathbf{h}_{i*}$ instance. This softmax output is then used as the input for the argmax operator, which selects the class j that has the highest probability from the softmax output.

Attention MIL

Both of the attention-based MIL pooling functions with and without gating [128], calculate $\mathbf{z}$, the slide-level representation, through the aggregation of the patch feature embeddings and corresponding weights:

$$\boldsymbol{z} = \sum_{k=1}^{K} a_k \boldsymbol{h_k},$$

where h_k is the k^{th} patch instance embedding and a_k is the weight that is derived from the neural network's attention backbone. The attention weights (a_k) add to one to be invariant of the number of patch embeddings in a slide and ensure that the weights are interpretable as probabilities. This means that the instances (h_k) with higher attention weights (a_k) contribute more to the aggregated slide-level representation ($\mathbf{z}$). The weights for the attention mechanism without gating are formulated by:

$$a_k = \frac{\exp\left(\boldsymbol{w^T} \tanh(\boldsymbol{V h_k^T})\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{w^T} \tanh(\boldsymbol{V h_j^T})\right)},$$

where $\mathbf{w} \in \mathbb{R}^{d \times 512}$ and $\mathbf{V} \in \mathbb{R}^{512 \times 256}$ are the first and second FC layers of the neural network, respectively and d represents the dimensionality of the input feature embeddings (1024 for the ResNet50 and 512 for the ResNet18 feature embeddings). Hyperbolic *tanh* element-wise operations allow for gradient flow of both positive and negative values.

Gated Attention MIL

Moreover, we implemented a gated attention mechanism, developed by [214] to introduce sigmoid non-linearity for learning more complex relationships:

$$a_k = \frac{\exp(\boldsymbol{w^T} \tanh(\boldsymbol{V h_k^T})) \odot sigm(\boldsymbol{U h_k^T})}{\sum_{j=1}^{K} \exp(\boldsymbol{w^T} \tanh(\boldsymbol{V h_j^T})) \odot sigm(\boldsymbol{U h_k^T})},$$

where $\mathbf{U} \in \mathbb{R}^{512 \times 256}$ and $\mathbf{V} \in \mathbb{R}^{512 \times 256}$ are stacked FC layers parameterised by the network and form the attention backbone and $\odot$ refers to element-wise multiplications. Following the stacked FC layers, each class has a parallel attention branch, with an attention-based pooling function. The values from the aggregators are then used as inputs for a softmax function, which determines the overall slide-level prediction ($\hat{y}$).

Clustering-constrained-attention MIL

Furthermore, we evaluated clustering-constrained-attention MIL (CLAM) developed by [200]. This mechanism was developed to improve feature learning between classes, by using pseudo labels, with a support vector machine (SVM) loss function to increase separation between the B most and B least attended patches within an image [200]. After the first fully-connected layer $\boldsymbol{w_1}$, a clustering layer is placed with 512 hidden units to perform binary classification for each class, using sampled patch instances within the slide. As there are no patch-level labels for the supervised clustering, during training of the model, pseudo labels are generated from the attention network and performance is evaluated using top-1 SVM loss [215].

Transformer based MIL

Additionally, we implemented transformer based MIL (TransMIL), which is based on a correlated framework by [178]. Self attention and multi-head attention are leveraged to incorporate spatial information into slide-level classifications.

Dual-stream MIL

The final MIL model we implemented was Dual stream MIL (DSMIL) [177], which is another variation of MIL framework, which models bag representations and instance-level representa-

tions separately using two different streams. Within the instance-level stream, instance-level classifiers are used on each feature embedding and a max pooling operation is used to determine the highest-confidence prediction - the critical instance. Subsequently, attention scores from DSMIL are computed based on the similarity between each instance and the critical instance, using a distance measure. These attention scores are then used to re-calibrate and attention-weighted bag representation to get the slide-level classification.

The original DSMIL paper [177] also incorporates SSL feature extraction in the pipeline, however, as we wanted to compare the features extracted using the methods described above, we do not use this within our pipeline.

The source code for the CLAM, Attention, Gated-attention and max-MIL models were adapted from `https://github.com/mahmoodlab/CLAM` GitHub repository (last accessed: November 2022). In addition, the source code for the DSMIL and TransMIL models was adapted from the following repositories: `https://github.com/binli123/dsmil-wsi`, `https://github.com/szc19990412/TransMIL` and `https://github.com/secrierlab/HistoMIL` (all last accessed: July 2023).

Model training

The networks were trained using a cross-entropy loss function, comparing the slide label with the predicted slide-level label, to derive the parameters. For all models, excluding the TransMIL model, model parameters were optimised using Adam optimizer with a learning rate of $2 \times 10e^{-4}$, and weight decay of $1 \times 10e^{-5}$. For the TransMIL model, following the author's implementation [178], we utilised a Lookahead optimizer [216] with rectified Adam (RAdam) optimizer [217] with a learning rate of learning rate of $2 \times 10e^{-4}$, and weight decay of $1 \times 10e^{-5}$. Dropout with a probability of 0.5 was used after each layer of the attention backbone of the CLAM, attention and gated-attention models and after the FC layer in the max-MIL model. Dropout was applied in the Nystrom self-attention module [218] of the TransMIL models. In addition, within the DSMIL models, dropout was applied after the first linear layer in the instance-level classifier and applied before the first linear layer in the bag-level classifier.

We trained models using the three immune subgroups found by Pozniak *et al.* (2019) [1] and also examined training models using only the 'high immune' and 'low immune' subgroups. We implemented 10-fold Monte Carlo cross validation, where each fold was split with 80% of data being used for training data and 10% being kept for both the test and validation datasets (see Table 3.3 & 3.4). Datasets were split at a patient-level, to prevent different slides from the same patient being in the train, test and validation sets. Furthermore, during training, to mitigate class imbalances between subtypes, a slide was sampled proportional to the inverse of the frequency of its ground truth class. The models were trained for a minimum of 50 epochs, with early stopping if the validation loss did not improve for 20 epochs continuously, to prevent overfitting. Additionally we experimented with increasing the stopping time from 50 to 300 epochs for a subset of models to test whether increased numbers of bag instances, due

to images being at a higher resolution, led to the slower convergence of models.

To assess model performance, we calculated the mean area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (CI). When calculating performance for the three immune subgroups, the AUC scores were calculated for individual classes by binarising classifications, then averaging the AUC for each class. A classification threshold of 0.5 was implemented. In addition, we calculated the balanced accuracy and F1 scores for the LMC models with the highest AUC values for the binary classification task. Balanced accuracy is a weighted accuracy measurement that accounts for differences in class sizes within the dataset, by using both the summation of the recall and specificity divided by two. The F1 score is a similar metric which takes the harmonic mean of precision and recall and is bound between [0,1], where 1 represents maximum precision and recall values, and 0 represents minimum precision and recall values. We calculated these additional metrics to assess how imbalance within the dataset was impacting performance. We defined the positive class as the minority 'high immune' class to observe whether the models were biased towards the majority 'low immune' class.

4.2.3 Visual attention maps

Visual attention maps were developed by using the attention weight scores (a_k) for the patch embeddings. The attention scores were then scaled with the highest (1.0), being the most highly attended patch for the predicted slide-level label and the lowest (0.0) being the least attended patch for the predicted slide-level label. The scores were then converted to an RGB colourmap which is overlaid over the WSI, with red tiles indicating highly attended patches and blue tiles indicating low attention, which contribute less to the subtype label prediction.

4.2.4 Survival analysis

Melanoma specific survival (MSS) years were calculated using the time of melanoma diagnosis to the time of melanoma death as stated by death certificate, general practitioner records, hospital records, summary care records, or obituary. Patients who were alive without these events or died from unrelated reasons were right censored. 'High' and 'low immune' labels were derived from each type of MIL model with the best AUC performance for classifying the LMC test set ($N = 230$) and the TCGA test set ($N = 89$). As Monte Carlo cross validation was used, not all cases were sampled in the test set. We examined up to 11 years follow-up, with median follow-up years for the LMC dataset being 5.43 years and 2.67 years for the TCGA dataset.

A Kaplan–Meier estimator [219] was then used to estimate the survival probability:

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where $S(t)$ denotes the estimated survival at time t, $\frac{d_i}{n_i}$ denotes the probability of an event

occurring at each observed time point, d_i represents the number of melanoma deaths observed at time t_i, n_i represents the number of individuals 'at risk' just before time t_i and $\prod_{t_i \leq t}$ represents the product over the observed times t_i. A pairwise log-rank test was then used to measure if there was a significant difference between the survival distributions of the 'high immune' and 'low immune' subsets.

Cox proportional-hazards model

A univariate Cox proportional-hazards model [144, 220] was implemented to assess the association between the predicted 'high immune' and 'low immune' subgroups and the MSS in the LMC dataset. Where the univariate Cox proportional model can be denoted by:

$$h(t|X) = h_0(t) \cdot \exp(\beta X)$$

where $h(t|X)$ represents the hazard rate at time t for an individual with covariate values X, $h_0(t)$ is the baseline hazard function, which is unspecified and is estimated non-parametrically, β is the regression coefficient estimated by the model and X is the covariate value (the predicted immune subgroup) for the individual.

Multivariate Cox regression analysis was carried out with American Joint Committee on Cancer (AJCC) stage [39], age, sex and melanoma site (limbs vs. rest) as possible confounding factors. The multivariate Cox proportional-hazards model can be represented as:

$$h(t|X) = h_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

Where $X_1, X_2, \ldots, X_p$ are the covariates for the individual and $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients estimated from the model. Patients with missing data in any covariate were dropped from the models. P values < 0.05 were considered statistically significant. The performance of models was assessed using concordance index (C-index) [221, 222], which quantified whether patients are listed in the correct order based on the lengths of their of follow-up time.

4.3 Datasets

For the work presented in this chapter, the primary dataset we used was a set of 667 digitised WSIs of melanoma primaries (667 slides, 667 patients) with corresponding immune subgroup labels [1], obtained from the LMC [188, 189]. The second independent dataset, was 285 WSIs (285 slides, 257 patients) from both primary (177 slides, 176 patients) and metastatic melanomas (107 slides, 81 patients) that was obtained from the TCGA [195] (data available from the NIH genomic data commons [`https://portal.gdc.cancer.gov`]). More detail on these datasets and how the ground truth immune subgroup labels [1] were generated for each dataset is described in Chapter 3.

4.4 Results

4.4.1 Predicting high, intermediate and low immune subtypes

Initial experiments were carried out by training the models using the three subgroups determined by [1]. The three subgroups are the 'high immune' class, which corresponds to patients with a greater inferred immune cell infiltration in the primary tumour and better associated patient survival outcomes, the 'intermediate immune' class which corresponds to less inferred immune cell infiltrate in the primary tumour and the 'low immune' class which had the least inferred immune cell infiltrate in the tumour and worst survival response of patients. Moreover, each subgroup was enriched for different pathways, for example the 'high immune' class was enriched for pathways involved with immune cell signalling and the 'low immune' class was enriched for pathways cell-cycle, metabolism and immune suppression. Therefore we worked under the assumption that each group had a distinct immune genetic signature, which would confer a distinct histological pattern, that could be determined using a ResNet CNN for feature extraction.

We found that the model with the highest AUC for classifying the three subtypes using the LMC dataset, was an attention MIL model with a mean test AUC of 0.61 (95% CI 0.57–0.65) (Fig. 4.2A & 4.3). Whereas, for the TCGA dataset, a gated attention MIL model had the best AUC performance, with a mean test AUC of 0.68 (95% CI 0.63–0.73) (Fig. 4.2B & 4.3). Both models used features from 10x resolution patches, extracted using the SSL histopathology specific ResNet18. Due to the limited performance when classifying the three immune subtypes (Fig. 4.3), we examined confusion matrices for the top performing models, noticing the main misclassifications were between 'intermediate immune' and 'high immune' and the 'intermediate immune' and 'low immune' cases. Due to these errors, we simplified the task to a binary problem of classifying the 'high immune' against 'low immune' subtypes (Fig. 4.3), as we believed 'high' and 'low' images were more likely to have discriminable features, due to inferred immune infiltrate within the tumour being much more polarised, compared to the 'intermediate' subtype. Therefore, we decided to look at whether MIL models could classify 'high' and 'low immune' cases into prognostic groups.

4.4.2 Comparison of MIL models

We evaluated the performance of six different MIL models on the classification tasks. In general, the results indicate that the standard max pooling MIL model gives the poorest performance for our subtyping tasks (Fig. 4.3A-D). We also observed that during the Monte Carlo 10-fold cross validation process, the max pooling MIL model exhibited the greatest variability in performance. This was evident through the presence of wide 95% CI (Fig. 4.4 & 4.5), as well as substantial interquartile ranges (Fig. 4.3). Within our results, the max pooling MIL model trained using 20x ResNet18 ImageNet features, generated the largest 95% confidence intervals. This model achieved a mean test AUC of 0.50 through 10-fold cross validation, with a 95%

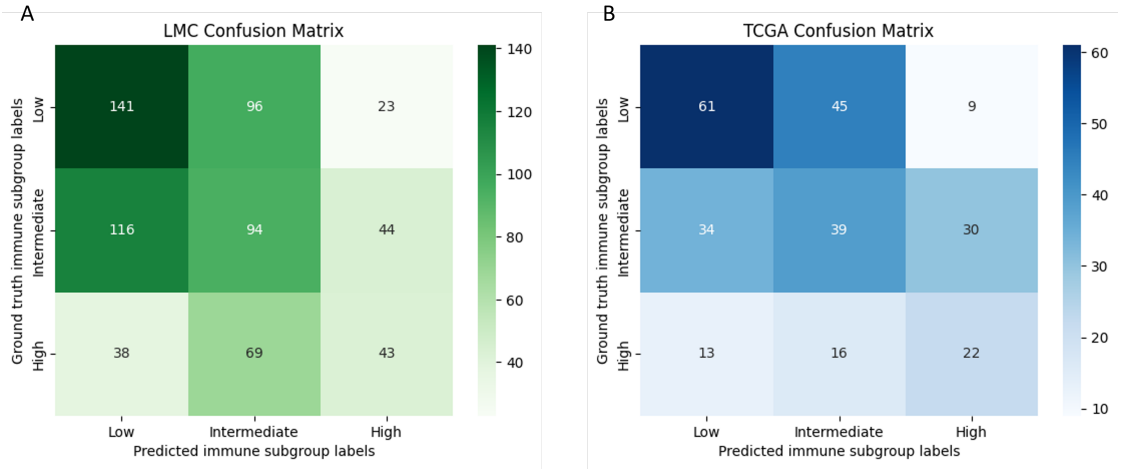


Figure 4.2: Confusion matrices for the model predictions of the 'low', 'intermediate' and 'high' cases on the 10-fold test sets. A. Predictions from the best performing model for classifying the three subtypes using the LMC dataset. B. Predictions from the best performing model for classifying the three subtypes using the TCGA dataset. LMC, Leeds Melanoma Cohort; TCGA, The Cancer Genome Atlas.

confidence interval spanning from 0.37 to 0.64. Moreover, we found the difference in test mean AUC values for this model type had a wide range, with the lowest test AUC being 0.46 (95% CI 0.40, 0.52) and the highest being 0.71 (95% CI 0.67, 0.75) for classifying the 'high' and 'low immune' cases in both datasets (Fig. 4.4A & 4.5A). max pooling MIL relies on the highest probability scored patch for the positive class being used to represent the slide-level prediction and can have a high sensitivity to outliers. As melanoma is a highly heterogeneous tumour, it might not be possible for a single patch to fully represent the diverse immune context for the different immune subtype classes, leading to poor performance in some cases.

In contrast the best performing model, for classifying 'high' and 'low immune' subtypes within the LMC dataset, was a DSMIL model, which utilised 40x ResNet50 ImageNet features and achieved a mean test AUC value of 0.80 (95% CI 0.76, 0.84) (Fig. 4.4F). However, we found that the TransMIL and DSMIL models, trained and tested with 10x ResNet18 SSL features from the LMC had similarly high performance, with both models achieving a mean test AUC of 0.79 (95% CI 0.74, 0.84) for classifying the 'high' and 'low immune' WSIs (Fig. 4.4E-F). Whereas for the TCGA tumours, the best performing model was an attention MIL model, which achieved a mean test AUC value of 0.82 (95% CI 0.75, 0.89) for classifying 'high immune' and 'low immune' from 10x patches (Fig. 4.5B).

When examining the difference in performance between the two datasets we also note that for the two immune subtype classification task, the TCGA dataset (Fig. 4.3B & D) test AUC values have a greater interquartile range (excluding the 10x max pooling MIL model results) compared to the LMC dataset results. This could be reflective of the dataset size, as the

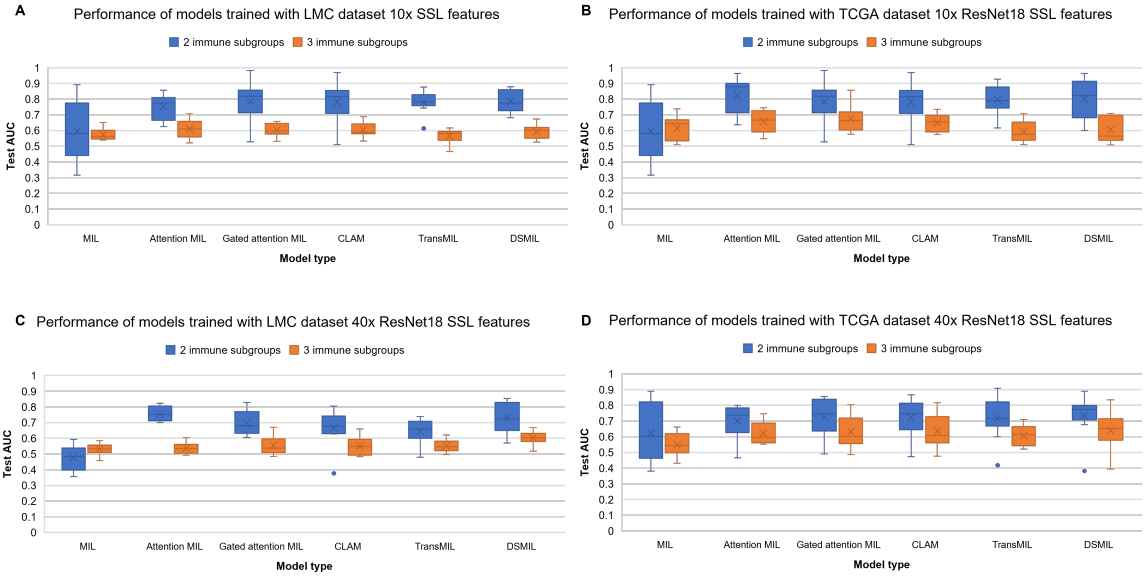


Figure 4.3: Performance comparison of the six different types of MIL models when classifying the melanoma immune subtypes in the two datasets. Box-plots showing the 10-fold test AUCs for each model type, when carrying out the 2 immune subtyping task ('high immune' and 'low immune' [in blue]) and three immune subtyping task ('high', 'intermediate' and 'low immune' [in orange]). The boxes show the quartile values and the whiskers extend to data points within 1.5× of the interquartile range. The input features for the models were extracted using the SSL ResNet18. AUC, Area Under the receiver operating characteristic Curve; LMC, Leeds Melanoma Cohort; MIL, multiple instance learning; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas.

TCGA dataset only has five 'high immune' cases in each fold, therefore, can be more sensitive to noise or different subsets within the heterogeneous dataset. Additionally, the classification task can be more challenging for this dataset as the cases come from both primary tumours and metastatic lesions, which are more advanced tumours from different body sites, so will contain heterogeneous imaging biomarkers. Moreover, the TCGA images come from multiple hospitals across the world, so the tissue processing and imaging protocols may be different, leading to the introduction of non-salient artefacts when classifying the images [198].

4.4.3 Comparison of feature extraction methods

When deploying MIL models, feature extraction is a key component within the pipeline. Often off-the-shelf CNNs pre-trained with the ImageNet dataset are utilised for this step, therefore we wanted to examine how using different feature extraction methods affected model performance. We found that using an SSL ResNet18 pre-trained with pathology specific images increased AUC performance at 10x and 20x patch resolutions for all the models, with the exception of the max pooling MIL models, DSMIL models and the TransMIL model trained and tested with

the TCGA dataset (Fig. 4.4B-E & 4.5B-D). In addition, the SSL ResNet18 pre-trained with pathology specific images improved performance for the CLAM (Fig. 4.4D) and attention MIL (Fig. 4.4B) models trained and tested with 40x resolution LMC WSIs and the CLAM (Fig. 4.5C) and gated attention MIL (Fig. 4.5D) models trained and tested with the TCGA dataset. We also saw a competing trend, when using ResNet50 ImageNet features, as for all the LMC models, excluding the LMC CLAM model, AUC performance steadily increased as resolution increased. Conversely, the best resolution when utilising ResNet50 ImageNet features from the TCGA dataset appears to peak at 20x for the majority of models (Fig. 4.5B-F). Models that utilise ResNet18 ImageNet features appear to under-perform for most resolutions, but the LMC MIL, LMC attention MIL, LMC gated attention MIL, LMC CLAM and TCGA MIL models have better performance when utilising 5x patches with this feature extraction method (Fig. 4.4A-D & 4.5A).

4.4.4 Comparison of resolutions

We compared how input patch resolution affected model performance for the two datasets. For the LMC dataset, when using features from the ResNet50 pretrained with ImageNet and the SSL ResNet18 pretrained using pathology images, performance improves when using 10x, 20x or 40x patches (Fig. 4.4 and 4.5). Furthermore, as mentioned previously, different feature extraction methods, appeared to perform the better at different resolutions. This is exemplified in Fig. 4.4C, which shows the mean test AUC values with 95% CIs of the gated attention MIL models. The models using the pathology specific SSL ResNet18 features have the highest performance using 10x resolution patches (mean AUC = 0.79 [95% CI 0.74, 0.84]), the models using feature embeddings extracted with a ResNet50 pretrained with ImageNet have the best performance when using 40x patches as input (mean AUC = 0.80 [95% CI 0.76, 0.84]) and yet the models using ResNet18 ImageNet features have the best performance when using 5x patches (mean AUC = 0.69 [95% CI 0.68, 0.79]). We also see differences between the datasets, for example, models that utilise features from the ResNet18 pretrained with ImageNet tend to have the highest mean AUC performance when using 5x input patches from LMC WSIs (Fig. 4.4A-D). Whereas the optimum resolution for the TCGA dataset WSIs is more variable for the ImageNet pretrained ResNet18 (Fig. 4.5).

We also wanted to examine whether the trend in increased performance using 10x input patches for features, could be due to faster model convergence during training caused by reduced numbers of patch embeddings in lower resolution images. To test this hypothesis, we experimented with increasing the minimum stopping count, when training the 40x embedding models, from 50 epochs to 300 epochs (Fig. 4.6). We believed that this would increase likelihood of model convergence, in models that were trained using a greater number of feature embeddings. When monitoring mean test AUC, we found that running the models for longer did lead to an incremental increase in performance for both TCGA and LMC max MIL, TransMIL, LMC CLAM and TCGA DSMIL models, nevertheless, in other cases it led to the opposite,

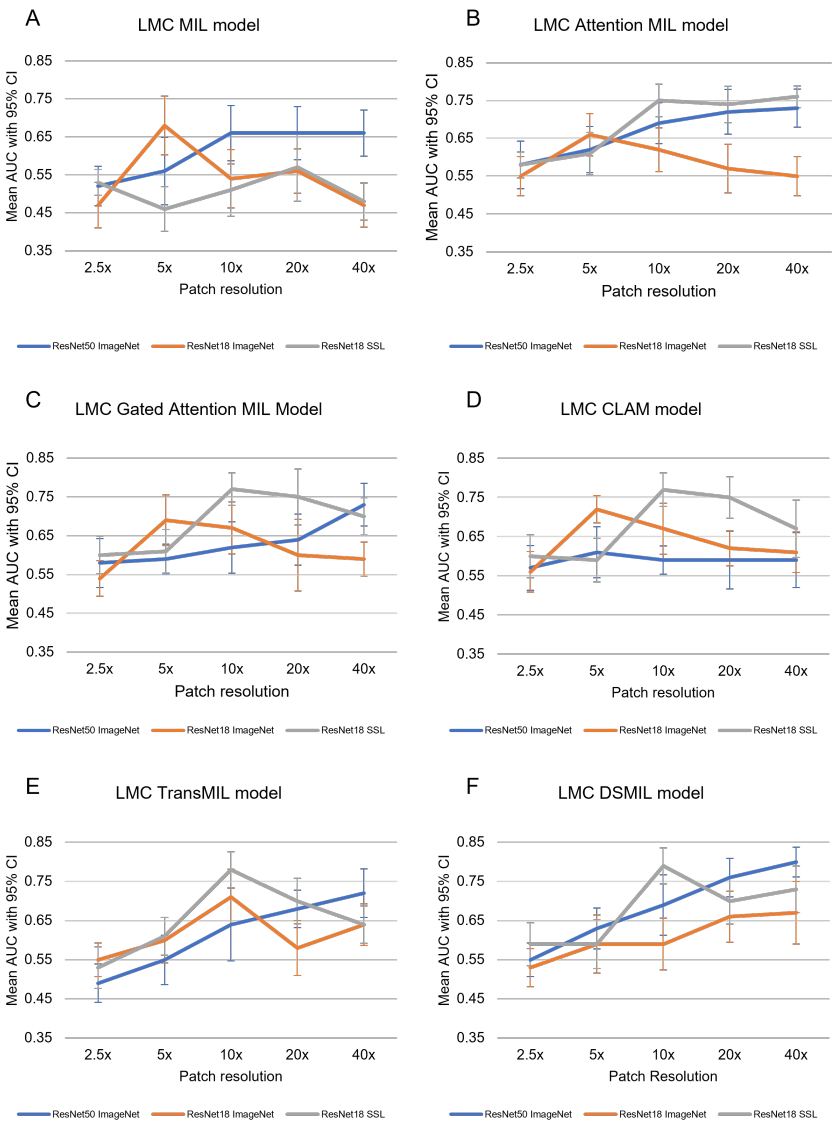


Figure 4.4: Model performance when altering the feature extraction method and resolution of input patches for models trained with the LMC dataset. Mean test AUC with 95% CI of MIL (A), attention MIL (B), gated attention MIL (C), CLAM (D) and TransMIL (E) and DSMIL (F) models for classifying 'high' and 'low immune' subtypes for the LMC primary melanoma dataset (n slides=667, n patients=667), when altering patch resolution and feature extraction method. AUC, Area Under the receiver operating characteristic Curve; CI, confidence interval; CLAM, Clustering-constrained-Attention Multiple instance learning; DSMIL, Dual Stream multiple instance learning; LMC, Leeds Melanoma Cohort; MIL, multiple instance learning; SSL, self-supervised learning; TransMIL, Transformer based multiple instance learning.

with the models trained for longer with higher resolutions, exhibiting a decrease in mean AUC performance. We found this for the DSMIL, attention and gated attention MIL models trained

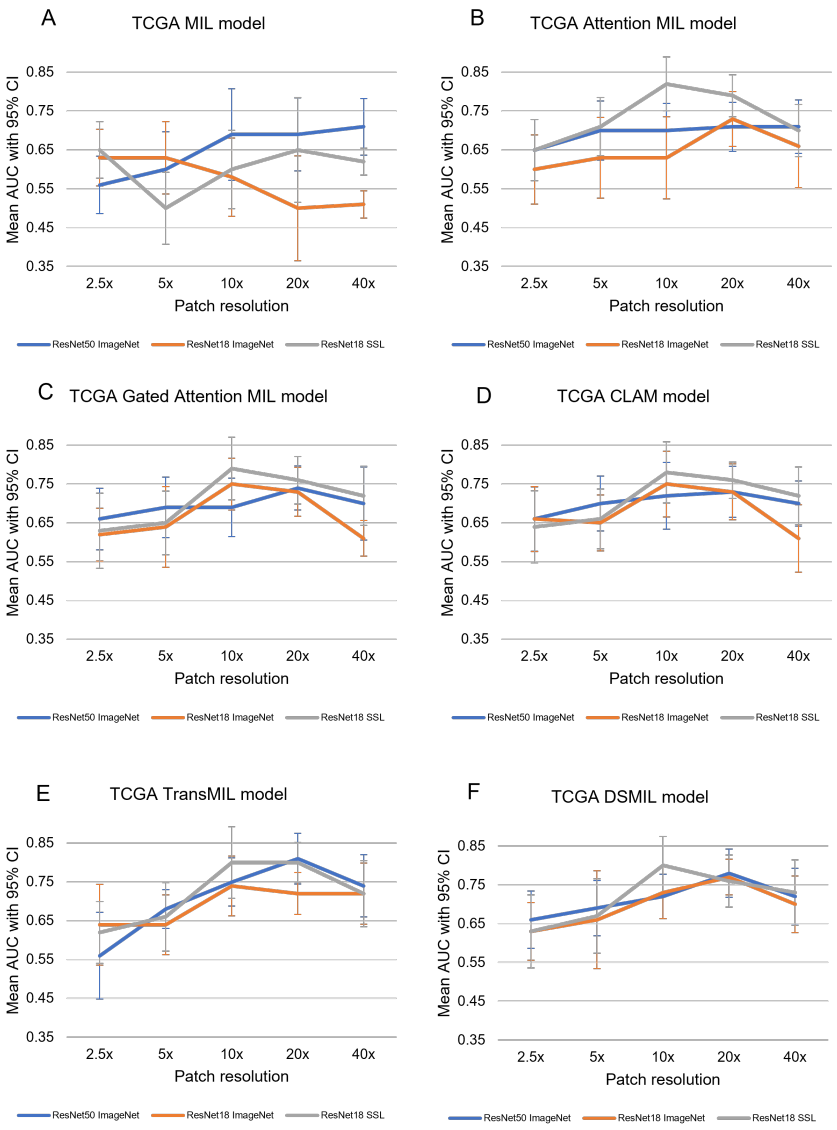


Figure 4.5: Model performance when altering the feature extraction method and resolution of input patches for models trained with the TCGA dataset. Mean test AUC with 95% CI of MIL (A), attention MIL (B), gated attention MIL (C), CLAM (D), TransMIL (E) and DSMIL (F) models for classifying 'high' and 'low immune' subtypes within the TCGA tumour dataset (n slides=285, patients=257), when altering patch resolution and feature extraction method. AUC, Area Under the receiver operating characteristic Curve; CI, confidence intervals; MIL, multiple instance learning; CLAM, Clustering-constrained-Attention multiple instance learning; DSMIL, Dual Stream multiple instance learning; TCGA, The Cancer Genome Atlas; TransMIL, Transformer based multiple instance learning.

with LMC data, with all models performing worse when increasing the number of epochs a

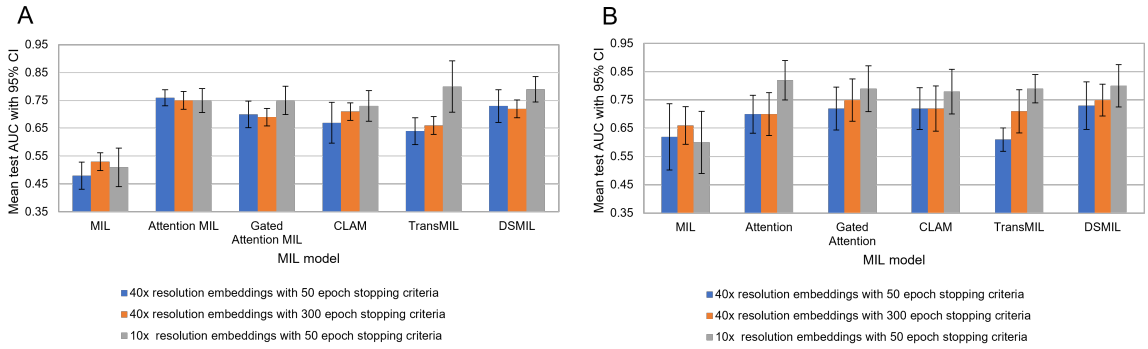


Figure 4.6: Mean 10-fold test AUC performance with 95% CI of MIL models classifying 'high' and 'low immune' subtypes with different epoch stopping lengths during training. A. Results from models trained with ResNet18 SSL feature extractor embeddings from the LMC dataset. B. Results from models trained with ResNet18 SSL feature extractor embeddings from the TCGA dataset. AUC, Area Under the receiver operating characteristic Curve; CI, confidence intervals; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas.

model was trained for (Fig. 4.6). Moreover, we found that most of the models trained with 10x input patch embeddings exhibited higher mean AUC performance compared to models utilising 40x feature inputs. This was despite the latter being trained for an extended duration, with the exception of the max pooling MIL models and LMC attention MIL models. This implies that the enhanced performance of models trained with 10x patch embeddings cannot be attributed to quicker model convergence and is more likely to be associated with improved feature representations at this resolution.

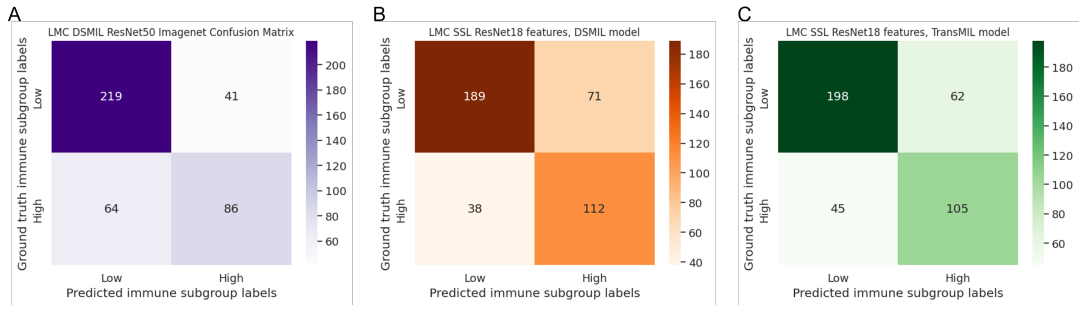


Figure 4.7: Confusion matrices for the model predictions of the 'low', and 'high' cases on the 10-fold test sets. A. Predictions from the LMC DSMIL model trained with 40x resolution ResNet50 ImageNet features. B. Predictions from the LMC DSMIL model trained with 10x resolution ResNet18 SSL features. C. Predictions from the LMC TransMIL model trained with 10x resolution ResNet18 SSL features. DSMIL, Dual-Stream multiple instance learning; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; TransMIL, Transformer-based multiple instance learning.

Table 4.1: Mean 10-fold test performance metrics with 95% CI for the three models with the highest mean test AUCs for classification of 'high' and 'low immune' LMC cases. The DSMIL 40x model, indicates the DSMIL model trained with ResNet50 ImageNet feature embeddings from 40x patches. The DSMIL 10x and TransMIL 10x model indicate the DSMIL and TransMIL models, which were trained using a SSL ResNet18 feature embeddings from 10x patches. AUC, Area Under the receiver operating characteristic Curve; CI, Confidence intervals; DSMIL, Dual Stream multiple instance learning; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; TransMIL, Transformer based multiple instance learning.

Model	AUC (95% CI)	Balanced accuracy (95% CI)	F1 score (95% CI)
DSMIL 40x	**0.80 (0.76, 0.84)**	0.71 (0.66, 0.76)	0.61 (0.53, 0.69)
DSMIL 10x	0.79 (0.74, 0.84)	**0.74 (0.69, 0.79)**	**0.67 (0.61, 0.73)**
TransMIL 10x	0.79 (0.74, 0.84)	0.73 (0.68, 0.78)	0.66 (0.60, 0.72)

Finally, to further investigate how resolution and feature extraction methods affected performance, we examined the confusion matrices of the three LMC models with the best AUC performance metrics (Fig. 4.7). Here, we saw that although the DSMIL model, which used 40x input features extracted using a ResNet50 pretrained with ImageNet had the highest AUC performance, this model also had the greatest number of misclassifications of 'high immune' cases with a 57.3% classification accuracy for this subtype (Fig. 4.7A). In contrast the TransMIL and DSMIL models which used ResNet18 SSL 10x features had accuracies of 70% and 74.7% respectively, for classifying the 'high immune' cases. Moreover, we calculated balanced accuracy and F1-scores for these models, to examine performance with respect to the imbalance in our dataset (Table. 4.1). Here, we found that the LMC DSMIL model trained and tested with 10x resolution ResNet18 SSL features had the highest mean balanced accuracy of 0.74 (95% CI 0.69, 0.79) and highest F1 score of 0.67 (95% CI 0.61, 0.73) of the three models. In contrast, the LMC DSMIL model trained and tested with 40x resolution ResNet50 ImageNet features had the lowest mean balanced accuracy of 0.71 (95% CI 0.66, 0.76) and F1 score of 0.61 (95% CI 0.53, 0.69), suggesting that this model is biased towards classifying the 'low immune' majority class. Here, we demonstrate a major weakness of AUC when used as a metric alone, as its insensitivity to class imbalance does not reflect the high false negative rate and weaker ability to recognise 'high immune' subtype cases of the DSMIL model trained with 40x pathology-agnostic features.

4.4.5 Immune attention heatmaps for explainability

Furthermore, we developed immune attention heatmaps, for three of the MIL models with highest mean AUC performance, showing where the patches with the highest and lowest attention weights were located. This was to help with explainability and to gain a better understanding of model performance. In Fig. 4.8, we compare heatmaps from a 'high immune' LMC case, that were generated from attention scores from DSMIL models trained with different feature embedding inputs. For both models we found that high attention patches are concentrated in

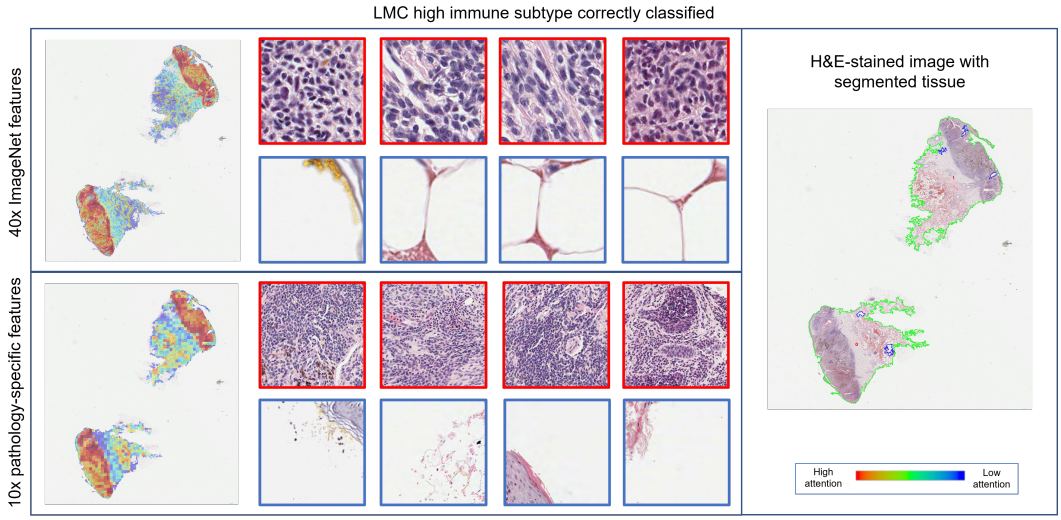


Figure 4.8: Comparison of correctly classified LMC 'high immune' tumours using different input features. Immune attention heatmaps with the original H&E stained WSI and four patches that contributed the most (red) to the 'high immune' subtype prediction and four patches that contributed the least (blue). The heatmap and high attention patches in the top panel are from a DSMIL model, which used 40x input features that were generated from a ResNet50 pretrained with ImageNet. The heatmap and high attention patches in the panel below are from a DSMIL model, which used 10x input features that were generated from a ResNet18 pretrained using SSL with histopathology WSIs. DSMIL, Dual Stream multiple instance learning; H&E, Haematoxylin and eosin; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas; WSI, whole slide image.

similar regions (the red areas), where there are tumour infiltrating lymphocytes (TILs), the smaller darkly stained circular cells, confluent with tumour cells. Moreover, we saw this replicated in the TCGA attention MIL model results, with high attention patches containing both TILs and tumour cells (Fig. 4.10).

We also examined correctly classified 'low immune' cases and found that high attention patches came from regions that contained tumour cells with large nucleoli, nests of melanocytes and an absence of TILs (Fig. 4.9 & 4.10). When examining the differences between the two DSMIL models trained and tested with different features, we show that the model trained with 40x features from the ResNet50 feature extractor, assigns high attention patches to both informative tumour, but also uninformative regions as shown in magnified section under the H&E-stained image in Fig. 4.9. This region is a large blood vessel in the dermis, which is unlikely to have prognostic value, suggesting models trained with pathology-agnostic features are less likely to attend to salient cellular regions and can be 'distracted' by regions with more staining. Whereas the DSMIL model trained using pathology-specific 10x embeddings, has less high attention scores in this area, with more high attention scores being learnt for embeddings from the tumour region. In addition, for all TCGA and LMC models we saw that low attention

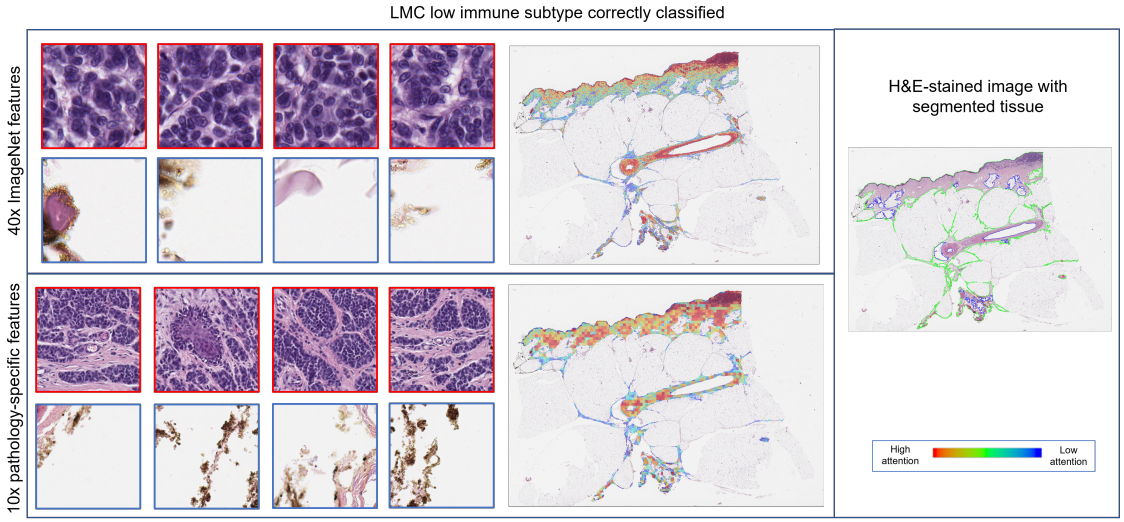


Figure 4.9: Comparison of correctly classified LMC 'low immune' tumours using different input features. Immune attention heatmaps with the original H&E stained WSI and four patches that contributed the most (red) to the 'low immune' subtype prediction and four patches that contributed the least (blue). We also highlight a region from the H&E stained image that the models show different levels of attention. The heatmap and high attention patches in the top panel are from a DSMIL model, which used input features that were generated from a ResNet50 pretrained with ImageNet. The heatmap and high attention patches in the panel below are from a DSMIL model, which used input features that were generated from a SSL ResNet18. DSMIL, Dual Stream multiple instance learning; H&E, Haematoxylin and eosin; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; WSI, whole slide image.

patches came from uninformative regions that had not been removed through thresholding, for example regions that contained a lot of white background, or black shadowy regions produced during image scanning.

Moreover, we also wanted to examine cases that were misclassified to further our understanding of MIL models and to see potential limitations of the experimental design. In Fig. 4.11, the DSMIL model which used 40x features extracted using a ResNet50 pretrained with ImageNet, misclassified the LMC 'high immune' case as 'low immune', even though the high attention patches appear to focus on darker stained regions which could be TILs. While this would indicate a 'high immune' case, the heavy staining can cause difficulty in pathological interpretation since it reduces the variation in staining within a nucleus, when variation in haematoxylin intensity within a nucleus can be a useful indicator of malignancy [102]. High variation in nuclear staining is caused by clumped chromatin suggests a disorder to the chromosomal arrangement. By contrast benign nuclei often have very a smooth chromatin pattern, which is shown as little variation in haematoxylin staining, therefore heavy staining creates difficulty in determining whether darker cells are benign melanocytes, malignant melanocytes, or TILs. Additionally, the stretched, elongated morphology of the cells that surround the darker

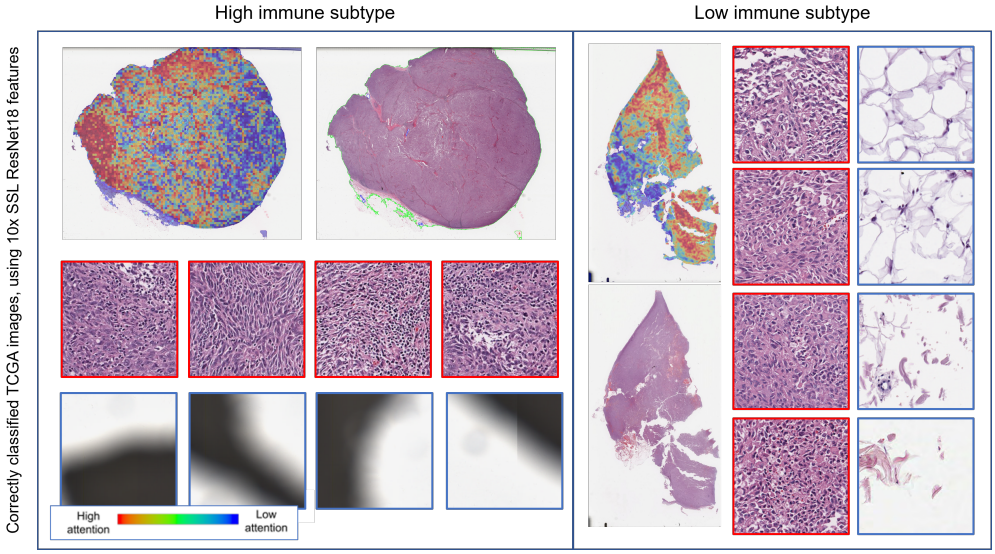


Figure 4.10: Correctly classified 'high' and 'low immune' TCGA tumours. Immune attention heatmaps with the original H&E stained WSI and four patches (red) that contributed the most and least (blue) to the subtype predictions. Input features were generated using 10x feature embeddings from the SSL ResNet18 and an attention MIL model was used for the slide-level classification. H&E, Haematoxylin and eosin; MIL, multiple instance learning; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas; WSI, whole slide image.

cells, indicate this case could be a rare spindle cell variant of melanoma, which are less represented in our dataset. We also show a TCGA 'high immune' case (Fig. 4.11), which was also misclassified as 'low immune'. Here, although the high attention regions show some lymphocyte infiltration, the high attention feature patches come from outside the darker tumour region in pink areas that show necrosis. Moreover, visual inspection of the tumour region shows there is less TIL infiltration in the dense tumour regions compared to nodal lymphocytes. Highlighting the significant challenge of representing the heterogenous landscape of metastatic and primary melanoma tumours, using labels derived from transcriptomic data sampled using a 0.6-mm microarray needle biopsy [84].

We also wanted to further understand why the LMC 40x DSMIL model had higher AUC performance compared to the LMC 10x DSMIL model, which had a higher balanced accuracy and F1 score. In Fig. 4.12, we show an image where the 40x DSMIL correctly classified the image as 'low immune', but the 10x DSMIL model predicted the image to be 'high immune'. The 40x DSMIL model that uses pathology-agnostic features, learns high attention scores for inert acellular regions, such as areas containing keratin (Fig. 4.12A). Whereas the pathology-specific model learns high attention scores for cellular regions, with more high attention (red) patches localised in the tumour region of the tissue compared to in the granular layer (Fig. 4.12B). These regions also appear to contain dark cells that could be TILs, suggesting the ground truth 'low immune' label may not be capturing the entirety of the tumour microenvironment. Moreover,

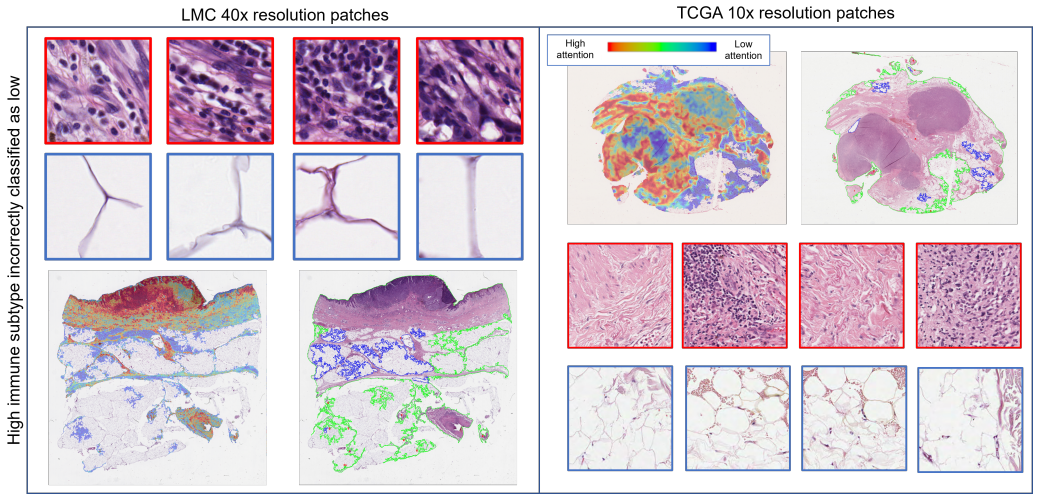


Figure 4.11: Incorrectly classified 'high immune' LMC and TCGA tumours. Immune attention heatmaps with the original H&E stained WSI and four patches (red) that contributed the most and least (blue) to the incorrect 'low immune' subtype prediction. For the LMC WSI, the input features were generated using 40x feature embeddings from the ResNet50 pretrained with ImageNet and a DSMIL model was used for the slide-level classification. For the TCGA WSI input features were generated using 10x feature embeddings from the SSL ResNet18 and an attention MIL model was used for the slide-level classification. DSMIL, Dual Stream multiple instance learning; H&E, Haematoxylin and eosin; LMC, Leeds Melanoma Cohort; MIL, multiple instance learning; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas; WSI, whole slide image.

this reinforces the evidence suggesting, the 40x DSMIL model using pathology-agnostic features is biased to predicting 'low immune' using non-salient features and not recognising immune cells within the tumour microenvironment.

4.4.6 Prognostic associations of model predictions

We assessed the prognostic ability of each of the different types MIL models, for stratifying the images into 'high immune' and 'low immune' subsets. We only show the MIL models which achieved the highest mean test AUC performance on the LMC dataset. The Kaplan Meier plot (Fig. 4.13) shows the survival distributions of the 'high' and 'low immune' subtypes were significantly different (log rank test $P < 0.05$), for all models apart from the attention MIL model (Fig. 4.13B), showing strong evidence that the models are able to stratify patients into groups associated with MSS.

Univariate and multivariate Cox proportional hazard models were also implemented to assess the prognostic ability of the models for stratifying patients into 'high immune' and 'low immune' subsets. The 'low immune' subgroup as predicted by all models apart from the attention MIL model, had a significantly higher hazard of melanoma death compared to the 'high immune'

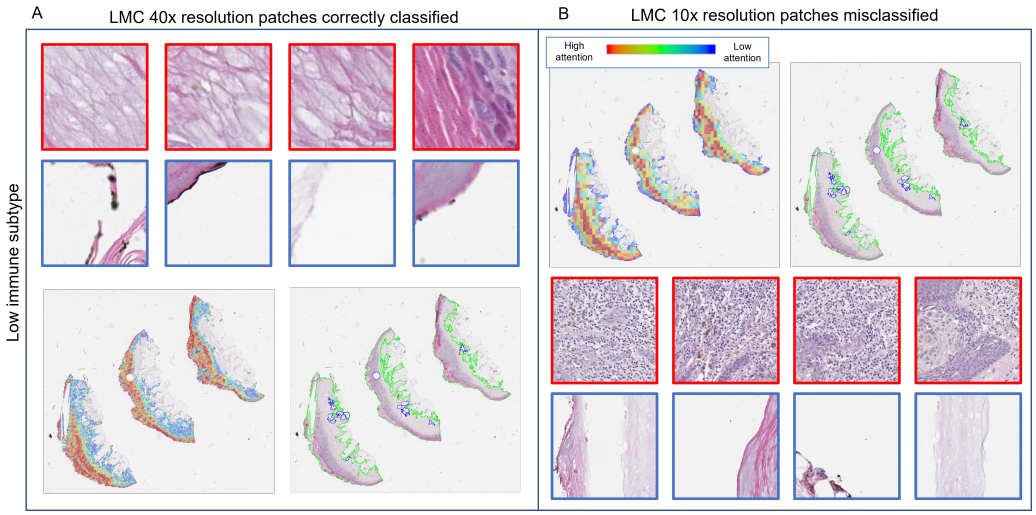


Figure 4.12: LMC 'low immune' image that was correctly classified by one model and misclassified by another model that used different input features. Immune attention heatmaps with the original H&E stained WSI and four patches that contributed the most (red) to the 'low immune' subtype prediction and four patches that contributed the least (blue). The heatmap and high attention patches in the top panel are from a DSMIL model, which used input features that were generated from a ResNet50 pretrained with ImageNet. The heatmap and high attention patches in the panel below are from a DSMIL model, which used input features that were generated from a SSL ResNet18. DSMIL, Dual Stream multiple instance learning; H&E, Haematoxylin and eosin; LMC, Leeds Melanoma Cohort; SSL, self-supervised learning; WSI, whole slide image.

patient subgroups (Table 4.2). Moreover when adjusting for clinical predictors (age, sex, tumour site, AJCC stage), by using a multivariate model, we also observed a significantly higher hazard of melanoma death for patients in the 'low immune' subgroups compared to 'high immune' subgroups (Table 4.2) predicted by the max pooling MIL, gated attention MIL, TransMIL and DSMIL models. Overall the TransMIL model, had the greatest prognostic ability for stratifying patients when adjusting for covariates, with a HR = 2.27 and $P < 0.005$ (95% CI 1.52–3.45). This exceeds the HR value, when using the ground truth subgroup labels to stratify patients using a multivariate model (HR = 1.72 [95% CI 1.15, 2.63]), as shown in Table 4.2, suggesting that image features could have a stronger association with survival, than the immune subgroup features derived from the transcriptomic data.

Moreover, we wanted to examine how MIL models performed when using the whole dataset, rather than just using the 'high' and 'low immune' cases within the test set. We believed it was important to do this, as the 'intermediate immune' subgroup represents around 40% of melanoma patient population, therefore, would represent a large proportion of patients in a real-world dataset. Consequently, being able to stratify these patients based on prognosis is still an important challenge, even if we are unable to classify these patients into the 'intermediate

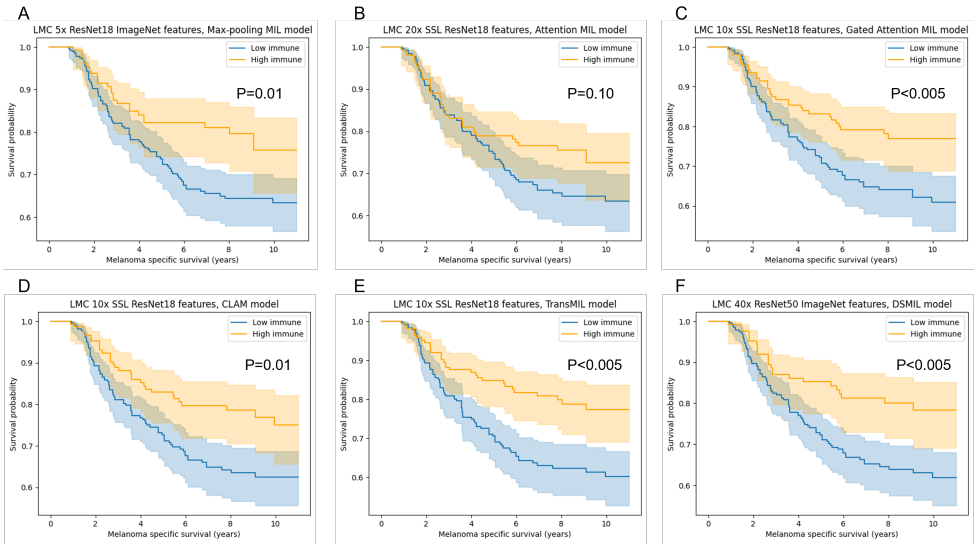


Figure 4.13: Prognostic associations of the 'high' and 'low immune' subtypes from the LMC dataset ($N = 230$), using labels generated from (A) max pooling MIL model trained with ResNet18 ImageNet embeddings from 5x patches, (B) attention MIL model trained with SSL ResNet18 feature embeddings from 20x patches, (C) gated attention MIL model trained with SSL ResNet18 feature embeddings from 10x patches, (D) CLAM model trained with SSL ResNet18 feature embeddings from 10x patches, (E) TransMIL model trained with SSL ResNet18 feature embeddings from 10x patches, (F) DSMIL model trained with ResNet50 ImageNet feature embeddings from 40x patches. Kaplan–Meier estimator with 95% CI was used to estimate the survival probability, and pairwise log-rank test was used to test the significance between the subgroups. CI, confidence intervals; LMC, Leeds Melanoma Cohort; MIL, multiple instance learning; SSL, self-supervised learning.

immune' subgroup based on imaging data. For our experiments, this meant including the 'intermediate' population in the test set and observing how a subset of the best performing MIL models trained using only 'high' and 'low immune' cases classified them in the binary task. As we show in Fig. 4.14D-F, we tested the prognostic ability of the three models that had the highest AUC performance when classifying the 'high' and 'low immune' subtypes, using a Kaplan–Meier estimator with 95% CI to estimate the survival probability. We also show in Table 4.3, the HR with 95% CI for these models when carrying out uni-variable analysis and multi-variable analysis with adjusting for clinical predictors (age, sex, tumour site, AJCC stage).

From Table 4.3, it appears that the model using 40x ResNet50 features is able to stratify all patients in the test set, including the 'intermediate subgroup', into 'high' and 'low immune' subgroups, with the highest HR (1.64 [95% CI 1.16, 2.33]) and C-index (0.72 [95% CI 0.685, 0.757]) when using multi-variate analysis. While this could indicate that the model is able to delineate image-based features that correlate with survival in the 'intermediate' subgroup, it could also be reflective of the model's bias towards the 'low immune' majority class, as this

Table 4.2: The prognostic value of the 'high immune' and 'low immune' subtypes in univariable and multivariable analyses from the LMC dataset, when using test set labels (N = 230), derived from the highest performing MIL models. HR and 95% CI for MSS are shown with P values. A HR > 1 implicates a higher risk of death in the predicted 'low immune' compared to the predicted 'high immune' subgroup of patients. P values < 0.05 are indicated by * and P values < 0.005 are indicated by **. The first row shows the HR for the ground truth labels in the test set. The model results that are shown are a max pooling MIL model trained with ResNet18 ImageNet embeddings from 5x patches, an attention MIL model trained with SSL ResNet18 feature embeddings from 20x patches, a gated attention MIL model trained with SSL ResNet18 feature embeddings from 10x patches, a CLAM model trained with SSL ResNet18 feature embeddings from 10x patches, a TransMIL model trained with SSL ResNet18 feature embeddings from 10x patches and a DSMIL model trained with ResNet50 ImageNet feature embeddings from 40x patches. AJCC, American Joint committee on cancer; CI, confidence intervals; CLAM, Clustering-constrained Attention Multiple instance learning; DSMIL, Dual Stream multiple instance learning; HR Hazard, ratio; LMC, Leeds Melanoma Cohort; MSS, melanoma specific survival; MIL, multiple instance learning; SSL, self-supervised learning; TransMIL, Transformer based multiple instance learning.

Method	Characteristic adjusted	HR (95% CI)	P
Ground truth labels	–	1.85 (1.22, 2.78)	0.005**
	Age, sex, tumour site, AJCC stage	1.72 (1.15, 2.63)	0.01*
MIL	–	1.69 (1.20, 2.63)	0.02*
	Age, sex, tumour site, AJCC stage	1.88 (1.20, 2.94)	0.01*
Attention MIL	–	1.39 (0.94, 2.41)	0.10
	Age, sex, tumour site, AJCC stage	1.04 (0.70, 1.56)	0.26
Gated attention MIL	–	1.72 (1.15, 2.63)	0.01*
	Age, sex, tumour site, AJCC stage	1.79 (1.19, 2.70)	0.01*
CLAM	–	1.75 (1.16, 2.63)	0.01*
	Age, sex, tumour site, AJCC stage	1.49 (0.98, 2.27)	0.06
TransMIL	–	2.00 (1.33, 3.03)	< 0.005**
	Age, sex, tumour site, AJCC stage	**2.27 (1.52, 3.45)**	< 0.005**
DSMIL	–	1.92 (1.22, 3.03)	< 0.005**
	Age, sex, tumour site, AJCC stage	1.96 (1.25, 3.13)	< 0.005**

model predicted the largest proportion (69%) of the 'intermediate' test cases as 'low immune' compared to the other models (Fig. 4.14G-I). We believe this bias may be leading to improved stratification, as in the study by Pozniak *et al.* (2019) [1], the 'low immune' and 'intermediate immune' had overlapping survival distributions, therefore, the pathology agnostic model which is more likely to predict 'low immune' is more likely to better predict the survival distribution of the 'intermediate immune' cases if classifying them as 'low immune'. Surprisingly, the TransMIL model which had the best HR when stratifying 'high' vs 'low immune' cases (Table 4.2) had the worst HR (1.49 [95% CI 1.10, 2.04], Table 4.3) when adding the 'intermediate' cases to the test set. This could be due to the TransMIL model overfitting to the distinctive features of the binary 'high' and 'low immune' cases, rather than classifying patients based on a prognostic signature that is also present in the borderline or more heterogeneous cases. This appeared to be the

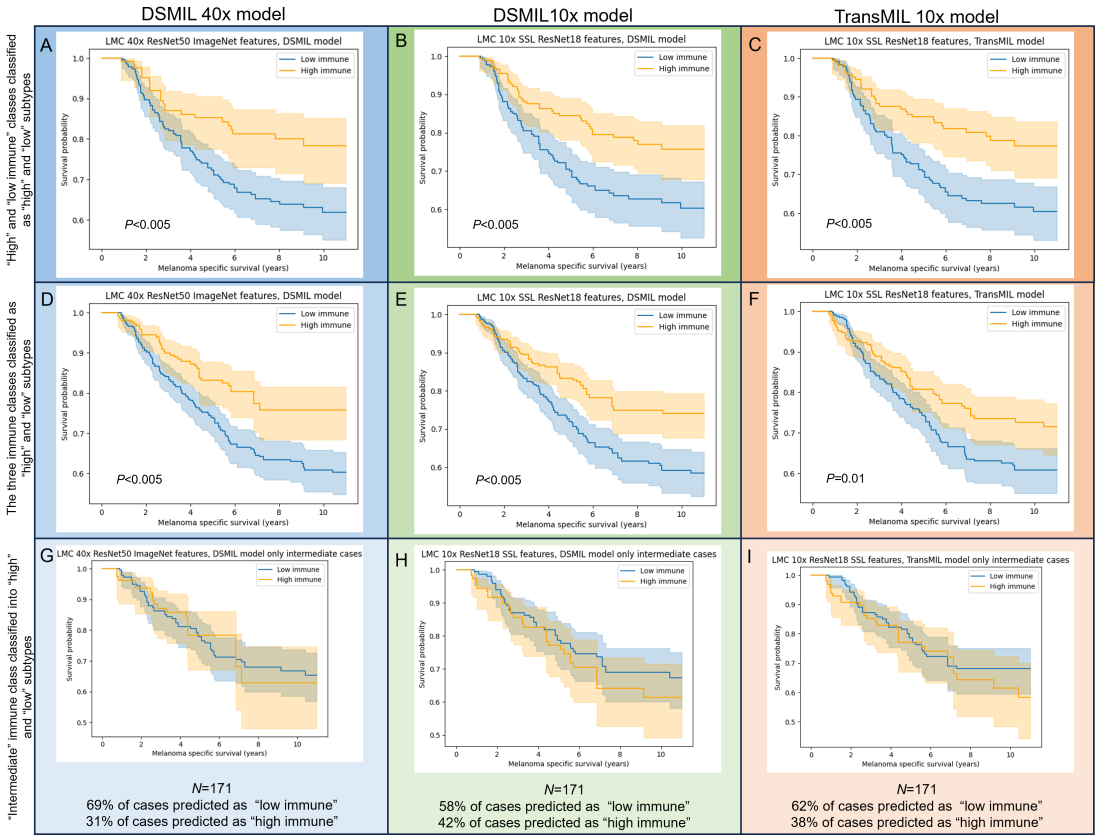


Figure 4.14: Prognostic associations of patients predicted as 'high' and 'low immune' subtypes from the LMC dataset ($N = 401$). The first row shows survival distributions of 'high' and 'low immune' patients predicted using labels generated from (A) a DSMIL model trained with ResNet50 ImageNet embeddings from 40x patches (DSMIL 40x), (B) a DSMIL model trained with SSL ResNet18 feature embeddings from 10x patches (DSMIL 10x) and (C) TransMIL model trained with SSL ResNet18 feature embeddings from 10x patches (TransMIL 10x). The second row shows survival distributions of the whole patient population including 'intermediate' immune cases classified as 'high' and 'low immune' subtypes by (D) DSMIL 40x , (E) DSMIL 10x and (F) TransMIL 10x models. The third row shows survival distributions of only the 'intermediate immune' cases classified as 'high' and 'low immune' subtypes by (D) DSMIL 40x, (E) DSMIL 10x and (F) TransMIL 10x models. Kaplan–Meier estimator with 95% CI was used to estimate the survival probability, and pairwise log-rank test was used to test the significance between the subgroups. CI, confidence intervals; DSMIL, Dual Stream multiple instance learning; LMC, Leeds Melanoma Cohort; MIL, multiple instance learning; SSL, self-supervised learning; TransMIL, Transformer based multiple instance learning.

case for all three high performing MIL models, with the 'intermediate immune' test cases that were stratified into 'high' and 'low immune' subgroups having overlapping survival distributions (as shown in Fig. 4.14G-I). This suggests that the MIL models trained with the binary 'high' and 'low immune' cases have limited generalisability to the intermediate patient population.

Table 4.3: The prognostic value of the 'high immune' and 'low immune' subtype classifications in univariable and multivariable analyses from the LMC dataset, when using test set labels (N = 401) for the whole datset (including 'intermediate immune' subtypes), derived from the highest performing MIL models. HR and 95% CI for MSS are shown with P values. A HR > 1 implicates a higher risk of death in the predicted 'low immune' compared to the predicted 'high immune' subgroup of patients. P values < 0.05 are indicated by * and P values < 0.005 are indicated by **. DSMIL 40x indicates a DSMIL model trained with ResNet50 ImageNet embeddings from 40x patches, DSMIL 10x indicates a DSMIL model trained with SSL ResNet18 feature embeddings from 10x patches and TransMIL 10x indicates a TransMIL model trained with SSL ResNet18 feature embeddings from 10x patches. C-index values for the multivariable models with 95% CI from 1000 iterations of bootstrapping are also shown. CI, confidence intervals; C-index, Concordance index; DSMIL, Dual Stream multiple instance learning; TransMIL, Transformer based multiple instance learning.

Model	Univariable HR (95% CI)	Multivariable HR (95% CI)	C-Index (95% CI)
DSMIL 40x	1.82 (1.22, 3.03)**	1.64 (1.16, 2.33)**	0.720 (0.685, 0.757)
DSMIL 10x	1.72 (1.28, 2.33)**	1.56 (1.34, 2.13)**	0.717 (0.680, 0.753)
TransMIL 10x	1.54 (1.14, 2.08)*	1.49 (1.10, 2.04)*	0.717 (0.680, 0.755)

Although these binary MIL models can still stratify all patients into prognostically significant subgroups (Fig. 4.14 and Table 4.3D-F), the 'intermediate immune' patient population may be more accurately represented as a continuum rather than a binary classification.

4.5 Discussion

4.5.1 Overview of findings

Recent studies [1, 84] have shown that melanoma patients can be stratified into subgroups, with added prognostic value compared to the current melanoma staging system [39]. However, these studies are carried out using transcriptomic data, which can be expensive and time consuming to analyse. Here we show that routinely used H&E images can be used to develop models that classify patients into these immune subgroups. We show that image-based MIL models can be developed to stratify patients into 'high' and 'low immune' subgroups in two independent datasets, with high performance. Moreover, by utilising MSS data, we show that MIL models, can stratify patients into prognostically significant groups with higher HRs than the original ground truth labels (Fig. 4.13 & Table 4.2). We also show feature inputs are important for improving model performance, in terms of both patch resolution and feature extraction methodology (Fig. 4.4 & 4.5). We highlight the importance of attention-based methods for improving model performance, with the attention MIL model achieving superior results for the TCGA dataset and the DSMIL model generating superior results for the LMC dataset. Nevertheless, we also acknowledge that using AUC to determine superior performance, does not take into account the effects of having an imbalanced dataset, therefore, we show how additional metrics are vital for further investigating model performance (Table. 4.1 & 4.2).

4.5.2 Challenges using transcriptomic ground truth labels

We outline a unique challenge associated with using ground truth labels derived from transcriptomic data. The immune subtype labels exhibit spatial bias, because they originate from a small region within the tumour where a 0.6-mm microarray needle was used to extract cores containing mRNA. This presents a challenge for our study, as MIL models utilise patches from the entirety of the segmented tissue, yet the ground truth labels come from a small region in the tumour tissue. This issue is further complicated by the highly heterogeneous nature of the melanoma landscape, in both primary and metastatic lesions. Moreover as mentioned by Nsengimana *et al.* (2018) [84], the transcriptomic sample can be contaminated by tissue outside of the tumour region, as might have been the case for the misclassification illustrated in Fig. 4.11. Here, it is possible that the sample was contaminated by nodal lymphocytes, leading to the ground truth label of 'high immune', when the tumour region in the WSI actually contained few TILs, leading to the image being misclassified by the attention MIL model. Nevertheless, our findings equally show that attention MIL models can overcome this in some cases, by focusing on regions that give contextual information. For example in Fig. 4.8 and 4.10, the 'high immune' 10x attention patches contain both tumour cells and immune cells. This indicates that the model hasn't solely focused on lymphocyte-rich regions outside of the tumour area, but may have utilised regions that contain tumor cells for context.

4.5.3 MIL models attend to regions with prognostic value

In 1989, Clark *et al.* [223] attempted to characterise TILs into categories based on their presence and positioning within primary melanomas, finding they had prognostic value. Subsequently, a meta-analysis by Sun *et al.* (2020) [224] corroborated that a 'brisk' TIL grade, where there is robust infiltration of TILs throughout the entire tumour or surrounding the tumour base, is associated with improved patient prognosis. As discussed in Chapter 2, despite the discernible prognostic value of TILs, they continue to be excluded from the current AJCC staging system due to lack of standardisation and interobserver variation during TIL grading [62]. Nevertheless, in 2023, a study by Chatziioannou *et al.* [130], demonstrated how deep learning can be harnessed to standardise the scoring of TILs in primary melanomas and provide a prognostic tool that is complementary to AJCC staging. Our results add to this evidence, as we show that MIL models attend to regions containing TILs in the correctly classified 'high immune' WSIs (Fig. 4.8 & 4.10), reinforcing this evidence that TILs have a prognostic signal.

In addition, in Table 4.2 and Fig. 4.13 we show that the TransMIL model, which was trained and tested with 10x SSL ResNet18 features, had the highest HR for stratifying the 'high' and 'low immune' subtypes. This suggests that the classifications from this model are based on stronger prognostic indicators than the other MIL models, as they lead to improved patient stratification. We believe that this improved prognostic ability could be from the combination of cellular context from 10x resolution SSL ResNet18 features and the introduction of

neighbourhood information through the model's self-attention mechanism. This neighbourhood information and lower resolution context together could be vital for capturing spatial relationships and configurations of TILs. Moreover, the higher HR of the TransMIL model compared to the original ground truth labels (Table 4.2) suggests that using MIL models with all the patches from a WSI, can improve patient stratification, due to reduced sensitivity to noise compared to the spatially biased ground truth labels.

4.5.4 The importance of feature inputs and model selection

A previous rigorous benchmarking study by Ghaffari Laleh *et al.* (2022) [108] examined the difference in performance of different MIL models for different cancer subtyping tasks. However, our study examines the often overlooked effects of how feature extraction methods and input patch resolution can impact the performance of six different MIL frameworks. Here we systematically show, for the task of immune subtyping of melanomas, that MIL models which use attention mechanisms have higher AUC performance than max pooling MIL models. Meanwhile, there does not appear to be a significant difference in mean test AUC performance between the attention-based MIL model methods, suggesting that more rigorous benchmarking studies are required to determine which MIL models are superior and should be used as appropriate baselines for cancer subtyping tasks.

Our results also indicate that, for both datasets and the majority of MIL models, 10x resolution input patches lead to the best performance when classifying 'high' and 'low immune' cases. This is an important finding, as resolving the optimum resolution used for prediction of molecular tumour biomarkers, remains an open question for many classification tasks with no clear consensus [225]. Moreover, many studies cite using 20x or 40x input patches to carry out their tasks, which may both increase processing and model training time and decrease model performance. Using high attention patches and immune attention heatmaps, we show the importance of tumoural context as well as cellular detail, and how 10x resolution patches provide a balance of lower-level immune cell features and higher-level tissue architecture. Nevertheless, we also concede that the SSL ResNet18 developed by Ciga *et al.* (2021) [125] was trained on images of 10x, 20x, 40x and 100x resolutions, which may have led to compromised performance when classifying 2.5x and 5x resolution images with the SSL feature extractor. In addition, melanoma tumours vary greatly in inter and intra-histological appearance, therefore pretraining with a dataset of melanoma histology slides could improve upon our results.

A surprising outcome from our study, was that the DSMIL model that had the highest mean AUC, for the classifying 'high' and 'low immune' cases from the LMC dataset, was trained using 40x resolution embeddings, from a modified ResNet50 pretrained on the ImageNet dataset. This result aligns with the general trend shown in Fig. 4.4, which showed increased performance as resolution increased for MIL models trained using LMC ResNet50 ImageNet embeddings. These findings also suggest that the modified ResNet50 feature extractor pretrained with ImageNet, can excel at capturing lower-level features in 40x patches, which may be more informative for

classification of melanoma WSIs, compared to using SSL with pathology-specific datasets. This result is further supported by the heatmaps shown in (Fig. 4.8 & 4.9), where the 40x DSMIL model is shown attending to prognostic areas within the tumour. Yet, we also saw in Fig. 4.9 and Fig. 4.12A that the model learned high attention scores for uninformative regions, such as blood vessels and keratin, suggesting this model is overfitting to noise caused by heavy staining or other artefacts in the image (as seen in Fig. 4.11 also). Furthermore, when comparing the F1 score and balanced accuracy of this model to the same model that had been trained using 10x pathology specific embeddings, we found that the F1 score, and balanced accuracy were superior for the model trained with 10x SSL ResNet18 features. This suggests that the DSMIL model, trained with 40x features from the modified ResNet50, has a bias towards 'low immune' cases and a poor ability to classify 'high immune' cases, which could be caused through mistaking TILs for melanocytes at the high resolution (Fig. 4.11), or making classification errors due to non-salient features (Fig. 4.9). Additionally, when examining the prognostic ability of the models, the TransMIL model which was trained and tested with 10x SSL ResNet18 features, had a higher HR for stratifying the 'high' and 'low immune' subtype (Table. 4.2). Here we determine that while the DSMIL model using 40x ResNet50 features had the highest mean test AUC performance, it shows weaker performance for classifying 'high immune' cases and stratifying patients based on MSS compared to models that use 10x SSL ResNet18 features.

The superiority of 10x SSL ResNet18 features for classifying the 'high' immune subtype in both datasets also supports the evidence [21, 210, 211, 212] that the feature extraction methodology can be important for capturing subtle differences in biological subtypes and improve classifier performance. Moreover, using a pathology-specific pretrained network addresses some of short-comings of MIL frameworks compared to classical weakly supervised models which are trained directly on images, while having the benefit of using a MIL model which focuses on tumour regions [108]. These results indicate that while using a MIL model with an attention mechanism is a major improvement over a standard max pooling MIL mechanism, the model inputs used are also equally important.

4.5.5 Limitations and future directions

While we show that MIL models can stratify all three patient populations into two subgroups with significantly different hazards of melanoma death (Fig. 4.14D-F & Table 4.3), deciphering the 'intermediate immune' subgroup remains a challenge. Although the 'high' and 'low immune' subgroups appear to have distinct image-based phenotypes (Fig. 4.8, 4.9 & 4.10), we believe that due to the tumour heterogeneity and complexity within the 'intermediate immune' subgroup, it is a harder task to delineate homogeneous features to represent this subset of patients. To tackle this problem we may need to look at further dividing this subgroup, as a previous study by Nsengimana *et al.* (2018) [84] found two distinct subgroups which overlap within this 'intermediate' group, or use techniques to learn more discriminant representations of the images, such as deep Fisher Discriminant analysis [226]. Moreover, it may be important to consider

how we frame this problem, as although we are treating it as a classification task, due to overlap of the groups, especially the 'low' and 'intermediate' subgroups, it could be more useful to look at it as a regression-based problem. A study by Nahhas *et al.* (2023) [110], showed how a weakly supervised regression method could be applied to predict multiple molecular biomarkers across nine cancer types. Conversely, implementing an ordinal loss function, which would rank the subtypes, penalising misclassification between 'high' and 'intermediate' or 'low' and 'intermediate' less, may prevent these errors as seen in Fig. 4.2.

We also believe that a multi-resolution approach, which incorporates multiple levels of cellular and tissue detail could further improve on our approach. We have seen that both 10x and 40x DSMIL models are able to correctly classify different cases, suggesting there is informative morphology at both resolution levels. Li *et al.* (2021) [177], showed how combining embeddings from multiple resolutions through concatenation led to improved performance of the DSMIL model and this has also been shown in other MIL models which we did not explore in this work [207, 227]. We also recognise that using a patch-based approach can prevent a model learning contextual information that is important in heterogeneous tissues. In future work, we aim to modify our approach using a graph neural network, by connecting adjacent patches and creating a hierarchy across resolutions, similar to Zormpas-Petridis *et al.* (2019) [173] when implementing cell-based melanoma graphs. In this way, we will also develop a tool that works in a similar way to a clinician, examining cellular detail and making diagnostic decisions based on the surrounding information from multiple magnifications, rather than an individual local patch-level.

Moreover while we were able to classify patients into significant groups with clinical outcomes in the LMC dataset (TransMIL model: Log rank $P < 0.005$, HR = 2.27, [95% CI 1.52, 3.45]), we were unable to replicate a significant result in the TCGA dataset. This could be due to insufficient data, as we had a reduced test dataset ($N = 84$) compared to the transcriptomic dataset used in the original paper ($N = 189$) [1], or due to the regions outside of the tumour, leading to misclassifications. However, the significant prognostic value when using the LMC trained TransMIL model, suggests there could be an even stronger signature if training a model with the imaging data and survival data alone. Therefore, in future works, we also hope to compare the prognostic value of the transcriptomic subgroups against de novo subgroups found through using the imaging biomarkers from the WSI image data.

4.6 Conclusion

To the best of our knowledge our work is one of the first extensive studies using MIL to subtype melanoma histopathology images into subgroups based on immune expression signatures with differential associated survival outcomes and potential treatment targets. Through survival analysis we show how MIL models can be used with clinical utility to stratify patients into prognostic groups. This is also one of the first studies to comprehensively show the importance

of resolution for context and pathology-specific feature extraction methods for improving MIL model performance. On the other hand we also highlight potential pitfalls to using transcriptomic ground truth labels and give examples of potential errors in image classification tasks. Overall we show how MIL models can be used as a tool to flag potential prognostic biomarkers and stratify patients into prognostic groups, without the need for additional genetic tests.

In the next chapter, we explore how computational graph representations of WSIs can be used to add context to classification models.

Chapter 5

Multi-level Graph Representation of Melanoma Whole Slide Images for Identifying Immune Subgroups

In the previous chapter we examined how multiple instance learning (MIL) can be applied to the task of identifying whole slide images (WSIs) into classes based on inferred immune cell presence within the tumour. In this chapter, we present our own method to improve on MIL, using novel graph representations of WSIs. Here, we aim to introduce further context to classifications beyond the individual patch-level, by using graph-representations of WSIs as our hypothesis is that using graph based methods could alleviate some of the limitations of MIL. For example, by adding more context to improve the representation of immune cells within the tumour microenvironment and reducing the number of false classifications caused by patches outside of the tumour region.

5.1 Introduction

As discussed in previous chapters, WSIs are multi-resolution gigapixel images, which means they can be computationally and memory intensive to process, especially on GPUs. Moreover, most WSIs only have slide-level labels, as pixel-wise labelling can be time consuming for a pathologist. In Chapter 4, we showed how MIL frameworks, where an image is treated as a bag with instances (patch embeddings) which inherit the bag or slide-level label [106], can been applied with high accuracy for classifying WSIs [109, 128, 178, 177]. A limitation of these MIL methods is their assumption that each patch-instance or patch-feature is independent of others, leading to

classifications based solely on local visual and morphological patterns within the patch region. Consequently, classifications rely on limited contextual information, which can be important as local cellular information is sometimes not enough in certain prediction tasks. For example, classifying and understanding the roles of immune cells within the tumour microenvironment (TME) requires long-range contextual information, including the spatial arrangement of cells, their location within the TME, and their interactions with other cell types. Moreover, as we demonstrated in Chapter 4 the ground truth labels for classifying the immune subgroups are derived from small 0.6-mm regions from the melanoma tumours, which means it is important that the visual features and patches are also from this region. Here, we found a trend in both TCGA and LMC datasets, where using 10x input patches which have a mixture of higher resolution cellular features and lower resolution tissue context, led to improved performance, suggesting models require contextual information for reduced misclassifications.

We also showed that Transformer based MIL (TransMIL) [178] and Dual Stream MIL (DSMIL) [177] approaches led to the highest mean AUC performances for the LMC dataset (0.79 and 0.80, respectively). These approaches both leverage attention-like mechanisms, which work by examining the relationship between a patch-instance with another patch-instance, with DSMIL comparing a critical instance with other instances within a WSI using a similarity metric and TransMIL using a self-attention mechanism which calculates the importance of each patch embedding to all other patch embeddings within the WSI on a pairwise basis. This suggests that determining the relationships between patch-instances within the WSI, whether that be through similarity (DSMIL) or spatial context (TransMIL), is important for improving classification accuracy when subtyping WSIs into immune subgroups. Meaning that the MIL assumption, which treats patches as independent and identically distributed, while being suitable for classifications involving fine-grained morphological detail, may not work as well for classification tasks based on the immune infiltrate in the TME. In this case, we need a WSI representation or model which is able to capture intra-entity relationships and provide the cellular context within the tumour.

Recent works have shown how graph neural networks (GNN) can be a powerful tool for WSI analysis and subtyping [24, 135, 228]. Graphs allow us to encode spatial and relational data, and have been applied to represent WSIs as cell-graphs [137, 138], patch-graphs [36, 133, 139], tissue-graphs [132, 136] and hierarchical cell-to-tissue graphs [134, 135]. However, to construct the WSI graphs, these methods use clustering or neighborhood similarity, which adds hyperparameters and inference time for running models, or use sampling techniques to select patches due to memory constraints which can lead to loss of the overall WSI structure.

Moreover, previous works [24, 34, 134, 135, 138, 229], have shown how hierarchical graphs can be used to aggregate multi-resolution structural information, improving model performance through incorporating contextual features. Hou *et al.* [133], demonstrated how multi-scale WSI graph representations constructed from 'Thumbnail', $\times 5$ patch and $\times 10$ patch embeddings, could be used for grading and subtyping esophageal cancer and kidney cancer datasets from

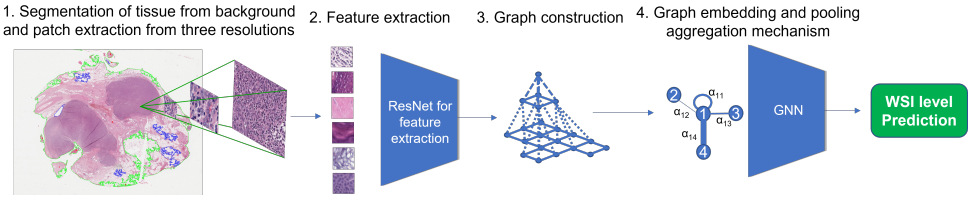


Figure 5.1: Experimental framework for classifying the WSIs, using graph representations.

the TCGA using an Iterative Hierarchical Pooling module. Here we build on this idea, by presenting one-hot encodings for node and edge features between different resolutions that we hypothesis could generate a more context-aware approach.

In this Chapter, we present a novel patch-based graph method, that exploits the intrinsic spatial positioning of all patches within a WSI and is also memory efficient, using patch-level feature embeddings from a CNN. In addition, we take inspiration from hierarchical graphs [24, 34, 135, 138, 133], developing a multi-resolution approach, which exploits the inherent hierarchical relationships between features extracted from patches at different resolutions. WSIs have a pyrimidal structure, where multiple versions of the image are represented at different resolutions, with the bottom image having the highest resolution. This allows a pathologist to examine cellular and tissue pathology at varying levels of detail. We try to capture this structure, by developing multi-resolution WSI graphs, using node and edge attributes to encode multi-scale relationships between patches. Moreover, following the results outlined in Chapter 4, we leverage an attention mechanism in our aggregation function, to enhance how informative features contribute to slide-level classifications. We present results demonstrating that our graph-based approach increases the contextual information learned by the models for classifying melanomas into immune subgroups and demonstrate how this leads to improved classification performance over state-of-the-art (SOTA) MIL methods.

5.1.1 Contributions

In this chapter, we demonstrate how GNNs and graph representations of WSIs can be used to classify melanomas into the subgroups delineated by Pozniak *et al.* (2019)[1].

Our contributions are as follows: (1) We show that GNNs are superior compared to MIL models at classifying immune subgroups. (2) We demonstrate that multiple GATv2 layers contribute to enhanced information flow, when using larger graph inputs. (3) We show that GNNs and attention GNN models are able to stratify LMC patients into prognostic groups. (4) We reiterate the importance of attention mechanisms, as the highest mean AUC performance for classifying 'high' and 'low immune' classes, was generated by a model with an attention mechanism.

Table 5.1: One-hot encoding dictionary for the edge features in symmetric and asymmetric multi-level graphs.

Edge feature connection	One-hot encoded vector	
	Symmetric	Asymmetric
10x ↔ 10x	[1,0,0,0,0]	[1,0,0,0,0,0,0]
10x → 20x	[0,1,0,0,0]	[0,1,0,0,0,0,0]
10x ← 20x		[0,0,1,0,0,0,0]
20x ↔ 20x	[0,0,1,0,0]	[0,0,0,1,0,0,0]
20x → 40x	[0,0,0,1,0]	[0,0,0,0,1,0,0]
20x ← 40x		[0,0,0,0,0,1,0]
40x ↔ 40x	[0,0,0,0,1]	[0,0,0,0,0,0,1]

5.2 Methods

5.2.1 Segmentation and feature extraction

For these experiments we used 667 digitised images from the LMC dataset and the corresponding immune subgroup labels (as described in Chapter 3, Table 3.1). The H&E tissue in the WSI was segmented from the background using the thresholding and morphological operations, according to a preprocessing protocol designed by Lu *et al.* [109] and described in Chapter 3. The tissue from the segmented images was then split into 256 pixel × 256 pixel non-overlapping patches at three different resolutions. We chose to carry out experiments using 10x, 20x and 40x, as these resolutions had generated the best performance for the MIL models in the previous chapter. Moreover we chose to use the 512-vector features from the ResNet18 which had been pretrained by self-supervised learning on multiple histopathology datasets [125], as we saw in the previous Chapter that these features generated best overall classification performance when using multiple metrics (Table 4.1).

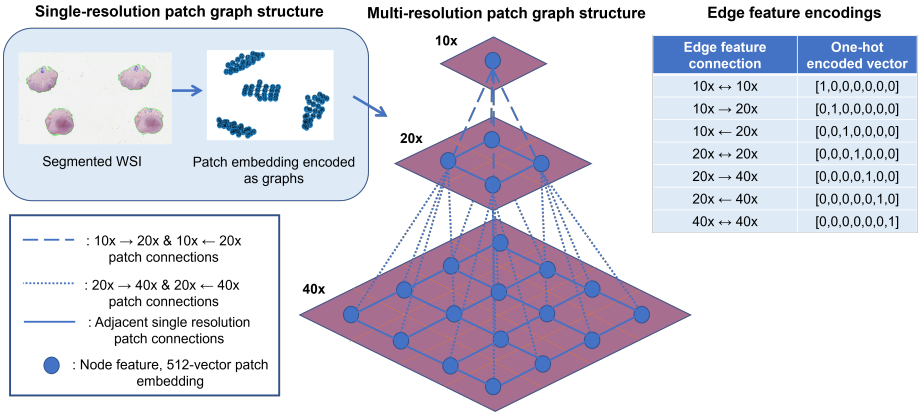


Figure 5.2: Single and multi-resolution patch graph representation construction and edge feature encodings used for multi-resolution graphs. Self-loops are not visualised.

5.2.2 Graph construction

We constructed global graph representations $G = (V,E)$ of the WSIs, where G is the graph, $V = \{1,...,n\}$ is the set of n vertices or nodes, which correspond to the n patch instances in a WSI. $E \subseteq V \times V$ is the set of edges, where $(i, j) \subseteq E$ is an edge that connects node i to node j. Node features for the graphs consisted of the 512-dimensional feature embeddings extracted from the patches. For experiments using single-resolution graph features, edge connections were defined as being between any neighbouring patch and a patch that is to the left, right, above, or below (diagonally neighbouring patches were not included). To represent uniform edges between adjacent patches in single-resolution graph representations, edge features were set to 1. We also wanted to build what we term 'multi-level' graph representations, where node features consisted of the extracted features from each of the three resolutions, and edges included those between adjacent patches within a single resolution and between patches from different resolutions. This involved defining edge connections between 10x patch nodes and the four spatially corresponding 20x patch nodes in the higher resolution level. Additionally, edge connections are defined between the 20x patch nodes to the four spatially corresponding 40x patch nodes to create hierarchical multi-resolution patch graphs. We also experimented with adding one-hot encodings to the node embeddings to distinguish between nodes of different resolutions (10x: [1,0,0], 20x: [0,1,0] and 40x: [0,0,1]). Here, binary vector representation was used to align with the scale of the existing pre-trained ResNet18 features, which are within the range of [-1,1]. The resolutions are represented by a vector of 0s and 1s and the unique positioning of the 1 indicates the exact resolution. These encodings were concatenated to the node feature embeddings giving a 515-dimensional embedding. All nodes were connected in both directions, and self-loops were included. To distinguish between edges that connect patches from different resolutions and edges that connect adjacent patches within the same resolution, edge features were encoded using a one-hot encoding scheme (Fig. 5.2 & Table 5.1). Moreover, for the ablation studies carried out, the 'one-hot node', 'asymmetric' and '10x-20x' multi-level representations all use the one-hot encoded vectors shown in the asymmetric column for edge features shown in Table 5.1, whereas the 'symmetric' graph representations edge features use the vectors shown in the symmetric column (Table 5.1). More detailed depictions of the multi-level graph structures used in the ablation studies are shown in Fig. 5.3.

5.2.3 Model architecture

When implementing the GNNs, we used *GATv2Conv* (GATv2) graph attention layers with a single attention head (default configuration from PyTorch Geometric) [141]. Node features, edge indices and edge features were passed through these GATv2 layers. Within the layers, a scoring function $e : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ was used to evaluate the importance of each edge (j,i) and the importance between the features of the neighbour node j to node i. The attention coefficients were formulated using:

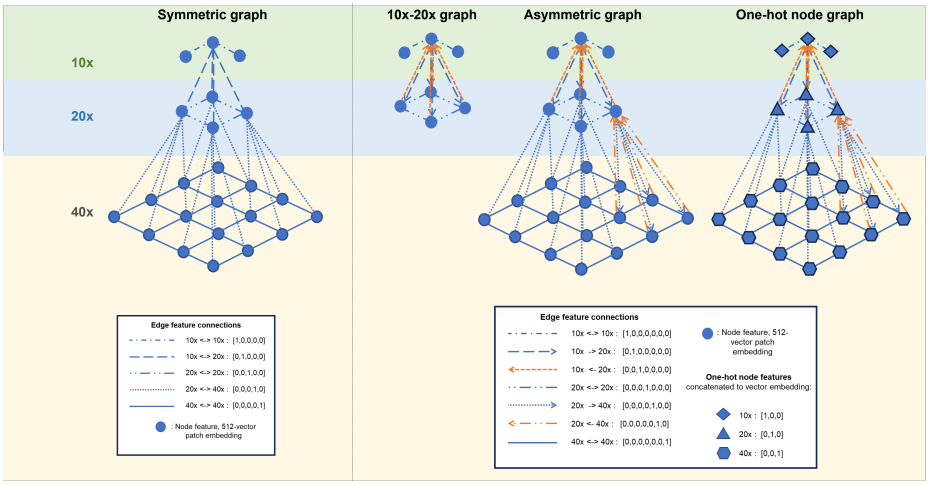


Figure 5.3: Types of multi-level graph structures, that were used in the ablation studies.

$$e(\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{e}_{ij}) = LeakyReLU((\boldsymbol{a} \odot \boldsymbol{W})^T [\boldsymbol{h}_i || \boldsymbol{h}_j || \boldsymbol{e}_{ij}]) \tag{5.1}$$

Where $\boldsymbol{h}_i$ is a node representation, $\boldsymbol{h}_j$ is a neighbouring node and $\boldsymbol{e}_{ij}$ represents the edge features between them. $\mathbf{W}$ represents the weights matrix, $\mathbf{a}$ represents learnable attention weights, $\odot$ denotes element-wise multiplication and $\|$ denotes vector concatenation. The attention scores are then scaled for all nodes within an image $j \in \mathcal{N}_i$, using a *softmax* function in the attention mechanism:

$$softmax_j(\boldsymbol{a}_{ij}) = \frac{exp(e(\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{e}_{ij}))}{\sum_{k \in \mathcal{N}_i} exp(e(\boldsymbol{h}_i, \boldsymbol{h}_k, \boldsymbol{e}_{ik}))} \tag{5.2}$$

Where $\mathcal{N}_i$ represents the set of neighbours of node $\boldsymbol{i}$ and $\boldsymbol{a_{ij}}$ is the attention coefficient between node $\boldsymbol{i}$ and node $\boldsymbol{j}$. $\boldsymbol{k}$ represents a node index from the set of all neighbors $\mathcal{N}_i$ of node $\boldsymbol{i}$. The output of the GATv2 layer is given by the equation:

$$\boldsymbol{h_i'} = \sigma \left(\sum_{j \in \mathcal{N}_i} e(\boldsymbol{h}_i, \boldsymbol{h}_i, \boldsymbol{e}_{ij}) \cdot (\mathbf{W} \odot \mathbf{a}) h_j \right) \tag{5.3}$$

Where $\boldsymbol{h_i'}$ represents the updated representation of node $\boldsymbol{i}$ after the GATv2 layer operation. σ represents an activation function, for our model each GATv2 layer we used the Exponential Linear Unit (ELU) function [230] and followed this with global mean pooling operation. The number of GATv2 layers was varied to assess their impact on performance. Initially to get the classification output, the outputs from each global mean pooling operation, when experimenting with multiple GATv2 layers, were then concatenated together and passed through two fully connected layers and a *softmax* activation function (Fig. 5.4).

In addition, to mitigate for uninformative node features contributing to the slide level label,

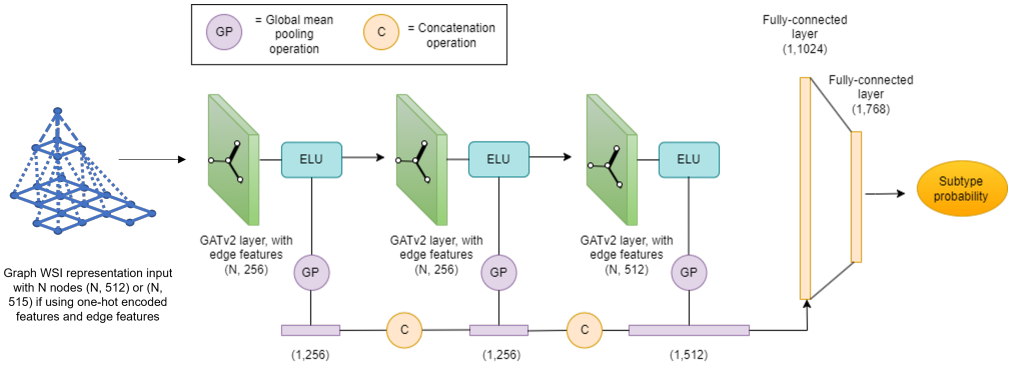


Figure 5.4: Network architecture for a three GATv2 layer graph neural network.

experiments were carried out implementing an attention mechanism to learn soft attention weights for the node embeddings. This involved using an additional GATv2 layer, to compute unnormalised attention scores for each node embedding, which are then normalised using a *softmax* operation. The node embeddings from the final GATv2 layer (excluding the GATv2 layer used to calculate the attention weights) are then updated using weighted aggregation of node embeddings with the attention weights. Two fully-connected layers followed by a *softmax* are then used to get a subtype probability output.

The GNN model performance was compared to the performance of the six weakly supervised MIL methods described in the previous chapter, including the max-pooling MIL model, the gated-attention MIL model (gated AttMIL) [128], the attention MIL (AttMIL) [128] and CLAM [109], DSMIL and TransMIL models. The MIL models shown in this chapter, all used the SSL ResNet18 patch embeddings and were trained as described in Chapter 4.

5.2.4 Model training

The networks were trained using a cross-entropy loss function, comparing the slide label with the predicted slide-level label. After experimenting with multiple learning rates ($2 \times 10e^{-3}$, $2 \times 10e^{-4}$, $2 \times 10e^{-5}$), we applied a learning rate of $2 \times 10e^{-4}$ for GNNs without an attention mechanism and a learning rate of $2 \times 10e^{-5}$ was applied for GNN models with an attention mechanism due to improved model convergence. A weight decay of $1 \times 10e^{-5}$ was applied to all models. Model parameters were updated and optimised using an Adam optimizer [231]. Dropout with a probability of 0.5 was used after each global mean pooling layer for the models without an attention mechanism. During training, to mitigate class imbalances between subtypes within the training data, a slide was sampled proportionally to the inverse of the frequency of its ground truth class. Model performance was evaluated using a 10-fold Monte Carlo cross validation approach to calculate the mean area under the curve (AUC) with 95% confidence intervals (CI). For each fold the data was split with 80% of the data being used for training data and 10% being kept for both the test and validation datasets. When calculating

Table 5.2: Mean test AUC for classification of three immune subtypes with 10 fold cross validation with 95% CI (mean ± 95% CI), for different MIL models compared to our proposed GNN models using single and multi-resolution graph respresentations. The GNN models here are implemented with 3 GATv2 layers.

Experiment	Method	Resolution		
		10x	20x	40x
SOTA MIL models	Max-pool MIL	$0.57_{\pm 0.022}$	$0.55_{\pm 0.026}$	$0.48_{\pm 0.023}$
	AttMIL	$0.61_{\pm 0.039}$	$0.59_{\pm 0.027}$	$0.54_{\pm 0.023}$
	Gated AttMIL	$0.60_{\pm 0.026}$	$0.57_{\pm 0.035}$	$0.55_{\pm 0.036}$
	CLAM	$0.60_{\pm 0.028}$	$0.57_{\pm 0.034}$	$0.55_{\pm 0.035}$
	TransMIL	$0.56_{\pm 0.028}$	$0.59_{\pm 0.037}$	$0.51_{\pm 0.029}$
	DSMIL	$0.59_{\pm 0.020}$	$0.60_{\pm 0.025}$	$0.60_{\pm 0.026}$
Proposed GNNs	Single-resolution	$\mathbf{0.63_{\pm 0.023}}$	$0.62_{\pm 0.021}$	$0.61_{\pm 0.017}$
	Multi-level		$0.62_{\pm 0.017}$	

performance for the three immune subgroups, the AUC scores were calculated for individual classes by binarising classifications, then averaging the AUCs for the three classes. The models without attention mechanisms were trained for a minimum of 50 epochs, with early stopping if the validation loss did not improve for 20 epochs continuously. Due to slower convergence of the models with an attention compared to those without the mechanism, we increased the minimum epoch stopping time for the models to 300 epochs.

Graph models were implemented using PyTorch 1.13.1 [232], PyTorch geometric version 2.2.0 [233] with CUDA 11.6 [234], using one Nvidia V100 32GB GPU from the JADE2 HPC facility. Segmentation, feature extraction and MIL models were implemented using PyTorch version 1.7.1 [232] with CUDA version 11.0 [234], using one Tesla V100 32G NVLink 2.0 GPU from the Bede N8 HPC facility.

5.3 Results

5.3.1 GNN models outperform MIL models

For classifying the three immune subtypes, the combination of graph and visual features led to improved performance over MIL models, when using 10x and 20x single resolution graphs and multi-resolution graphs with one-hot encoded node embeddings (multi-level) as input (Table 5.2). With the best performing model achieving a mean AUC of 0.63 (95% CI 0.61 to 0.65), using a 10x resolution input patch to get node and edge embeddings. However, as we saw with the MIL models in Chapter 4, the 'intermediate' subgroup presents a challenge. This is demonstrated by Fig. 5.5, as we show that the majority of misclassifcations are between the 'low immune' and 'intermediate immune' cases and the 'high immune' and 'intermediate immune' cases, with 66% of the 'intermediate' cases being misclassified.

Due to the heterogeneity and complexity within the 'intermediate immune' subgroup, we

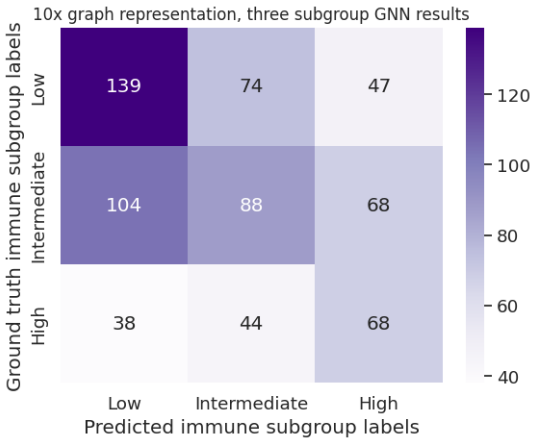


Figure 5.5: Confusion matrix for the GNN model prediction of the 'low', 'intermediate' and 'high immune' cases on the 10-fold test sets. The predictions are from the GNN model which achieved the highest AUC performance, which utilises 10x single resolution graph representations.

decided to simplify the task to assess the performance of GNN models when training and testing the models using only the 'high' and 'low immune' as these groups are more well defined compared to the 'intermediate' subgroup. Furthermore the 'high' and 'low immune' are more informative for immunotherapy options, so it is still valuable to predict these classes. We found using models trained with both 20x single resolution graphs and multi-resolution graphs outperformed current state-of-the-art MIL models (Table 5.3). Here, we found that the best performing GNN models both achieved a mean AUC of 0.80 and were trained using 20x patch WSI graph representations, or multi-level WSI graph representations which included one-hot encodings to represent different resolution patches in the embeddings.

Notably, for the three and two immune subtype classification tasks, the types of graph representations that generated the best performance were not consistent (Table 5.3 & 5.4). For the three immune subtype classification task, the 10x patch graph (mean test AUC of 0.63 [95% CI 0.61 to 0.65]), followed by the 20x and multi-level graph representations (mean AUC of 0.62), generated the best test mean AUC performances (Table 5.3). We believe this could be due to 10x patch feature nodes containing a balance of cellular and structural detail, but also due to 10x graphs containing less nodes. As melanoma whole slide images contain highly heterogeneous tissues, this can lead to increased node hetererogeneity and increased noise being introduced to classifications leading to errors. This heterogeneity is also increased when including the 'intermediate' subtype in the classification task, as this subgroup is less defined than the 'high' and 'low immune' subtypes. Therefore, lower resolution graph representations will contain less noisy or heterogeneous nodes and may generate better classification performance. Meanwhile, for the 'high' v.s. 'low immune' subtyping task, the 20x and one-hot node graph representations generated the joint highest mean AUC performance (Table 5.4). For this task the presence of immune cells in the TME is important for classification of the 'high immune' class, therefore,

Table 5.3: Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for different MIL models compared to our proposed GNN models. Models were tested at different using single and multi-resolution graph representations. The 10x and 20x single resolution models utilised three GATv2 layers, the 40x uses five GATv2 layers, the the model trained with multi-level representations utilised four GATv2 layers.

Experiment	Method	Resolution		
		10x	20x	40x
SOTA MIL models	Max-pool MIL	$0.51_{\pm 0.069}$	$0.57_{\pm 0.090}$	$0.48_{\pm 0.049}$
	AttMIL	$0.75_{\pm 0.043}$	$0.74_{\pm 0.048}$	$0.76_{\pm 0.029}$
	Gated AttMIL	$0.77_{\pm 0.042}$	$0.75_{\pm 0.072}$	$0.70_{\pm 0.048}$
	CLAM	$0.77_{\pm 0.043}$	$0.75_{\pm 0.053}$	$0.67_{\pm 0.073}$
	TransMIL	$0.79_{\pm 0.050}$	$0.72_{\pm 0.052}$	$0.64_{\pm 0.058}$
	DSMIL	$0.79_{\pm 0.046}$	$0.70_{\pm 0.059}$	$0.73_{\pm 0.059}$
Proposed GNNs	Single-resolution	$0.78_{\pm 0.033}$	$\mathbf{0.80_{\pm 0.052}}$	$0.79_{\pm 0.077}$
	Multi-level		$\mathbf{0.80_{\pm 0.048}}$	

cellular and spatial detail is important. As graph representations provide spatial context and higher resolution patches provide cellular detail, a trade-off between sufficient cellular detail and avoiding an increase of uninformative patch node embeddings at higher resolutions, is required. Hence, in Table 5.4 for single-resolution graphs, the most likely explanation for 20x patch graphs generating the best performance, is that they provide more cellular detail than 10x patches for 'high' v.s. 'low immune' task, without being adversely affected by the noise seen with the 40x patch graph performance. Conversely, the 'one-hot node' multi-level graph representations, whilst having the disadvantage of more noisy and heterogeneous nodes, also have additional multi-resolution with structural information from the edge and node one-hot encodings, which may be the reason that they generate the joint highest mean AUC when used as an input (0.80).

Although the highest performing model differed when using single resolution graphs and multi-reolution graphs, there appeared to be a greater affect on performance when changing the resolution of inputs for using MIL models rather than GNNs. For example in Table 5.3, the TransMIL model mean AUC performance decreases from 0.79, to 0.72 to 0.64 as the input patch resolution is increased from 10x to 20x to 40x, whereas for the GNN model it remains consistently high with mean AUC performance of 0.78-0.80, depending on input resolution. This suggests graph models may be more robust to changes in input resolution, whereas when classifying melanoma patients into immune subgroups 10x resolution patches to consistently generate the best performances in MIL models (Table 5.3). This could be due to GNNs ability to encode structural relationships and model global and local context, so although 40x or 20x patch features alone may be less discriminative than 10x for immune subtyping, the connections learned by the GNN between the higher resolution instances may still preserve critical contextual information from the tissue that could be lost by a MIL model that independently aggregates features.

Table 5.4: Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for GNNs when increasing the number of GATv2 layers in the network. 'One-hot node' indicates multi-level graphs with one-hot encodings added to the node patch embeddings.

Experiment	Graph representation	Number of GATv2 layers				
		1	2	3	4	5
Single Resolution	10x	$0.76_{\pm 0.050}$	$0.76_{\pm 0.041}$	$0.78_{\pm 0.033}$	$0.78_{\pm 0.038}$	$0.78_{\pm 0.045}$
	20x	$0.73_{\pm 0.061}$	$0.78_{\pm 0.045}$	$\mathbf{0.80_{\pm 0.052}}$	$0.78_{\pm 0.068}$	$0.79_{\pm 0.076}$
	40x	$0.68_{\pm 0.061}$	$0.75_{\pm 0.050}$	$0.75_{\pm 0.084}$	$0.77_{\pm 0.056}$	$0.79_{\pm 0.077}$
Multi-level ablation studies	10x-20x	$0.73_{\pm 0.056}$	$0.76_{\pm 0.042}$	$0.79_{\pm 0.033}$	$0.77_{\pm 0.044}$	$0.77_{\pm 0.045}$
	Symmetric	$0.73_{\pm 0.071}$	$0.77_{\pm 0.052}$	$0.78_{\pm 0.049}$	$0.79_{\pm 0.053}$	$0.79_{\pm 0.072}$
	Asymmetric	$0.77_{\pm 0.091}$	$0.78_{\pm 0.042}$	$0.77_{\pm 0.088}$	$0.77_{\pm 0.088}$	$0.79_{\pm 0.046}$
	One-hot node	$0.73_{\pm 0.065}$	$0.75_{\pm 0.077}$	$0.78_{\pm 0.076}$	$\mathbf{0.80_{\pm 0.048}}$	$0.79_{\pm 0.081}$

5.3.2 Increasing GATv2 layers improves performance for models trained with larger graph representations

We also tested how increasing the number of GATv2 layers affected model performance. Additionally we carried out several ablation studies, exploring different configurations of the multi-level representations. This included using only the 10x and 20x node embeddings to generate graph representations, to test whether the noise from 40x features were leading to a decrease in performance (referred to as '10x-20x' in Table 5.4). Moreover we wanted to see how different edge encodings influenced model performance, so as well as using the 'Asymmetric' edge feature connections described in the methods section, we also used simplified 'Symmetric' edge feature encodings (Table 5.1). These simplified encodings only encoded whether the connections were between two different resolution levels, not the direction of the encoding, for example whether the edge is between a lower and higher resolution node embedding, or higher and lower resolution node embedding.

We found that overall, having three or more GATv2 layers produced the best performance for models trained with single and mutli-resolution graph representations. Three layers generated the best AUC performance for the models trained with single (20x) resolution graph inputs, whereas the model with four GATv2 layers, using a multi-resolution one-hot encoded node representation, achieved the equally highest mean test AUC score of 0.80 (95% CI 0.75 to 0.85). When further increasing the number of GATv2 layers to five, we did not see a further increase in performance for the highest performing models, but found there was a convergence in performance at this point for most models.

From the multi-level ablation studies, the model using 'Asymmetric' graph representations had the best performance when there was only one GATv2 layer, and the performance of models using asymmetric inputs remained consistently high as the number of GATv2 layers were increased (Table 5.4). Suggesting models using multi-level graphs with asymmetric edge features, do not require the additional GATv2 layers to learn improved WSIs representations for

classification of the WSIs. Conversely, both the 'One-hot node' and 'Symmetric' models showed increases in AUC classification performance as the number of GATv2 layers were increased to four layers. This was also the case for models trained with 40x graph representations, suggesting these larger graph structures, containing 40x node embeddings, benefited more from increased message passing layers for long-range information transfer. Whereas models trained using 10x and 20x single resolution WSI graph representations had peaks in performances when using three GATv2 layers.

5.3.3 Attention mechanism generates the best AUC performance for model trained with 40x single resolution graph representations

To combat noise from uninformative nodes within the WSI graph representations, we implemented a learned attention mechanism, which learns soft attention weights for the node embeddings. In Table 5.5, we demonstrate that the mean test AUC performance for 10x and 20x single resolution graphs, when adding the attention mechanism, is mixed with both incremental increases and decreases in performance. However, when using 40x single resolution graphs as an input, there was a consistent increase in AUC performance for both three and four layer GATv2-layer models with an attention mechanism. Moreover, the three layer GATv2 attention GNN trained with 40x graph representations, generated the highest overall mean test AUC of 0.81 (95% CI 0.78, 0.84) for classifying the 'high' and 'low immune'. This increase in performance when using 40x inputs may be a consequence of these graph representations containing the largest number of nodes and subsequently more uninformative node embeddings. Thus adding an attention mechanism can counteract the adverse affect of these noisy instances. Nevertheless, if this were the case it would be expected that the multi-level graph representations would have improved performance, yet for the majority of models which were trained using multi-level representations, the addition of the attention mechanism led to a large decrease in AUC performance. This may in part be due to the mechanism adding additional parameters

Table 5.5: Mean test AUC for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI), for GNNs when using an attention mechanism instead of global mean pooling operation. 'One-hot node' indicates multi-resolution graphs with one-hot encodings added to the node patch embeddings.

Experiment	Graph representation	Number of GATv2 layers		
		3	4	5
Single resolution	10x	$0.78_{\pm 0.051}$	$0.80_{\pm 0.040}$	$0.78_{\pm 0.050}$
	20x	$0.78_{\pm 0.032}$	$0.78_{\pm 0.043}$	$0.75_{\pm 0.063}$
	40x	$\mathbf{0.81_{\pm 0.034}}$	$0.80_{\pm 0.034}$	$0.79_{\pm 0.040}$
Multi-level ablation studies	10x-20x	$0.79_{\pm 0.050}$	$0.74_{\pm 0.060}$	$0.73_{\pm 0.072}$
	Symmetric	$0.73_{\pm 0.064}$	$0.72_{\pm 0.026}$	$0.74_{\pm 0.034}$
	Asymmetric	$0.71_{\pm 0.046}$	$0.74_{\pm 0.038}$	$0.72_{\pm 0.031}$
	One-hot node	$0.73_{\pm 0.051}$	$0.72_{\pm 0.032}$	$0.76_{\pm 0.043}$

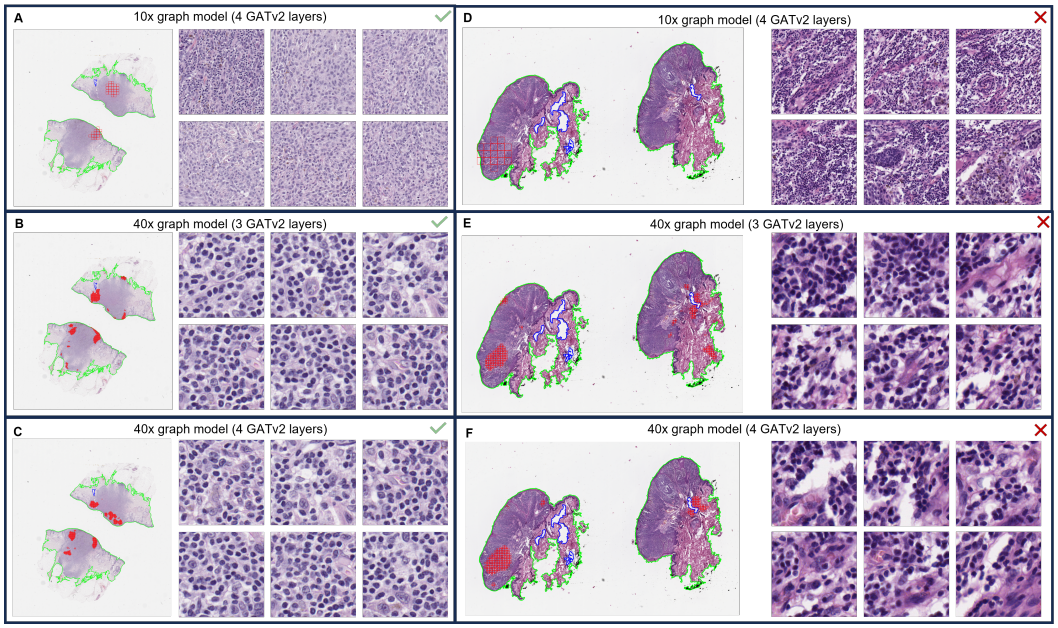


Figure 5.6: Correctly classified high immune images (A-C) and incorrectly classified low immune images (D-F) from the top three performing attention GNNs. H&E stained images with segmented tissue are shown annotated with the top 5% of patches which generated features with high attention scores. Next to the top six patches which generated high attention scores.

and complexity to the GNN models, causing overfitting and reduced performance. It is also possible that the mechanism we use is unable to learn the hierarchical relationships within the graph representations. For example, by aggregating all node embeddings together there may be dilution of the expressivity of the features at each resolution and loss of the the hierarchical structure of the graphs.

In Fig. 5.6, we examined where the regions of high attention in the highest performing attention GNN models in order to better understand how the resolution of the input patch features and number of GATv2 layers affected the learned attention scores. For the correctly classified high immune cases we show that the top 5% of high scoring immune attention features come from patches (highlighted in red Fig. 5.6A-C) within the darker stained tumour region. Interestingly the highest performing model, is the only model that learns high attention scores for features from patches that are deeper in the dermis towards the invasive front (Fig. 5.6 B), which could suggest that this model is utilising features that are indicative of the invasive properties of the tumour. The patches with high immune attention scores shown for the models trained with 40x features are florid with the darkly stained circular immune cells, which possibly also contain cells like histiocytes or granulocytes from pathologist interpretation. Whereas in Fig. 5.6A only the first high attention patch shown has a large volume of immune cell infiltrate, with fewer immune cells in the other patches, suggesting that the 10x model may not only be classifying images based on high proportions of TILs, but also tumour regions which contain

fewer TILs. In addition the top 5% of high attention patches are in different regions for all three models, which again highlights that different resolution attention GNN models may focus on different tumour features for classifying melanoma patients. To further understand which features could be contributing to model classifications, we also show the patches which generated the high attention features in the misclassified cases (Fig. 5.6.D-F). In Fig. 5.6D-F we show a 'low immune' case was misclassified as 'high immune' by all three models. The tissue in this case appears to have been over-stained, which as shown in the previous chapter can contribute to erroneous classifications (Chapter 4, Fig. 4.11), by obscuring chromatin patterns that can help to determine which cells are tumour cells. Over-stained malignant cells are more likely to resemble lymphocytes, appearing small, dark, and circular, hence the model may have classified this case as 'high immune' mistaking overstained tumour cells for immune cells. Additionally, melanocytes often imitate other cell types [199], which might lead the model to incorrectly learn that these features are immune cells. However, the pathologist examining these cases noted that while this case was categorized as 'low immune' and the over-staining means it is difficult to distinguish melanocytes from immune cells, the cells in the high-attention regions are TIL-like in appearance. As discussed in Chapter 4, if these cells are lymphocytes, this underscores the discrepancy between the spatially biased ground truth labels [1] and the heterogeneous tumor environment represented in the whole slide images.

In addition, we examined the outputs from the best performing multi-level attention GNNs (Fig. 5.5, mean AUC 0.76 [95% CI 0.72, 0.80]), as shown in Fig. 5.7. The top 5% of high attention patch features, are from a mixture of all three image resolutions. Unlike the single resolution attention GNN models shown in Fig. 5.6, the multi-level model was unable to correctly identify the high immune case shown in Fig. 5.7A, with the six high attention patches shown next to the image of the tissue including blurry patches and patches which are from outside the tumour region, suggesting this model is more impacted by out of focus regions compared to the single resolution models. Moreover, for Fig. 5.7A, there doesn't appear to be much overlap between the top 5% of high attention patches from different resolutions, which could also be contributing to an incorrect classification. Unlike the cases Fig. 5.6A-C, the multi-level model appears to be learning high attention weights for features from different scales from different regions, this added complexity could be responsible for the 'multi-level' model overfitting and worse performance compared to single resolution GNNs. The multi-level model, also misclassified the case with overstained tissue (Fig. 5.7). The six high attention patches in Fig. 5.7B, appear to be similar to the single resolution high attention patches (Fig. 5.6D-F) and contain darkly stained cells which could be lymphocytes, which could explain the 'high immune' classification, but as described in Chapter 4, heavy staining can cause difficulty in pathological interpretation [102]. In Fig. 5.7C we show a correctly classified 'high immune' case, although there are some 40x high attention patches in the brown regions which contain 'pigment-laden' macrophages (a type of immune cell), the top six high attention patches are from slightly blurry patches which contain small 'bland' tumour cells. It is possible that the

tumour cell features from these patches may have been mistaken as immune cell features by the model, as another correctly classified 'high immune' image shown in Appendix A.1, has high immune patches containing lymphocytes. However, the pathologist assessing the image and the high attention regions in Fig. 5.7C commented on the difficulty reviewing this case, due to out of focus regions and crushed tumour tissue, highlighting that the multi-level models may be less resilient to the noise of imaging artefacts as shown with the misclassification of the case shown in Fig. 5.7A compared to the correct classifications of models trained with single resolution graphs (Fig. 5.6A-C).

5.3.4 Additional metrics and prognostic associations of models

As we explored in the previous Chapter 4, while AUC is generally considered to be resilient to class imbalance within the data compared to other metrics like accuracy, it does not always provide a complete picture of a model's performance. In order to investigate whether the GNN models were biased towards the 'low immune' majority class, we calculated additional F1-score and balanced accuracy (BA) metrics for models that generated similarly high AUC values. Moreover, as the immune subgroups have prognostic ability, we wanted to measure

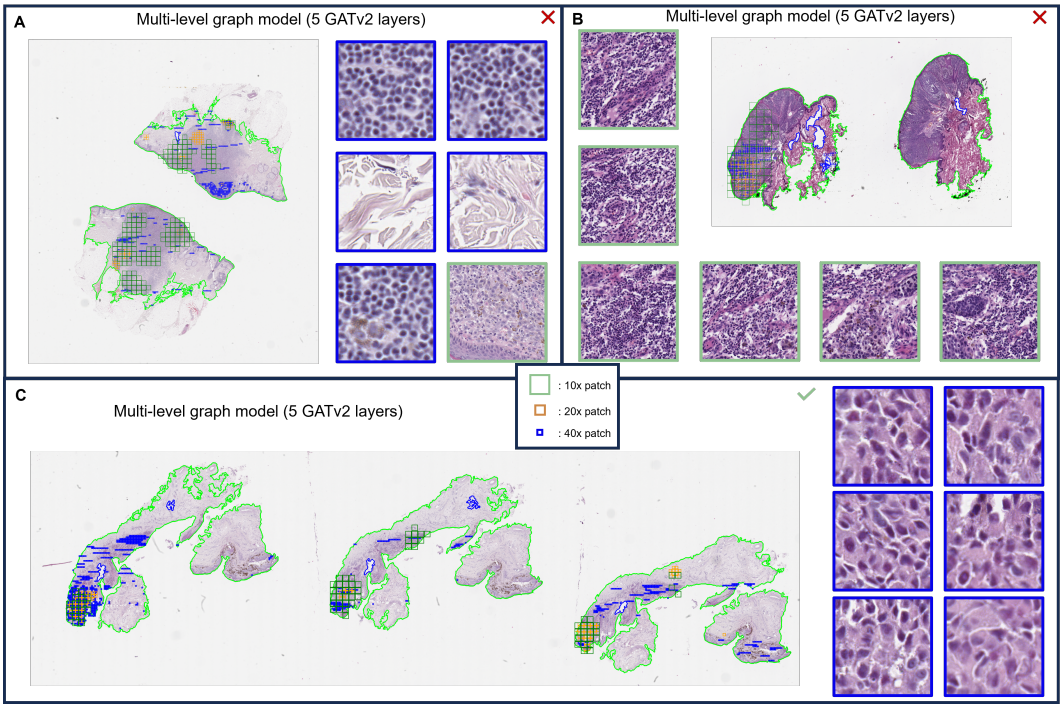


Figure 5.7: High attention regions from the multi-level attention GNN. H&E stained images with segmented tissue are shown annotated with the top 5% of patches which generated features with high attention scores. The images are shown next to the six patches which generated the highest attention scores.

Table 5.6: Mean test AUC, BA and F1-score for high v.s. low classification with 10 fold cross validation with 95% CI (mean ± 95% CI). The results shown are from GNNs with and without an attention mechanism and the number of GATv2 layers implemented in the models is shown in the GATv2 column. The metrics are also shown for the best performing MIL models, where the input column shows the type of input features, which were from a ResNet50 pretrained with ImageNet (ImageNet) or ResNet18 trained with SSL and histology images (Histo). The highest performance metrics are highlighted in bold.

Method	Input	GATv2	AUC	BA	F1 score
GNN	20x graph	3	$0.80_{\pm 0.052}$	$0.72_{\pm 0.059}$	$0.64_{\pm 0.075}$
models	Multi-level	4	$0.80_{\pm 0.048}$	$0.70_{\pm 0.061}$	$0.63_{\pm 0.073}$
GNN	40x graph	3	$\mathbf{0.81_{\pm 0.034}}$	$0.73_{\pm 0.039}$	$0.66_{\pm 0.046}$
attention	40x graph	4	$0.80_{\pm 0.034}$	$0.73_{\pm 0.043}$	$0.66_{\pm 0.051}$
models	10x graph	4	$0.80_{\pm 0.040}$	$\mathbf{0.74_{\pm 0.026}}$	$\mathbf{0.67_{\pm 0.033}}$
Method	Input	MIL model	AUC	BA	F1 score
SOTA	40x ImageNet	DSMIL	$0.80_{\pm 0.038}$	$0.71_{\pm 0.053}$	$0.61_{\pm 0.078}$
MIL	10x Histo	DSMIL	$0.79_{\pm 0.046}$	$\mathbf{0.74_{\pm 0.050}}$	$\mathbf{0.67_{\pm 0.056}}$
models	10x Histo	TransMIL	$0.79_{\pm 0.050}$	$0.73_{\pm 0.051}$	$0.66_{\pm 0.063}$

the prognostic associations of the 'high' and 'low immune' test set labels as predicted by the GNN models. A multivariate Cox proportional model was implemented to generate HRs for the predicted 'low' and 'high immune' test set patients, adjusting for clinical predictors (age, sex, tumour, site and 7th edition AJCC staging). In addition a Kaplan–Meier estimator with 95% CI was used to estimate the survival probability, and pairwise log-rank test was used to test the significance between the subgroups generated by the GNN models with attention mechanisms.

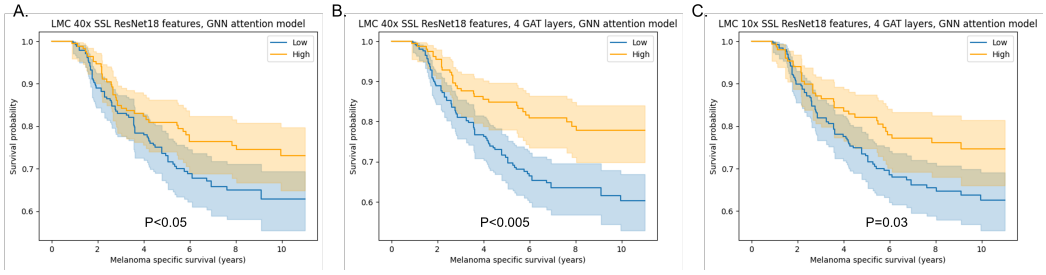


Figure 5.8: Survival distributions for the 'high' v.s. 'low immune' test set predictions from GNN models which utilise an attention mechanism. (A) GNN model with 3 GATv2 layers, trained with 40x graph WSI representations. (B) GNN model with 4 GATv2 layers, trained with 40x graph WSI representations. (C) GNN model with 4 GATv2 layers, trained with 10x graph WSI representations.

In Table 5.6 and 5.7 we demonstrate that although all three top performing GNN models with attention mechanisms, show similar AUC, BA and F1 score metrics, only the model with four GATv2 layers, that uses 40x graph representations can predict 'high immune' patients that have a significantly lower hazard of melanoma death subgroup compared to the 'low immune'

subgroup (HR of 1.89 [95% CI 1.27, 2.86]). Moreover, when we examined the Kaplan Meier curves for all three models with attention mechanisms, we see that for the 40x graph input with a three layer GNN model and 10x graph input four layer GNN model, there is more overlap in the survival distributions, from around 0-4 MSS years (Fig. 5.8). This suggests these models are less able to classify patients at an early stage of follow-up. Additionally out of the five GNN models with highest AUC metrics, the GNN model that used multi-level inputs generated the highest HR for stratifying 'high' and 'low immune' test set patients, when using multi-variable analysis (Table 5.7). This suggests that multi-level representations have greater prognostic value, compared to graphs of a single resolution. However, when comparing the GNN models to SOTA MIL models, the TransMIL model which used 10x patch feature inputs appears to have superior prognostic ability with a HR of 2.27 (95% CI 1.52, 3.45). Here, the top performing GNN, DSMIL and TransMIL models all generate higher HR values, when carrying out multi-variable analysis, compared to the ground truth label HR (Chapter 4, Fig. 4.2).

Moreover, we examined the prognostic ability of the binary GNN classifiers when including the 'intermediate immune' subgroup cases in the test set. The rationale behind this was to see if the GNN models still had prognostic ability when classifying the whole dataset including the 'intermediate immune' subgroup, as this subgroup constitutes around 40% the real world

Table 5.7: The prognostic value of the 'high immune' and 'low immune' subtype classifications in multivariable analyses for the LMC dataset. The HR (95% CI) column indicates when classifying only 'high' and 'low immune' cases (N=410) the 'HR (95% CI) with Intermediate' column indicates when the same model is used to classify all cases from the whole dataset (N=667) including the 'intermediate immune' subgroup. The results shown are from GNNs with and without an attention mechanism instead of global mean pooling operation and the number of GATv2 layers in the model is shown in the GATv2 column. All GNN models with and without attention mechanisms use features extracted using a ResNet18 trained with SSL and histology images. The HRs are also shown for the best performing MIL models, here the DSMIL 40x uses 40x patch features extracted using ResNet50 pretrained with ImageNet and the DSMIL 10x and TransMIL 10x methods use 10x patch features extracted using a ResNet18 trained with SSL and histology images. A HR > 1 implicates a higher risk of death in the predicted 'low immune' compared to the 'high immune' subgroup of patients. Significant P values of < 0.05 are indicated by * and P values < 0.005 are indicated by **. The HRs with the greatest magnitude are highlighted in bold.

Method	GATv2	HR (95% CI)	HR (95% CI) with Intermediate
GNN 20x	3	1.82 (1.19, 2.70)**	1.43 (1.09, 1.85)*
GNN multi-level	4	1.89 (1.27, 2.86)**	1.54 (1.19, 2.00)**
GNN attention 40x	3	1.35 (0.92, 2.00)	**1.79 (1.37, 2.33)****
GNN attention 40x	4	1.81 (1.19, 2.78)*	1.61 (1.25, 2.13)**
GNN attention 10x	4	1.33 (0.89, 2.00)	1.35 (1.04, 1.75)*
DSMIL 40x	-	1.96 (1.25, 3.13)**	1.64 (1.16, 2.33)**
DSMIL 10x	-	1.89 (1.27, 2.78)**	1.56 (1.34, 2.13)**
TransMIL 10x	-	**2.27 (1.52, 3.45)****	1.49 (1.10, 2.04)**

patient population. Table 5.7, shows how the GNN attention model using 40x patch features generates the highest HR when including the 'intermediate immune' cases. This is an unexpected outcome as this model generates one of the lowest HRs when classifying 'low immune' and 'high immune' cases, suggesting that this model may have greater discriminative ability for borderline 'intermediate' cases and less ability for the more polarised 'high' and 'low immune' cases. Conversely, as discussed in Chapter 4, the TransMIL model, which generated the best HR for the binary cases, generates one of the lowest HRs when the 'intermediate' cases are included. This suggests that the model is less capable as classifying more heterogenous cases and may be classifying based on simplified features such as presence of lymphocytes.

5.4 Discussion

In this chapter, we demonstrate how GNN and graph representations of WSIs can improve performance over current state-of-the-art MIL methods for classifying WSIs into immune subgroups. While we still show limited performance for classifying the three immune subtypes (mean test AUC of 0.63), the improved AUC performance over SOTA MIL models, suggests that the additional context from GNNs could be important in deciphering the heterogeneity within the 'intermediate' subgroup.

We also show how the number of graph attention message passing (GATv2) layers can impact model performance (Table 5.4). Similar to results by [138], we see that increasing the number of message passing steps in the model appears to improve information transfer through the network, leading to increased performance, when using one-hot encoded node embedding and 40x single resolution graph inputs. Showing that larger graph representations require additional GATv2 layers to capture longer-range spatial relationships between node embeddings. We also demonstrated how the addition of an attention mechanism to replace graph mean pooling led to mixed results. Here, we found that attention models trained with 40x single resolution graph inputs generated the best overall performance, yet the attention models which were trained with multi-resolution graph inputs had worse performance than when training without the attention mechanism. We believe this may be due to the attention mechanism not being able to retain the hierarchical features of the graphs, due to the large size and imbalance of nodes within the representations. In order to address this, we could take inspiration from a study by Huang *et al.* [235] who developed *AttPool*, a module to concurrently learn local and global attention, to better learn hierarchical representations, or Deng *et al.* (2024) who developed an attention mechanism for a cross-scale MIL to explicitly model inter-scale relationships [236].

Moreover, to generate features for all graph representations we used the histopathology specific ResNet18 developed by Ciga *et al.* (2022) [125]. This model was trained with 10x, 20x and 40x, resolution images. It is possible that the features, which were generated from a model already trained with multiple resolutions, may already contain multi-resolution information, which means that using multiple resolutions as input is redundant compared to using single

resolution patches. In the future to test this, it would be useful to examine the performance of graphs, which utilised features generated by specific resolution feature extraction models.

Again, as we demonstrated in Chapter 4, we show that a combination of metrics is important for deciphering the best performing model for predicting patient outcomes. Here, we show that while five of the GNN models generate similar AUC performance values (0.80-0.81) (Table 5.6), the HR of the 'high' and 'low immune' subgroups predicted by these models are much more variable, ranging from 1.33 to 1.89 (Table 5.7). With the GNN model with no attention mechanism that was trained with multi-level graph representations, having the greatest HR of all the GNN models for the high and low patients. Whereas, the GNN attention model trained with 40x graph representations had the greatest HR out of all MIL and GNN models when classifying all LMC patients (Table 5.7), including the 'intermediate' immune patients, suggesting that both high resolution cellular detail and context are important when stratifying a heterogeneous patient population. On the other hand, we also add to the evidence from Chapter 4 that models trained with 10x features generate the best balanced accuracy and F1 scores, showing that models trained with these lower resolution input features have reduced model bias to the 'low immune' majority class.

Recent studies [237, 238] have shown how elements of both graph and vision transformer models can be combined to classify WSIs, retaining positional encodings with detailed patch level features. As we see in Table 5.6, the TransMIL model, which leverages self-attention and multi-head attention from vision transformers, has the greatest prognostic ability when classifying 'high' and 'low immune' WSIs. Although our models were trained to predict the immune subgroups, so having the highest AUC, balanced accuracy and F1 score values is more relevant to the task than HR, classifying patients into groups that retain or generate more prognostic information is still important. Therefore in future studies we hope to explore using transformer layers in combination with GNNs as another way to improve model performance, reduce errors and improve the prognostic ability of models.

Nevertheless the ability to classify images into more prognostic subgroups could suggest that the MIL models with higher HR values, could be performing shortcut learning. For example the model may be performing a simplified task of classifying images based on the presence of lymphocytes, rather than locating TILs in the tumour region. This is further demonstrated with the large drop in HR (from 2.27 to 1.49 [Table 5.7]), for the TransMIL model, when more heterogeneous 'intermediate' cases are included in the test set. On the other hand, due to contamination when sampling the transcriptomic data used to generate the ground truth labels, there may be some labelling errors as described in Chapter 4. These labelling errors could create noise in the data, which if learned, would lead to a superior classification performance, but lower HR value.

5.5 Conclusion

Overall, our results demonstrate that using context-aware approaches that utilise graph representations with GNNs can result in better performance compared to the standard weakly supervised MIL methods when performing immune subtyping of melanoma WSIs. These results lead us to believe that graph-based methods could be applied to a wide variety of other problems where MIL is state-of-the-art, especially where spatial relationships in the TME are important.

In the next chapter, we compare our results from classifying the LMC WSIs using the immune subgroup labels [1] to classifying the WSIs using survival as the target for the neural networks.

Chapter 6

Survival prediction for melanoma patients, using graph-based whole slide image representations

In this chapter, we explore how neural networks can be used to identify new prognostic subgroups within the LMC dataset, using patient survival as a target. We then compare the prognostic value of these risk groups to the AJCC staging system version 7 [38] and the immune subgroups identified by Pozniak *et al.* (2019) [1] to determine whether these new signatures are independent or related to established melanoma prognostic factors.

6.1 Introduction

The AJCC staging system [39] is currently the standard for prognostically staging melanoma patients, incorporating various histopathological characteristics (detailed further in Chapter 2). Breslow thickness has been recognised as the most reliable predictor of patient survival for over five decades [239]. However, this system displays wide levels of variance [40, 41] and fails to address thinner stage I-II tumors that may progress to metastatic disease. Notably, a significant number of melanoma-related deaths occur in patients initially diagnosed with early-stage disease [240]. Thus, it is crucial to develop new validated biomarkers to identify aggressive tumors in early-stage patients and enhance the current AJCC staging system for better patient management.

As described in previous Chapters (2, 4 and 5), genetic signatures can add value to the current staging systems [1, 11, 84] and have been thoroughly explored in other types of cancer, such as breast [241] and colorectal cancer [242]. Yet, the heterogeneity of melanoma driven by differences in gene expression [243], can make the development of meaningful genetic signatures

challenging. Moreover, melanoma primary tumours are small, making it difficult to collect adequate samples for large scale gene expression studies and leading to a sampling bias for larger tumours, as described in Chapter 4 in reference to the immune subgroup ground truth labels derived by Pozniak *et al* (2019) [1]. In addition, these types of genetic analyses can be expensive and time consuming.

In contrast, H&E stained slides are routinely used and collected during the treatment of melanoma patients and contain a wealth of phenotypic information beyond the biomarkers currently used in the AJCC staging systems. Moreover, developing automated image analysis algorithms, reduces the burden on pathology departments, which have worldwide staffing shortages [18]. In this chapter, we explore how neural networks, driven by image and survival data alone, can be implemented to discover novel biomarkers within these images and stratify patients into risk groups. We then test if these new signatures have added prognostic value compared to the established staging systems and the immune subgroups delineated by Pozniak *et al.* (2019) [1].

6.1.1 Related works

While many studies applying deep learning image analysis techniques to cutaneous melanoma WSIs focus on lesion subtyping [244, 245, 246], detecting malignancy [247, 248, 249, 250] or classification of driver and suppressive mutations [157, 158, 159, 185], fewer delve into the application of deep learning for survival prediction. This scarcity of research in melanoma survival prediction may be attributed to the adoption of immunotherapy for advanced stage melanoma patients which began in 2012 [163]. The challenge lies in accurately gauging the biological effects of the primary tumor independently of the treatment intervention. Consequently, studies are more likely to utilise survival proxies such as distant metastatic recurrence [176] and sentinel node biopsies [251] to assess patient prognosis. These studies are often biased towards later stage patients, which means they are less likely to examine early-stage biomarkers. Moreover, there are not many large-scale well-annotated melanoma datasets available to carry out this analysis. However, because the LMC study began enrolling patient in the year 2000, before the advent of immunotherapies for the treatment of melanoma, we can directly examine survival of patients to underpin whether visual features in our images contribute to patient survival. In addition, most studies that look at patient survival only report overall patient survival [159, 171, 172, 173], rather than melanoma specific survival, which means these studies are at risk of confounding due to other causes leading to patient deaths. Whereas the LMC dataset has melanoma-specific survival status for patients with long follow-up and has a high proportion of early stage patients, who would benefit from new prognostic biomarkers.

A study by Kulkarni *et al.* (2020) looked at creating a binary classifier for distant metastatic recurrence as a proxy for melanoma specific survival [176]. Patches were randomly sampled, then patches with high cell density, determined using the image analysis tool QuPath [160], were added to a sequence and fed through a recurrent neural network, generating an AUC of 0.905 and

0.880 for predicting metastatic recurrence on two external datasets (n=104 and n=51 patients respectively). Moreover, when carrying out multivariate Cox proportional Hazards (CPH) analysis, the authors found that the score generated from their RNN model, was significantly associated with disease specific survival (HR = 58.7 with 95% CI 4.83–713, p= 0.001). However, the authors acknowledge that these results are limited by the fact the training and two validation datasets are small, with the training cohort containing 108 patients and two validation sets containing under 200 patients. The method is also limited to specific skin types, as regions with high melanin content disrupted the input patch selection stage, by preventing the identification of cellular features, so were excluded from analysis. Following the work by Kulkarni *et al.* (2020) [176], Brinker *et al.* (2021) also used QuPath software [160] to extract cellular features which were combined with patient clinical data, as an input for an MLP with a squeeze-and-Excitation model, paired with patch features extracted using a ResNeXt50 [251]. The authors were attempting to predict lymph node status, which can be treated as a proxy for survival, due to its close association with it. They found that combining imaging features with clinical and cellular features did not improve results when just using imaging features alone, as both methods yielded AUROC scores of 0.61. These results suggest that clinical features and hand-crafted cellular features are highly correlated with the histopathological patch features from CNNs, therefore not adding further prognostic value.

In 2022, Comes *et al.* (2022) [252], proposed their *DeepSVM* model, a binary classifier for predicting disease-free survival, using 5-fold cross validation with 43 melanoma WSIs from the Clinical Proteomic Tumor Analysis Consortium Cutaneous Melanoma (CPTAC-CM) public database. Again the authors used QuPath [160] for preprocessing the slides, dividing the images into patches and only selecting patches with high cellular density for further analysis. They then extracted 2048-dimensional feature embeddings from the patches using three ImageNet pretrained [98] CNNs. Selected features from these CNNs were then used to build three SVM classifiers which returned a classification score for each patient at patch and slide-level. The ensemble of all three models was used to develop the DeepSVM model. The authors [252] also assessed whether using clinical features as an input improved disease-free survival classification, but found that while using the imaging features from the *DeepSVM* model improved on using clinical features alone, the combination of clinical and *DeepSVM* image features did not improve on using the imaging features alone, which adds to the evidence of Brinker *et al.* (2021) [251] that quantitative image-based features may be correlated with clinical features. Although the group achieved an AUROC score of 0.67 for predicting disease-free survival on an independent dataset, the independent cohort had a small sample size of only 11 patients, again demonstrating the challenge of identifying large melanoma datasets with imaging and clinical data for model training and evaluation.

In 2023 Couetil *et al.* developed a machine learning framework to predict survival and metastasis of melanoma patients, using interpretable H&E histopathology image features [253]. Initial experiments were carried out using an InceptionNet [154] which was trained to clas-

sify tumour tissue from image patches based on manual annotations from a pathologist using QuPath [160]. They achieved an F1 score of 0.72 for both predicting 5-years survival and 5-years metastasis, experimenting with three quality controlled datasets from the CPTAC (n=19 patients, n=45 slides), TCGA (n=71 patients, n=81 slides) and Indiana University School of medicine (n=70 patients, n=92 slides). Yet, this study was limited by having few AJCC stage I and stage II patients, for whom new prognostic tools would have the most benefit, moreover they did not carry out multivariate analysis with other clinicopathologic variables, so it is uncertain if these new features are correlated with established prognostic melanoma features or recapitulating them.

In our work we take inspiration from these methods, but instead of examining patient survival using cellular features from selected annotated regions, we leverage patch-graphs for melanoma WSI representations, which we presented in Chapter 5. Here we aim to incorporate contextual information into our patient-level WSI representations, which can then be used to derive risk scores, using a Cox loss function [34, 254]. Furthermore, we explore whether our novel WSI representations that combine features from multiple resolutions generate survival risk groups that are more prognostic than groups from single resolution graph representations.

6.1.2 Contributions

In this Chapter our contributions are as follows: (1) We demonstrate how Survival Graph Neural Networks (SurvGNNs) can be used to generate patient-level risk scores for stratifying melanoma patients from the LMC into prognostic subgroups. (2) We show that models trained with 10x graph representations of WSIs, generate risk groups with the greatest prognostic ability. (3) We highlight how reduction in hidden layer size can improve model performance. (4) We show how our attention GNN and SurvGNN model risk groups are on average sufficient to represent the immune subgroups derived by Pozniak *et al.* (2019)[1].

6.2 Methods

6.2.1 Patient-level representations and ground truth labels

For the experiments described in this chapter, the WSI graph representations described in Chapter 5.2.2, from the 667 LMC WSIs described in Chapter 3 were utilised. Specifically, the single resolution graphs constructed from 10x, 20x and 40x patches and the one-hot-node multi-level graph representations (referred to as multi-level graphs) were used. These WSI graph representations are shown in greater detail in Fig. 6.2. Graph representations were used as inputs for the survival experiments, as previously when using the entire LMC dataset (n=667), the models trained with 40x graph representations stratified patients into immune subgroups with the largest HR and with the highest AUC (Table 5.6 & 5.7), demonstrating the utility of graph representations and GNNs.

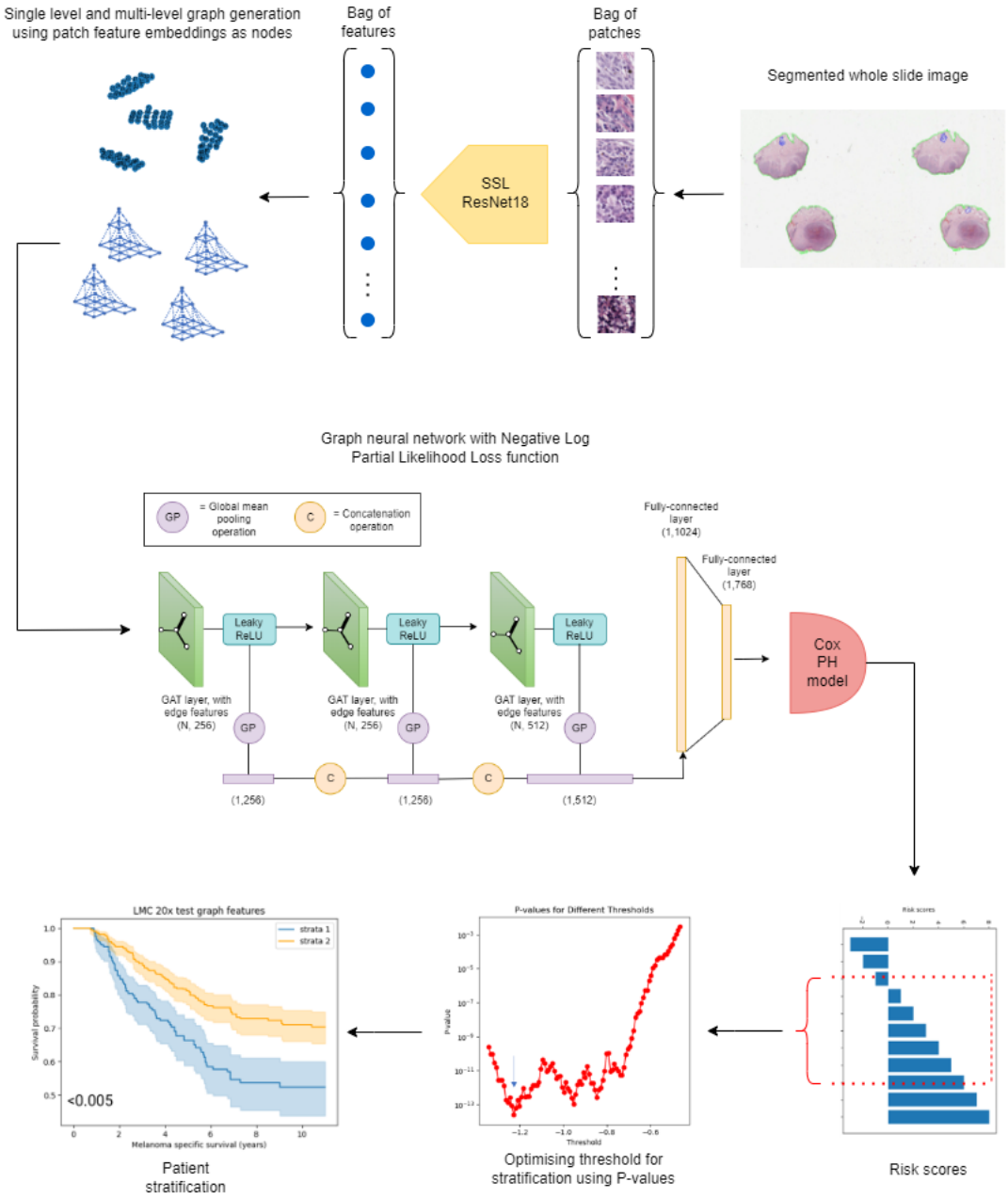


Figure 6.1: Survival analysis pipeline.

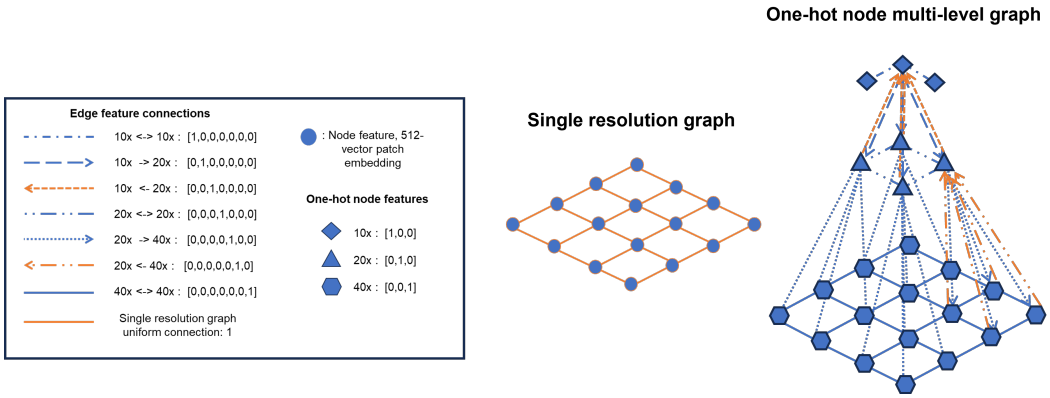


Figure 6.2: A reminder of the graph representations of the WSIs, which are inputs for the SurvGNNs. Single resolution graph representations contain either 10x, 20x, or 40x resolution patch-embedding nodes and have uniform edge connections between vertical and horizontal neighbouring patch-embedding nodes in a grid-like structure. The one-hot multi-level graphs (referred to as multi-level graphs) also contain edge connections between patch-embedding nodes of different neighbouring resolutions, which have one-hot encodings to represent connections between different resolution patch-embedding nodes in different directions. In addition, the nodes of the one-hot node multi-level graph representations have one-hot encodings concatenated to the patch-embeddings to encode additional hierarchical information. WSIs, whole slide images; SurvGNNs, Survival Graph Neural Networks.

Slide-level labels were derived from the follow-up time and whether the patient had died from melanoma, as stated by death certificate, general practitioner records, hospital records, summary care records, or obituary, or had been censored. Here the i^{th} patient was denoted as (t_i, δ_i), where t_i is the follow-up time recorded and δ_i denotes the event data, whether or not melanoma specific death has occurred (yes = 1, or no =0). The follow-up time being denoted as from the time of diagnosis to the time of melanoma death or censorship.

6.2.2 Model architecture

For the experiments described in this chapter, we modified the GNN architectures described in the previous chapter (Chapter 5, Fig. 5.4), experiments were carried out with three *GATv2* layers [141] with a single attention head. Following each GATv2 layer [141] an *ELU* activation layer [230] was applied, as this activation function is smooth and continuously differentiable, so can capture both positive and negative information in the data and facilitate more stable and efficient learning of the model. Dropout regularisation of $P = 0.2$ was applied, during training to prevent model overfitting. Initial experiments aimed to maximise batch size and maintain hidden layer sizes, within the constraints of fitting all graphs from a batch into the memory of a single *Nvidia* V100 32GB GPU. The hidden layer size was maximised to enhance the network's expressivity. The batch size was increased as an attempt to increase the number of patients who had died from melanoma being in the batches, as this was the minority of cases, but uncensored

Table 6.1: The hidden layer size and batch size for graph inputs of different sizes.

Input graph	Hidden layer size	Batch size
10x	256	64
20x	256	15
40x	128	10
Multi-level	64	8

patients are essential for calculating patient risk scores. For the larger input graphs (e.g. those using higher resolution 40x patches), this meant that both the batch size and the hidden layer size for the models were decreased (Table 6.1). However, we also trained models with hidden layers size 64 and batch size of 8 to allow comparison with the larger multi-resolution graphs, which have greater memory constraints.

A global mean pooling mechanism followed by two linear layers was implemented, with the last fully connected layer generating a single score, the patient's hazard risk, denoted as $\hat{R}(\mathbf{X}_i)$. This is the network's prediction of the hazard rate at a specific time for the i^{th} patient based on their corresponding WSI graph representation.

We also tested how adding an attention mechanism instead of global mean pooling mechanism affected results, as in the previous chapter we found that the model which stratified all patients (including intermediate patients) with the highest HR, was a GNN model with an attention mechanism. The attention mechanism consists of an additional GATv2 layer (as previously described in Chapter 5.2.3), which is used to compute unnormalised attention scores for each node embedding, which were then normalised using a *softmax* operation. The node embeddings from the third GATv2 layer were then updated using a weighted aggregation of the node embeddings and the normalised attention weights. These attention weights were then used to determine the high attention patches, from the best performing models, by selecting the node embeddings or patches which have the highest normalised scores. Two fully connected layers output logits, which were used for raw patient-level risk score values ($\hat{R}(\mathbf{X}_i)$).

6.2.3 Loss function

As we described previously, the Cox proportional hazards model [144] is one of the most common methods for survival analysis and can be used to relate the hazard for an individual at a certain time undergoing the event of interest (melanoma death). The model assumes the hazard function is comprised of $h_0(t)$, a baseline hazard function and $R(\mathbf{X}_i)$ a relative risk function denoting the effects of an individual's covariates, which is a weighted sum of covariates ($\mathbf{X}_i$) multiplied by their corresponding regression coefficients:

$$h(t|X_i) = h_0(t) \cdot \exp R(\mathbf{X}_i)$$

where $h(t|X_i)$ represents the hazard rate at time t, for an individual with covariate values X_i.

The Cox partial likelihood [144], can then be defined as the product of conditional probabilities of observing the event for those who experienced the event, given the individuals who are still at risk (not censored) at each event time:

$$\mathrm{L} = \prod_{i:\delta_i=1} \frac{\exp(\hat{R}(\mathbf{X}_i))}{\sum_{j:t_j \geq t_i} \exp(\hat{R}(\mathbf{X}_j))}$$

for all individuals i with $\delta_i = 1$. Here, the product is taken over all individuals i who experienced the event ($\delta_i = 1$). The summation ($\sum$) is taken over all individuals j who are still at risk ($\delta_j = 1$ or $\delta_j = 0$) at each event time ($j : t_j \geq t_i$). $\exp(\hat{R}(\mathbf{X}_i))$ represents the estimated hazard risk from the model's predictions, which is the exponential of the raw risk scores from the GNN model.

The corresponding negative log partial likelihood can them be derived and used as the loss function for our model [144, 149, 150]:

$$L(\theta) = -\frac{1}{N_{\delta=1}} \sum_{i:\delta_i=1} \left(\log(\hat{R}(\mathbf{X}_i)) - \log \left(\sum_{j:\delta_j=1} \hat{R}(\mathbf{X}_j) \right) \right)$$

During training we aim to adjust the parameter values (θ) of the GNN model, by minimising the negative log-likelihood, to maximise the likelihood of finding the network parameters that best describe how the covariates influence the hazard risk. The loss function takes the raw risk values ($\mathbf{X}_i$) from the GNN models as inputs along with a binary tensor, which indicates the censoring status of the patient ($\delta_j = 1$ or $\delta_j = 0$). A normalisation factor $\frac{1}{N_{\delta=1}}$ is used to ensure that the loss is consistent with different numbers of uncensored cases in batches and to stabilise model training. The survival loss function focuses on modeling the likelihood of observed events ($\delta = 1$), but censored cases influence the loss indirectly via the risk set $\delta_j = 1$. However, their contribution is not tied to a specific event time, so they are not used directly in the numerator or for normalisation.

6.2.4 Model implementation

For training, an Adam optimizer [231] with a learning rate of $2 \times 10e^{-4}$, and weight decay of $1 \times 10e^{-5}$ was implemented. The performance of models was assessed using concordance index (C-index) [221, 222], which quantified whether patients are listed in the correct order based on the lengths of their of follow-up time. It is calculated as:

$$C\text{-index} = \frac{\text{Number of Concordant Pairs}}{\text{Number of Concordant Pairs + Number of Discordant Pairs}}$$

Here the proportion of correctly ordered (concordant) pairs among all possible pairs is calculated, providing a measure of how well the model discriminates between patients with different survival outcomes. In the context of survival analysis a higher C-index indicates better discriminatory power of the model, in terms of ranking the relative risks of patients. It

is worth noting, however, that the C-index, has several limitations, for example, the C-index does not have the ability to discriminate between survival times based on differing magnitudes and therefore may penalise the incorrect ranking of two subjects with similar survival times. In addition, the C-index can be affected by a high-degree of censoring and imbalances within data, which is likely to affect our results as 71% of LMC dataset is censored data. Nonetheless, the C-index [221, 222] is a widely used metric that can give valuable insight into the ranking ability of the models, so we implemented it with caution and use complementary metrics in subsequent analysis. We show the C-indices for multivariable models with 95% confidence intervals (CI) generated from 1000 iterations of bootstrapping.

Models were trained for 1000 epochs, and the model with the best C-index on the validation set for each fold in the 10-fold cross validation was selected. The SurvGNN models were implemented using PyTorch version 1.13.1 [232], PyTorch geometric version 2.2.0 [233] with CUDA 11.6 [234], using one Nvidia V100 32GB GPU from the JADE2 HPC facility. The *lifelines* Python package version 0.27.8 [255] was used for the survival analysis described in the subsequent sections.

6.2.5 Generating de-novo survival-based subgroups

In order to stratify patients based on the estimated risk scores, a cut-point search via the minimal-p-value approach was carried out [256]. This method aims to find the optimal threshold for stratifying patients based on the smallest p-value from log-rank tests for the validation set. This method was chosen, because during preliminary tests using the continuous risk values in univariate analysis did not generate significant results, suggesting there was not a smooth relationship between risk scores and survival. Moreover, while the minimal-p-value approach can lead to inflation of type I errors due to multiple testing [257], using arbitrary dichotimisation based on a median risk value, can lead to significant information loss.

To find the optimal cut-point, all patient-level predicted partial hazard risk scores generated from the best performing models were ranked in order. One hundred evenly distributed cut-points were then determined within the interquartile range of the calculated risks, excluding the lower 25% and the upper 25% of the patient-level risk scores, to prevent strata containing a large imbalance in the number of patients. The risk scores from the validation set into two strata based on these one hundred potential thresholds. Using the python lifelines package, log-rank p-value statistics were calculated for the survival distributions of the stratified groups generated from the cut points. The p-values for these one hundred different combinations of strata were then plotted and the risk score threshold with the lowest p-value was chosen, as we believed this would produce the superior patient stratification for the test set. This cut-point was then used to stratify the test set risk scores, which were used for the subsequent CPH analyses.

In addition, due to the heterogeneity of melanoma and previous genetic studies [1, 84] suggesting that as well as high and low risk patients there can be an intermediate subset

of patients, we decided to stratify patients into three subgroups as well. To do this, again the partial hazard risk scores generated by the 10 top performing models from the 10-fold cross validation were ranked. Nevertheless, as finding two minimal p-value cut-off points is not trivial, for the task of finding three risk groups, patients were trichotomised into three equal groups, without using the minimal-p-value approach. This meant determining the cut-point that separates the first tertile from the rest of validation set and a second cut-point that separates the second tertile from the third tertile of the validation set. As mentioned previously arbitrary cut-off points can lead to information loss [256], however, as the minimal-p-value approach we implemented requires each stratum to contain a minimum of 25% of the data, the cut-off point determined using a minimal p-value approach is likely to be similar to using two cut-points to divide the validation set into three equal subsets. These two cut-points from the validation set, were then used to determine the three risk strata for the test set.

6.2.6 Univariate and multivariate statistics

Kaplan-Meier estimators [219] with 95% CI were implemented to assess the survival distributions of the different SurvGNN risk groups, with a pairwise log-rank test being used to measure if there was a significant difference between the survival distributions. Univariate CPH regression [144] was used to assess the associations between the SurvGNN risk groups and MSS. Multi-variable survival analysis was conducted to assess the joint effects of the risk groups with up to five covariates (age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1], [Table 3.1]) in the Cox model. Age of diagnosis [145], sex [2] and site of melanoma [258] were included as there is evidence that these factors are independent predictors of patient outcome (shown in Fig. 6.3). When carrying out univariate and multivariate statistics for covariates with three levels, such as AJCC staging, immune subgroups and three risk groups, dummy variables were used and the lowest level was used as the reference group (e.g. Stage I, high immune, or the low risk strata). We also carried out step-wise regression with prognostic subgroupings from the previous chapters, and applied Ridge (L2) regularisation, with an L2 penalty of 0.1 to mitigate for multicollinearity between the model subgroup covariates [259]. CI and associated p-values for the HRs were calculated using the Wald tests, using *lifelines* (version 0.27.8) packages [260] in Python programming language.

Moreover, we show partial Akaike information criterion (AIC) [261], C-index [221, 222] and likelihood-ratio test statistics as calculated by the *lifelines* package, to evaluate the performance and goodness of fit of the CPH models. Where a model that has a smaller AIC indicates a better balance between model fit and complexity, as AIC penalises models with more parameters. The partial AIC is used for CPH models, because the model is maximising a *partial* log-likelihood. In addition, to assess whether the risk groups added prognostic information to the already established clinical variables, likelihood-ratio tests were carried out. This involved measuring the difference of how well two nested models fit the LMC data - using a simplified model where

the established clinical variables for MSS were used as covariates and a more complex model where the new risk groups of interest were also included as covariates. The test statistic is then calculated as twice the difference in log-likelihoods between the two models and compared to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

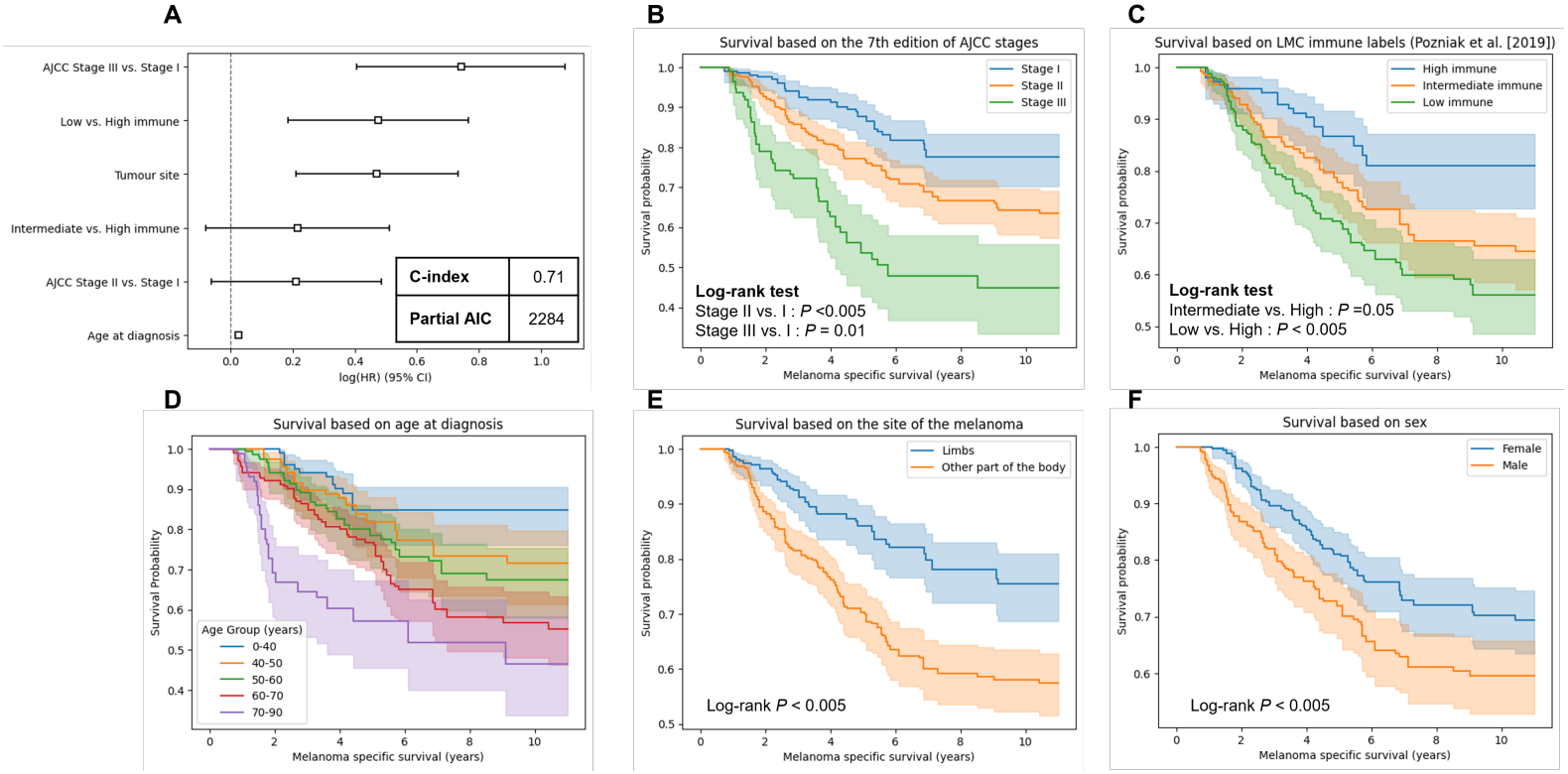


Figure 6.3: The clinical factors associated with MSS. (A) Log HR values with 95% CI for the clinical factors associated with MSS. The C-index and partial AIC values are also shown for the multivariate CPH model fitted with these covariates. (B-F) Kaplan- Meier estimators with 95% CI to show the survival distributions for different clinical factors, with two sided log-rank statistics. (D) Age at diagnosis is shown here as a categorical variable to demonstrate the adverse effect on survival, however, is treated as a continuous variable in our analyses. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, Hazard ratio; MSS, melanoma specific survival.

6.2.7 Treatment target expression analysis

To determine whether the risk groups generated were associated with differential expression of potential treatment targets, we also analysed the relative distributions of seven potential treatment targets (BATF3 [262], PD-L1 [7, 63], CTLA-4 [5, 7, 263, 17], TIM-3 [15], IDO-1 [264], LAG-3 [191] and PD-1 [5, 7, 63]) and an immune score value, which corresponds to the combined expression of the 27 immune cells used to generate the three immune subgroups by Pozniak *et al.* (2019) [1]. The gene expression values, which were measured as part of the whole transcriptome using DASL, were normalised as described in Nsengimana *et al.* (2018) [84]. These normalised values were then dichotomised based on the median expression value for each biomarker to establish whether each patient had high or low expression of the treatment targets. We then counted the number of patients in each risk group which had either a high or low expression of these targets, to see the distributions of immune target expression in the patient risk groups.

6.3 Results

6.3.1 Analysis of high and low risk groups

Initially, patients were divided into 'high' and 'low risk' strata using the risk scores derived from the SurvGNN models. When classifying patients into high vs. low risk groups, models which used the global mean pooling mechanism and had varied hidden layer sizes, performed the best when utilising 10x, 20x and multi-level graph representations (Table 6.2). These models all had a C-index of 0.72. The multivariate CPH model which used risk groups generated from 10x WSI representations had the lowest partial AIC value (2277) and largest multivariate HR value (1.56 [95% CI 1.18, 2.08]) for the risk groups (Table 6.2). This could be attributed to 10x patch features containing both cellular detail and tumoural structure, which give improved prognostic information for stratifying patients into groups associated with MSS, which we noted in Chapter 4. In addition, as previously explored in Chapter 5, 10x WSI representations contain fewer nodes, which means there is less node heterogeneity compared to 40x WSI graph representations and less noisy instances contributing to the patient-level representations. In agreement with this finding, the risk groups generated from the SurvGNN trained with 40x graph representations had worst performance metrics, which again suggests that increased noisy node instances leads to poorer patient-level representations. On the other hand, for the models with attention mechanisms and varied hidden layer sizes, the high vs low risk patient groups generated utilising multi-level graphs had the highest HR values (1.59 [95% CI 1.22, 204]), with the multi-level multivariate CPH model having the best performance metrics (C-index = 0.73, partial AIC=2274). Unlike the models trained using global mean pooling mechanisms, for the models which used an attention mechanism, 40x graph inputs had competitive performance with the multi-level and 10x graph input models (Table 6.3). The attention mechanism appeared

Table 6.2: Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models trained with different graph WSI representation inputs using global mean pooling mechanism. The models varied in the size of the batches (n=64, n=15, n=10, n=8) and hidden layers (n=256, n=256, n=128, n=64). The HRs are shown with 95% CI. The C-index with 95% CI and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input	Univariate HR	Multivariate HR	C-index	Partial AIC
10x	**1.69 (1.28, 2.22)** $P = 2.1 \times 10^{-4}$	**1.56 (1.18, 2.08)** $P = 1.8 \times 10^{-3}$	**0.72 (0.69, 0.76)**	**2277**
20x	**1.69 (1.32, 2.22)** $P = 5.5 \times 10^{-5}$	1.41 (1.08, 1.81) $P = 0.01$	**0.72 (0.69, 0.75)**	2280
40x	1.37 (1.06, 1.79) $P = 0.01$	1.35 (1.04, 1.72) $P = 0.02$	0.71 (0.68, 0.75)	2281
Multi-level	1.51 (1.16, 1.96) $P = 1.6 \times 10^{-3}$	1.43 (1.10, 1.85) $P = 7.0 \times 10^{-3}$	**0.72 (0.69, 0.75)**	2279

to disadvantage models trained with 20x inputs as the high vs. low risk patient subgroups generated using 20x WSI graph representations were not significant when using a multivariate CPH model. This again follows the results we presented in Chapter 5, suggesting that models which use attention mechanisms perform better for larger more heterogeneous graph inputs (40x and multi-level representations).

Furthermore, reducing the hidden layer (n=64) and batch sizes of the SurvGNN models with global mean pooling mechanisms, led to improvements in C-index and partial AIC metrics for the univariate and multivariate CPH models which used the 10x and 40x graph WSI representations as covariates. Here the multivariate CPH model using 10x WSI graph risk groups, generated the best overall AIC (2261) and C-index (0.73) values (Table 6.4). This was also the case for the attention models, trained using 40x and 20x inputs, with the 40x model having the best partial AIC (2267) and multivariate HR values out of the attention models (1.79 [95% CI 1.39, 2.33]) (as shown in Table 6.5). This indicates that reducing the size of hidden layers, as well as improving model efficiency, could be improving the prognostic value of the risk scores, by reducing model complexity and preventing model overfitting.

For the high and low risk groups generated from smaller (n=64 hidden layer size) 10x graph WSI representation models, we show Kaplan-Meier plots with 95% CI and pairwise log-rank tests (Fig. 6.4B), to demonstrate the survival distributions of LMC test set patients. Here we see the models are able to stratify patients into risk groups with significantly different survival distributions. In addition when examining the forest plot in Fig. 6.4A, the log HR for the high vs. low risk groups generated by the smaller model has the HR with the greatest magnitude,

Table 6.3: Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models with an attention mechanism trained with different graph WSI representation inputs. The models varied in the size of the batches (n=64, n=15, n=10, n=8) and hidden layers (n=256, n=256, n=128, n=64). The HRs are shown with 95% CI. The C-index with 95% CI and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input	Univariate HR	Multivariate HR	C-index	Partial AIC
10x	1.61 (1.27, 2.08) $P = 1.2 \times 10^{-4}$	1.45 (1.12, 1.85) $P = 4.3 \times 10^{-3}$	0.72 (0.69, 0.76)	2278
20x	1.33 (1.03, 1.69) $P = 0.027$	1.16 (0.90, 1.49) $P = 0.24$	0.71 (0.68, 0.75)	2285
40x	1.79 (1.39, 2.33) $P = 8.0 \times 10^{-6}$	1.49(1.15, 1.96) $P = 2.6 \times 10^{-3}$	0.72 (0.68, 0.75)	2278
Multi-level	**1.82 (1.41, 2.33)** $P = 4.0 \times 10^{-6}$	**1.59 (1.22, 2.04)** $P = 4.5 \times 10^{-4}$	**0.73 (0.70, 0.76)**	2274

when carrying out multivariate CPH analysis.

Overall, the smaller 10x model which used a global mean pooling mechanism appeared to

Table 6.4: The effect of reducing batch and hidden layer size on the SurvGNN models which use graph mean pooling modules. Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models using graph mean pooling, trained with different graph WSI representation inputs. All models were trained with a batch size of n=8 and hidden layer size of n=64 to accommodate for memory constraints. The HRs are shown with 95% CI. The C-index with 95% CI and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input	Univariate HR	Multivariate HR	C-index	Partial AIC
10x	**2.33 (1.82, 3.03)** $P = 7.5 \times 10^{-11}$	**1.96 (1.51, 2.56)** $P = 2.7 \times 10^{-7}$	**0.73 (0.70, 0.77)**	**2261**
20x	1.56 (1.22, 2.00) $P = 3.9 \times 10^{-4}$	1.35 (1.05, 1.72) $P = 0.02$	0.72 (0.69, 0.76)	2281
40x	1.64 (1.28, 2.08) $P = 9.6 \times 10^{-5}$	1.52 (1.19, 1.92) $P = 9.3 \times 10^{-4}$	**0.73 (0.69, 0.76)**	2275
Multi-level	1.51 (1.16, 1.96) $P = 1.6 \times 10^{-3}$	1.43 (1.10, 1.85) $P = 7.0 \times 10^{-3}$	0.72 (0.69, 0.76)	2279

Table 6.5: The effect of reducing batch and hidden layer size on the attention SurvGNN. Univariate and multivariate analysis in relation to MSS for the high vs. low risk subgroups, generated from the SurvGNN models with attention mechanisms trained with different graph WSI representation inputs. All models were trained with a batch size of n=8 and hidden layer size of n=64 to accommodate for memory constraints. The HRs are shown with 95% CI. The C-index with 95% CI and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input	Univariate HR	Multivariate HR	C-index	Partial AIC
10x	1.72 (1.33, 2.27) $P = 3.2 \times 10^{-5}$	1.49 (1.15, 1.92) $P = 2.8 \times 10^{-3}$	0.72 (0.68, 0.76)	2278
20x	1.59 (1.23, 2.04) $P = 3.0 \times 10^{-4}$	1.45 (1.06, 1.89) $P = 5.6 \times 10^{-3}$	0.72 (0.69, 0.76)	2279
40x	**2.08 (1.61, 2.70)** $P = 1.6 \times 10^{-8}$	**1.79 (1.39, 2.33)** $P = 1.1 \times 10^{-5}$	**0.73 (0.69, 0.76)**	**2267**
Multi-level	1.82 (1.41, 2.33) $P = 4.0 \times 10^{-6}$	1.59 (1.22, 2.04) $P = 4.5 \times 10^{-4}$	**0.73 (0.68, 0.76)**	2274

stratify patients into risk groups with superior performance metrics. This suggests that smaller less complex models, might lead to better performance, potentially by mitigating overfitting that larger more complex models with attention mechanisms may experience. This may also contribute to the poorer performance of models that use multi-level inputs, as their complexity may also be causing overfitting. In addition, although the multi-level graph contain one-hot nodes that attempt to represent the different resolutions, both the attention and pooling mechanism do not account for differences in resolution scale and the differences in proportions of the three different resolutions of node embeddings. This may be preventing hierarchical features being retained in the final risk score, contributing to reduced performance (as discussed in Chapter 5).

Meanwhile, the 40x attention model using n=64 size hidden layers has superior metrics to all other (20x, 40x, and multi-level) global mean pooling and attention-based models apart from the 10x model which used global mean pooling and smaller hidden layers. This suggests that there is still value in 40x graph WSI representations, especially when utilising an attention mechanism. One reason that the 40x attention models might have better performance than the 10x and 20x attention-based models, is that the higher resolution cellular features could be more prognostically relevant, but when using the global mean pooling mechanism, more noisy instances are included in the classification. This is probable given that 40x WSI graph representations have the greatest number of node embeddings, so are more likely to contain more noisy embeddings which are not relevant to the task. Another reason 40x graph representation models may benefit more from attention mechanisms than lower resolution representations

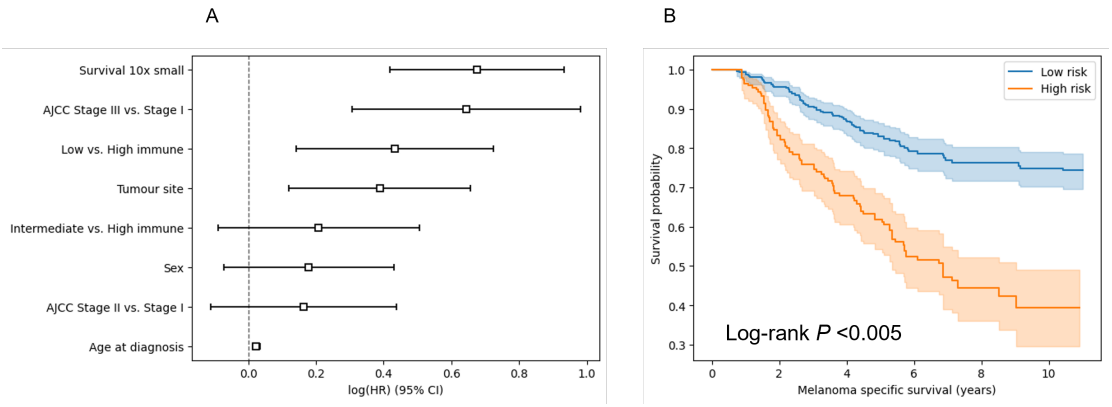


Figure 6.4: (A) Log HRs with 95% CI of clinicohistological features in relation to MSS. 'Survival 10x small' indicates the High risk (n = 203) vs. Low risk (n = 465) strata generated by the SurvGNN models (hidden layer size n=64, batch size = 8) trained and tested using 10x graph representations of LMC WSIs. Low risk patients are used as the baseline to generate the HR. (B) Kaplan-Meier survival analysis with 95% CI of the risk groups from the 10x SurvGNN with two-sided log-rank test. CI, Confidence Intervals; HR, Hazard Ratio; LMC, Leeds Melanoma Cohort; MSS, Melanoma Specific Survival; SurvGNN, Survival Graph Neural Network; WSI, Whole Slide Image.

could be due to the higher heterogeneity exhibited between 40x node embeddings. Since 40x patches span a small high resolution 256×256 pixel area within a WSI, they only contain certain tissue types and cells, compared to lower resolution patches which contain a greater mixture of both cells and different tissue types. While individual 10x patches may be more heterogenous, collectively they are more homogeneous, whereas for 40x graph representations although there is greater homogeneity within the confines of the embeddings generated from 256×256 pixel patches, there is a greater difference and diversity between individual patches meaning that as a collection of patches or embeddings they are more heterogeneous. This means that although 40x models are bigger and more complex, because the graph representations are more diverse they may be less prone to overfitting. Hence the 40x models may be generating superior performance, despite the attention mechanism introducing additional model complexity.

We also examined the patches which had the top attention scores from the attention-based 40x SurvGNN model with the highest HR and C-index and lowest partial AIC (Fig. 6.5). The high risk top attention patches (Fig. 6.5B) were located in tumour regions of the WSIs and although the morphology of tumour cells appears to differ between cases, the pathologist reviewing the high attention regions noted that the model appeared to focus on patches that contain more tumour cells relative to the surrounding stroma (supportive tissue). Tumour stroma ratio has been shown as strong prognostic indicator for various solid tumours [265], but is less studied in melanomas, where studies focus on prognostic indicators such as Breslow thickness or tumour microenvironment factors, such as gene expression profiles [1, 84, 168] and immune infiltration [224], instead of the tumour stroma ratio alone. On the other hand, the

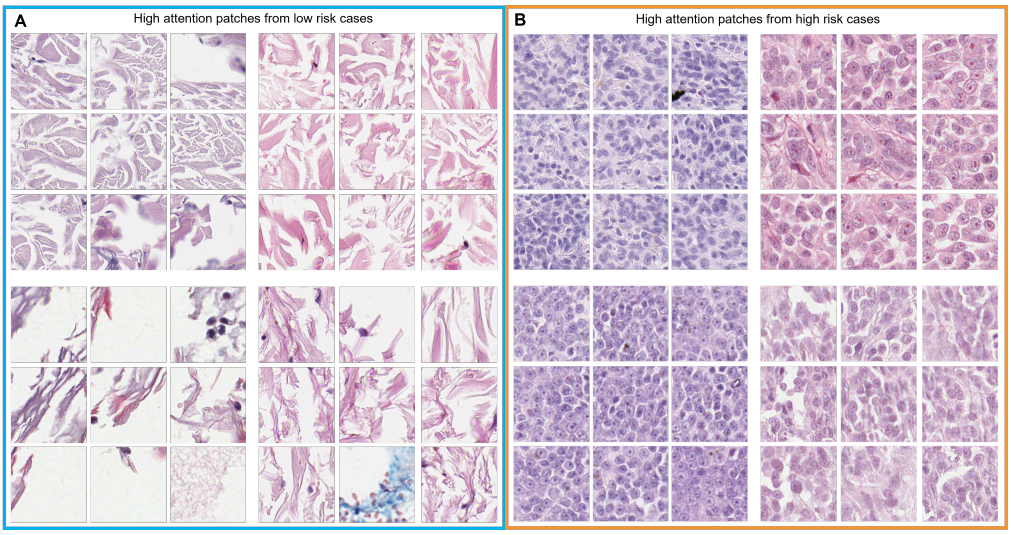


Figure 6.5: The top nine high attention patches from four different melanoma WSI cases classified, as (A) low risk or (B) high risk of melanoma death by the 40x Attention-based SurvGNN model. WSI, Whole Slide Image; SurvGNN, Survival Graph Neural Network.

low risk patient group high attention features, came from patches outside the tumour regions and focused on acellular regions (Fig 6.5A), which do not have prognostic significance. It is possible that this could be related to the attention-based SurvGNN models overfitting to noise, which could also explain why worse than the smaller 10x model with no attention mechanism performs better.

6.3.2 Analysis of high, low and intermediate risk groups

We also stratified patients into three risk groups, as previous studies indicated that there is an intermediate patient population [1, 84]. For models where a global mean pooling mechanism was applied, the multivariate CPH model with the highest C-index (0.73) and joint lowest AIC value (2277), which used the three risk groups as inputs, was the 20x SurvGNN model (Table 6.6 & Fig. 6.6). The HR for high vs. low risk strata generated by this 20x model was 1.57 (1.19, 2.08) ($P = 1.6 \times 10^{-3}$, Table 6.6) and the HR for the intermediate vs. low risk groups was 1.39 (1.04, 1.86) ($P = 0.03$, Table 6.6). Meanwhile, the multivariate CPH which used the three risk strata from the 10x model with smaller hidden layers and batch size had the same partial AIC value and a larger HR value (1.62 [95% CI 1.22, 2.15], $P = 8.0 \times 10^{-4}$) for the high vs. low risk strata from the multivariate CPH model. Yet, the HR for the intermediate vs. low risk strata from this multivariate CPH model was much smaller than the HR from the 20x model and not statistically significant (1.14 [95% CI 0.85, 1.15], $P = 0.38$). The majority of intermediate vs. low risk strata were not significant ($P < 0.05$) when carrying out multi and univariate analysis using the three prognostic groups from the SurvGNN models (Table 6.6, 6.7, 6.8, 6.7).

Table 6.6: Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the SurvGNN models with global mean pooling mechanisms, that were trained with different graph WSI representation inputs. The models varied in the size of the batches (n=64, n=15, n=10, n=8) and hidden layers (n=256, n=256, n=128, n=64). The C-index and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1] Multi refers to the multi-level graph inputs. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival; SurvGNN, Survival Graph Neural Network.

Input		Univariate HR	Multivariate HR	C-index	Partial AIC
10x	Intermediate vs. low risk	0.95 (0.71, 1.25) $P = 0.70$	1.01 (0.76, 1.34) $P = 0.95$	0.72 (0.68, 0.75)	2280
	High vs. low risk	1.53 (1.15, 2.03) $P = 3.0 \times 10^{-3}$	1.47 (1.17, 1.99) $P = 7.0 \times 10^{-3}$		
20x	Intermediate vs. low risk	1.27 (0.95, 1.69) $P = 0.11$	1.39 (1.04, 1.86) $P = 0.03$	**0.73 (0.69, 0.76)**	**2277**
	High vs. low risk	1.82 (1.28, 2.40) $P = 2.3 \times 10^{-5}$	1.57 (1.19, 2.08) $P = 1.6 \times 10^{-3}$		
40x	Intermediate vs. low risk	1.05 (0.79, 1.39) $P = 0.74$	0.94 (0.71, 1.25) $P = 0.69$	0.72 (0.68, 0.75)	2283
	High vs. low risk	1.46 (1.10, 1.93) $P = 8.3 \times 10^{-3}$	1.32 (0.99, 1.75) $P = 0.06$		
Multi	Intermediate vs. low risk	1.15 (0.86, 1.54) $P = 0.34$	1.18 (0.88, 1.58) $P = 0.26$	0.72 (0.69, 0.76)	2280
	High vs. low risk	1.70 (1.29, 2.23) $P = 1.7 \times 10^{-4}$	1.53 (1.17, 1.99) $P = 3.1 \times 10^{-3}$		

The attention-based models had comparable partial AIC and C-index scores to the models that used the global mean pooling mechanism and for all models the intermediate vs. low risk strata were not significant ($P < 0.05$) when carrying out multi and univariate analysis using the three prognostic groups (Table 6.7). This could be due to loss of information when dividing patients into three equally sized groups, but also reflect the heterogeneity of patients and prognostic features within this 'intermediate' subset of patients, which may overlap with the low risk subgroup.

The effect of reducing batch size and hidden layer size on the performance metrics and HR values was similar when assessing three risk groups compared to when analysing just the high vs. low risk groups (Table 6.8 & 6.6). The multivariate CPH model metrics for the risk groups from the 10x and 40x trained models with global mean pooling, improve when the hidden layer and batch sizes are decreased (Table 6.8). Suggesting that for the 10x and 40x models reducing the hidden layer size and therefore model complexity could be preventing model overfitting and improving the acuity of the prognostic risk scores. On the other hand, the 20x model shows the

Table 6.7: Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the attention-based SurvGNN models trained with different graph WSI representation inputs. The models varied in the size of the batches (n=64, n=15, n=10, n=8) and hidden layers (n=256, n=256, n=128, n=64). The C-index and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. Multi refers to the multi-level graph inputs. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival; SurvGNN, Survival Graph Neural Network.

Input		Univariate HR	Multivariate HR	C-index	Partial AIC
10x	Intermediate vs. low risk	1.17 (0.87, 1.57) $P = 0.29$	1.19 (0.88, 1.59) $P = 0.26$	**0.73 (0.69, 0.76)**	**2277**
	High vs. low risk	1.81 (1.38, 2.39) $P = 2.1 \times 10^{-5}$	1.61 (1.22, 2.13) $P = 7.4 \times 10^{-4}$		
20x	Intermediate vs. low risk	1.19 (0.90, 1.59) $P = 0.22$	1.23 (0.92, 1.65) $P = 0.15$	0.72 (0.68, 0.75)	2284
	High vs. low risk	1.49 (1.13, 1.97) $P = 4.5 \times 10^{-3}$	1.32 (1.00, 1.74) $P = 0.053$		
40x	Intermediate vs. low risk	1.05 (0.79, 1.41) $P = 0.72$	1.05 (0.78, 1.41) $P = 0.75$	0.72 (0.69, 0.76)	2279
	High vs. low risk	1.70 (1.29, 2.23) $P = 1.6 \times 10^{-4}$	1.47 (1.11, 1.95) $P = 6.7 \times 10^{-3}$		
Multi	Intermediate vs. low risk	1.00 (0.75, 1.35) $P = 0.98$	1.07 (0.80, 1.45) $P = 0.64$	**0.73 (0.70, 0.76)**	2278
	High vs. low risk	1.72 (1.21, 2.26) $P = 9.1 \times 10^{-5}$	1.55 (1.18, 2.05) $P = 1.7 \times 10^{-3}$		

opposite trend, with a reduction in model parameters leading to a decrease in the C-index value and increase in the AIC value. This suggests that generating risk scores and stratifying these values into risk groups with 20x graph WSI representations, may require a more complex model. This could be due to the features at this resolution being more complicated, as in the previous Chapters 4 and 5, models trained with 10x features and 40x features focused on the presence or absence of lymphocytes, small dark circular cells which are more easily discernible. Moreover, the 20x model with larger hidden layers was the only model where the intermediate vs low risk strata level was prognostically significant in a multivariate CPH model (Table 6.6). This could reflect that stratifying patients into three classes may require greater model complexity or that 20x features are better for deciphering an intermediate class of patients.

On the other hand, for all three single resolution attention-based models, decreasing the model hidden layer and batch size led to improved performance metrics, with the 40x model having the lowest partial AIC and joint highest C-index value (Table 6.9). Nevertheless, although the 40x attention-based SurvGNN model three risk groups generated a multivariate

Table 6.8: The affect of reducing batch and hidden layer size on SurvGNN models with global mean pooling mechanism that generate three risk groups.Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the SurvGNN models with global mean pooling mechanisms, that were trained with different graph WSI representation inputs. The models had a consistent batch size of n=8 and consistent hidden layer size of n=64. The C-index and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. Multi refers to the multi-level graph inputs. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input		Univariate HR	Multivariate HR	C-index	Partial AIC
10x	Intermediate vs. low risk	1.09 (0.81, 1.45) $P = 0.57$	1.14 (0.85, 1.15) $P = 0.38$	**0.72 (0.69, 0.76)**	**2277**
	High vs. low risk	1.92 (1.45, 2.53) $P = 4.0 \times 10^{-6}$	1.62 (1.22, 2.15) $P = 8.0 \times 10^{-4}$		
20x	Intermediate vs. low risk	1.07 (0.80, 1.42) $P = 0.65$	1.01 (0.76, 1.35) $P = 0.93$	**0.72 (0.68, 0.75)**	2284
	High vs. low risk	1.60 (1.21, 2.11) $P = 9.2 \times 10^{-4}$	1.34 (1.01, 1.78) $P = 0.04$		
40x	Intermediate vs. low risk	1.31 (0.98, 1.74) $P = 0.067$	1.23 (0.93, 1.64) $P = 0.15$	**0.72 (0.69, 0.76)**	2278
	High vs. low risk	1.68 (1.27, 2.22) $P = 2.7 \times 10^{-4}$	1.56 (1.18, 2.06) $P = 1.9 \times 10^{-3}$		
Multi	Intermediate vs. low risk	1.15 (0.86, 1.54) $P = 0.34$	1.18 (0.88, 1.58) $P = 0.26$	**0.72 (0.69, 0.76)**	2280
	High vs. low risk	1.70 (1.29, 2.23) $P = 1.7 \times 10^{-4}$	1.53 (1.17, 1.99) $P = 3.1 \times 10^{-3}$		

model with one of the highest C-index values (0.73) and lowest partial AIC values (2272) (Table 6.9), the intermediate and low subgroups had non-significant overlapping survival distributions.

Meanwhile, the three risk groups generated using our novel multi-level graph representations, did not appear to predict MSS better than the single resolution graphs. However, as we were unable to test different hidden layer sizes for models trained with these multi-level representations, it is possible that the smaller hidden layer size and reduced model complexity limits the performance of these models.

6.3.3 Determining the prognostic ability of the risk groups in the presence of established prognostic factors

We also carried out step-wise regression, using backward elimination with all potential prognostic groupings, including those from previous chapters (Chapter 4 & Chapter 5) in order to select the model covariates that might give the most accurate effect estimates from the LMC

Table 6.9: The affect of reducing batch and hidden layer size on attention-based SurvGNN models that generate three risk groups. Univariate and multivariate analysis in relation to MSS for the three risk subgroups, generated from the attention-based SurvGNN models trained with different graph WSI representation inputs. The models had a consistent batch size of n=8 and consistent hidden layer size of n=64. The C-index and AIC are shown for the multivariate CPH models, which are adjusted for age of diagnosis, sex, site of melanoma, AJCC staging version 7 and the immune subgroups defined by Pozniak *et al.* (2019) [1]. Multi refers to the multi-level graph inputs. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; MSS, melanoma specific survival.

Input		Univariate HR	Multivariate HR	C-index	Partial AIC
10x	Intermediate vs. low risk	1.12 (0.83, 1.49) $P = 0.46$	1.10 (0.82, 1.47) $P = 0.53$	**0.73 (0.69, 0.76)**	2275
	High vs. low risk	1.92 (1.46, 2.53) $P = 3.0 \times 10^{-6}$	1.68 (1.27, 2.21) $P = 2.8 \times 10^{-4}$		
20x	Intermediate vs. low risk	1.06 (0.80, 1.41) $P = 0.67$	1.07 (0.81, 1.43) $P = 0.62$	0.72 (0.68, 0.76)	2280
	High vs. low risk	1.65 (1.25, 2.19) $P = 4.6 \times 10^{-4}$	1.50 (1.13, 2.01) $P = 6.3 \times 10^{-3}$		
40x	Intermediate vs. low risk	1.15 (0.85, 1.54) $P = 0.067$	1.08 (0.79, 1.45) $P = 0.15$	**0.73 (0.69, 0.76)**	**2272**
	High vs. low risk	1.88 (1.43, 2.47) $P = 5.0 \times 10^{-6}$	1.66 (1.26, 2.19) $P = 3.2 \times 10^{-4}$		
Multi	Intermediate vs. low risk	1.00 (0.75, 1.35) $P = 0.34$	1.07 (0.80, 1.45) $P = 0.26$	**0.73 (0.69, 0.76)**	2278
	High vs. low risk	1.72 (1.21, 2.26) $P = 9.1 \times 10^{-5}$	1.55 (1.18, 2.05) $P = 1.7 \times 10^{-3}$		

dataset. We carried out this predictor selection, instead of adjusting for all possible covariates, because although the prognostic immune subgroups and SurvGNN risk group predictions may all be highly predictive of MSS, they are unlikely to all be necessary for controlling confounding, therefore, adjusting for all the predicted groupings could result in an effect estimate that is inflated or has unnecessarily wide confidence intervals [266]. In addition, we exercise confounding control as the causal status of our model predicted covariates is uncertain and as the ratio of the number of covariates to sample size increases there is greater risk of p-value and confidence interval distortion [266].

The covariates we adjusted for in our step-wise predictor selection, included the high and low immune subgroups identified from the whole LMC dataset (n=667 patients, including 'intermediate' patient subset), using the best performing MIL models from Chapter 4 - DSMIL 10x, DSMIL 40x and TransMIL 10x - and best performing graph models from Chapter 5 - the GNN attention 10x model, GNN attention 40x (4 GATv2 layers), GNN attention 40x (3 GATv2 layers), GNN 20x and GNN multi-level (shown in Table 6.10). Step-wise removal of the covari-

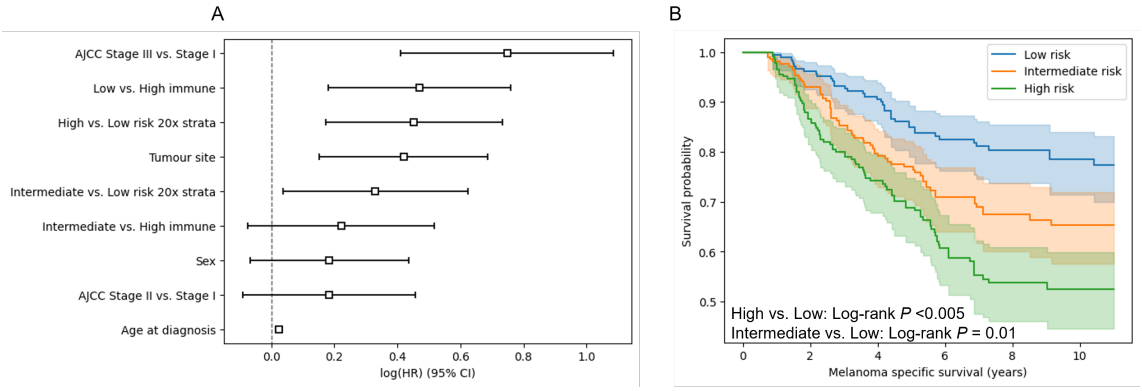


Figure 6.6: (A) Log HRs with 95% CI of clinicohistological features in relation to MSS. High vs Low risk strata and Intermediate vs. Low risk strata represent the HRs for the Low (n = 212), Intermediate (n = 219) and High (n = 236) risk groups generated by the SurvGNN model trained using 20x graph representations of LMC WSIs. Low risk patients are used as the baseline to generate the HRs. (B) Kaplan Meier survival analysis with CI of the risk groups from the 20x SurvGNN with two-sided log-rank test. CI, confidence interval; HR, hazard ratio; MSS, melanoma specific survival; LMC, Leeds melanoma cohort; SurvGNN, survival graph neural network.

ates (including the clinicohistological features [Fig. 6.3] and the model groups [Table 6.10]) with non-significant p-values (the coefficient p-value of the covariate is above $\alpha - level = 0.05$) was carried out and the final multivariate CPH model was fitted to only significant covariates. Here, the high and low risk strata generated by the SurvGNN trained with 10x representations and using a smaller model architecture and batch size remained significant, generating the highest HR (Fig. 6.7). Although AJCC stage II vs. I did not have a significant P-value ($P = 0.29$), it was retained in multivariate analysis of the LMC dataset, as it is a level of the AJCC staging and AJCC stage III vs. I remained significant (2.2×10^{-4}) with the second largest HR (1.89 [95% CI 1.35, 2.66]), showing that staging is a clinically valuable predictor of melanoma survival (Fig. 6.7). Out of all eight GNN and MIL models trained using 'high' and 'low' immune labels determined by Pozniak *et al.* (2019) [1] (Table 6.10), only the subgroups generated by the GNN attention model with three *GATv2* layers trained using 40x WSI representation remained significant ($P = 9.3 \times 10^{-5}$). Moreover, the original Pozniak *et al.* immune subgroups (low vs. high immune and intermediate vs. high immune), no longer remained significant when including the subgroups generated by the GNN attention 40x (3 *GATv2*) model, as previously the low vs. high immune strata were significant in the multivariate CPH with the survival 10x small model (Fig. 6.4). This suggests that on average, the subgroups generated by the GNN attention trained to classify the two immune subgroups and risk groups generated by SurvGNN models are sufficient to represent the prognostic information of the immune subgroups derived by Pozniak *et al.* (2019)[1]. While sex did not remain a significant predictor, both the site of tumour and age at diagnosis remained significant (Fig. 6.7). Although age at diagnosis appears

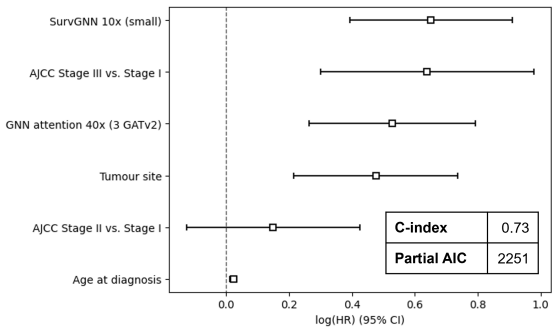
Table 6.10: The prognostic value of the 'high immune' and 'low immune' subtype classifications in multivariable analyses for the LMC dataset compared to the SurvGNN model risk group classifications. The HR (95% CI) column indicates the HR when carrying out multi-variable CPH analysis (adjusting for age at diagnosis, sex, AJCC stage edition 7 and tumour site) when classifying cases from the whole dataset (N=667). For the SurvGNN models, the immune Pozniak *et al.* subgroup labels [1] were also included in the CPH analysis, as these models were not trained with the labels, so are less likely to be correlated with them. A HR > 1 implicates a higher risk of death in the predicted 'low immune' compared to the 'high immune' subgroup of patients. Significant p-values of < 0.05 are indicated by * and p-values < 0.005 are indicated by **. The HR with the greatest magnitude is highlighted in bold. AJCC, American joint committee on cancer; CI, confidence interval; CPH, Cox proportional Hazards; DSMIL, Dual Stream Multiple-instance learning; HR, hazard ratio; GATv2, graph attention network version 2; GNN, graph neural network; MSS, melanoma specific survival; SurvGNN, survival graph neural network; SSL, self-supervised learning; TransMIL, Transformer based Multiple-instance learning.

Method	GATv2	Strata	HR (95% CI)
SurvGNN 10x	3	-	**1.96 (1.52, 2.56)****
SurvGNN 20x	3	Intermediate vs. low	1.39 (1.04, 1.86)*
		High vs. low	1.57 (1.19, 2.08)**
GNN 20x	3	-	1.43 (1.09, 1.85)*
GNN multi-level	4	-	1.54 (1.19, 2.00)**
GNN attention 40x	3	-	1.79 (1.37, 2.33)**
GNN attention 40x	4	-	1.61 (1.25, 2.13)**
GNN attention 10x	4	-	1.35 (1.04, 1.75)*
DSMIL 40x	-	-	1.64 (1.16, 2.33)**
DSMIL 10x	-	-	1.56 (1.34, 2.13)**
TransMIL 10x	-	-	1.49 (1.10, 2.04)**

to have the lowest absolute HR value (HR = 1.02 [95% CI 1.01, 1.03]), age at diagnosis is a continuous variable measured in years, therefore, a HR of 1.02 is indicative that for every increase in the year that a patient is diagnosed with melanoma there is a 2% greater risk of melanoma death.

The multivariate CPH model with only significant covariates from the step-wise removal (Fig. 6.7) had a higher C-index value (0.73) and lower partial AIC value (2251) compared to the CPH model adjusted without the risk groups generated by the SurvGNN or GNN models (Fig. 6.3A). This suggests that as well as being able to sufficiently represent the Pozniak *et al.* immune subgroups [1], adding the groupings from the SurvGNN 10x and GNN attention 40x model is generating greater predictive value for the CPH model. Moreover while the multi-level representations did not generate the best performance for the multivariate and univariate CPH models (Table 6.4 & 6.8), it is interesting to note that the significant model predictors are derived from models that use different resolutions (10x and 40x), showing that features from both 10x and 40x resolutions can be strong predictors of MSS.

In addition to further assess the prognostic ability of the risk groups and immune subgroups



Covariates	HR (95% CI)	*P*-value
SurvGNN 10x (small)	1.92 (1.48, 2.48)	7.8×10^{-7}
AJCC stage III vs. I	1.89 (1.35, 2.66)	2.2×10^{-4}
GNN attention 40x (3 *GATv2*)	1.70 (1.30, 2.21)	9.3×10^{-5}
Tumour site	1.61 (1.24, 2.09)	3.6×10^{-4}
AJCC stage II vs. I	1.16 (0.88, 1.53)	0.29
Age at diagnosis	1.02 (1.01, 1.03)	1.5×10^{-5}

Figure 6.7: Log HRs with 95% CI for significant clinicohistological features and SurvGNN and GNN attention risk groups in relation to MSS. SurvGNN 10x (small) indicates the High risk (n = 203) vs. Low risk (n = 465) strata generated by the SurvGNN models (hidden layer size n=64, batch size = 8) trained and tested using 10x graph representations of LMC WSIs. Low risk patients are used as the baseline to generate the HR. GNN attention 40x (3 *GATv2*), refers to the high (n=273) vs low (n=395) groups predicted for the 667 LMC cases (including 'intermediate' immune cases) by a GNN attention model trained to only classify 'high immune' and 'low immune' cases using the labels determined by Pozniak *et al.* [1]. 'High immune' predicted cases are used as the baseline reference level to generate the HR. The C-index and partial AIC for the multivariate CPH model fitted with the significant features are also shown. AIC, Akaike information criterion; AJCC, American joint committee on cancer; CI, confidence interval; C-index, Concordance index; CPH, Cox proportional Hazards; HR, hazard ratio; GATv2, graph attention network version 2; GNN, graph neural network; LMC, Leeds melanoma cohort; MSS, melanoma specific survival; SurvGNN, Survival Graph Neural Network; WSIs, whole slide images.

from the SurvGNN and GNN attention models, we carried out likelihood-ratio tests on two CPH models fitted with and without the subgroups of interest (Table 6.11). The likelihood-ratio test helps to determine whether the new subgroups significantly improve the fit of the model and therefore contribute to explaining the variability in the survival times, beyond what the clinical variables (AJCC staging, age at diagnosis and tumour site) provide alone. Here, we found that the model subgroups provided significantly independent prognostic information compared to when using the significant clinical variables alone, with the more complex model including the GNN attention 40x immune subgroups and SurvGNN 10x risk groups generating a likelihood-ratio test statistic of 50 and $p - value = 1.0 \times 10^{-11}$, suggesting these additional covariates are improving the model fit and prognostic ability (Table 6.11).

6.3.4 The distribution of potential treatment targets and clinical, histological and immune features in the predicted immune and risk groups

In order to explore if the risk groups generated by the SurvGNN models had underlying immunological correlates, we examined the expression of seven immune checkpoint and inhibitory

Table 6.11: Likelihood-ratio test statistics with p-values to assess the goodness of fit of the models with the new subgroups compared to the standard clinical variables for predicting MSS. The SurvGNN 10x column shows the likelihood-ratio statistics when comparing a CPH model with and without the SurvGNN 10x risk groups in the presence of standard clinical variables and the Pozniak *et al.* (2019) immune subgroups [1]. The GNN attention 40x column shows the likelihood-ratio statistics when comparing a CPH model with and without the immune classes classified by the GNN attention model in the presence of the clinical variables. The final column shows the likelihood-ratio statistics when comparing a CPH model fit with and without both the SurvGNN and GNN attention model biomarkers, with the baseline model containing including all significant clinical variables. CPH, Cox proportional Hazards; GNN, graph neural network; MSS, melanoma specific survival; SurvGNN, Survival Graph Neural Network.

	SurvGNN 10x (small)	GNN attention 40x	SurvGNN 10x (small) + GNN attention 40x
Test statistic	26.5	27.3	50.6
P-value	2.6×10^{-7}	1.8×10^{-7}	1.0×10^{-11}
Degrees of freedom	1	1	2

molecules based on the transcriptomic data. We also looked at a combined score called the 'immune score', which was a sum of all 27 immune cells used to generate the Pozniak *et al.* three immune subgroups [1]. The expression values for each molecule and the immune score were then dichotomised based on the median expression score into low and high expression groups. From Fig. 6.9, it is clear that there is a difference in the distribution of the immune biomarkers in the two different risk groups, with the 'high immune' class having a greater proportion of patients with high expression of the immune molecules, as is expected. Whereas, Fig. 6.8, shows less difference in the proportion of immune molecules between the two different risk groups, indicating that differences in immune cell expression may not be causing differing survival distributions for the high vs. low risk patients stratified by the 10x SurvGNN models.

When examining the clinical, histological and immune features of the high and low risk groups, shown in Table 6.12, it confirms that the SurvGNN risk groups appear to be less defined by immune features, with similar TIL classifications for the two risk groups and similar proportions of patients from the immune subgroups [1] in each risk strata. Regarding the 7th edition AJCC staging, 72% of patients predicted to be in the high risk group comprise of stage I and stage II patients compared to 87% of the low risk group. Although the low risk group has a greater proportion of early-stage patients, it is notable that the high risk group still includes a significant number of these patients. Understanding what makes these early stage patients at higher risk of melanoma death, through identifying new biomarkers for these individuals is crucial for enhancing melanoma treatment outcomes. Moreover, the high risk patient group has a larger proportion of patients with ulcerated tumours (46.8%), compared to the low risk patients (28.0%). Ulceration is a histological feature of tumours, where there is a breakdown of the skin on top of the melanoma. The presence of tumour ulceration in H&E images of patient tumours is used to refine the 'T category' in the 7th edition AJCC staging, as it is associated

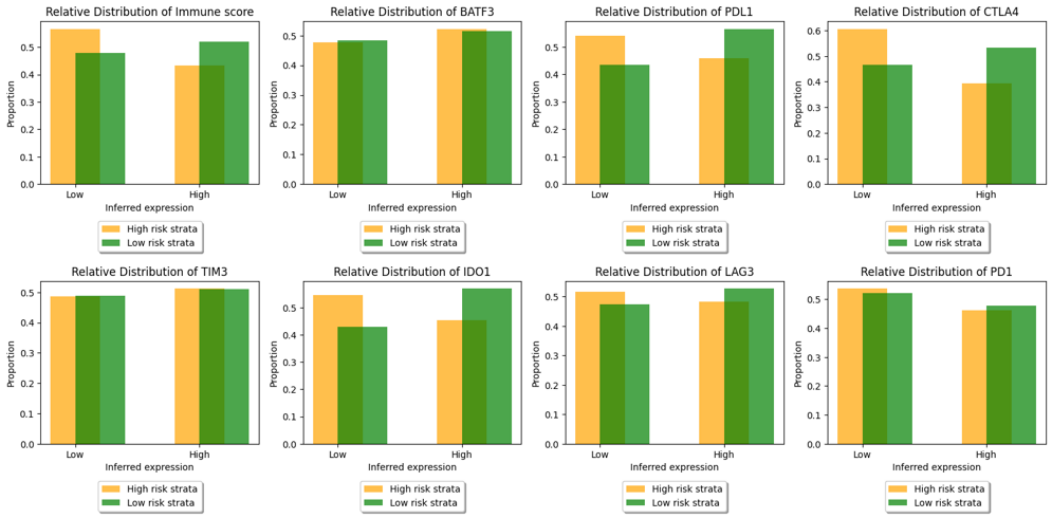


Figure 6.8: Relative distributions of potential treatments targets and immune biomarkers for patients in the high and low risk groups as predicted by the 10x SurvGNN model, using smaller hidden layers and batch size. SurvGNN, Survival Graph Neural Network.

with more aggressive tumour behaviour and poorer patient prognosis [38]. The results indicate that the risk groups could potentially be associated with ulceration, or associated features that are related to more aggressive tumours. We also compared this to the clinical features of the high and low risk groups from the best performing attention-based 40x small SurvGNN model (Table 6.9). Similarly the risk groups from this model were less associated with immune features, but 54% of the tumours in the high risk group were labelled as ulcerated compared to 25% of the predicted low risk group tumours (as detailed in Appendix Table B.1). This again indicates that the SurvGNN models, may be classifying patients based on histological features that are associated with ulceration or more aggressive tumours. Furthermore, the high attention patches for both risk groups in Fig. 6.5, show no immune infiltration again highlighting that the SurvGNN models are unlikely to be stratifying patients based on immune cell presence, but instead based on tumour histology (Fig. 6.5B).

On the other hand, the clinical features of the 40x attention GNN model predicted immune subgroups from the whole LMC dataset (including the intermediate subset of patients) shown in Table 6.13, are, as expected, highly associated with the high and low immune subgroups [1]. This is expected as the predicted high and low immune subgroups were generated from the 40x attention GNN model which was trained to predict the 'high' and 'low immune' ground truth labels [1]. This association further supports the assertion that the patient tumours in the predicted high and low immune subgroups have differing immune microenvironment compositions, which contributes to their prognostic value. Moreover, the TIL classifications shown in Table 6.13, show that the predicted high immune subgroup is comprised of a higher proportion (19%) of tumours labelled as Brisk (diffuse TILs throughout the tumour) compared to the

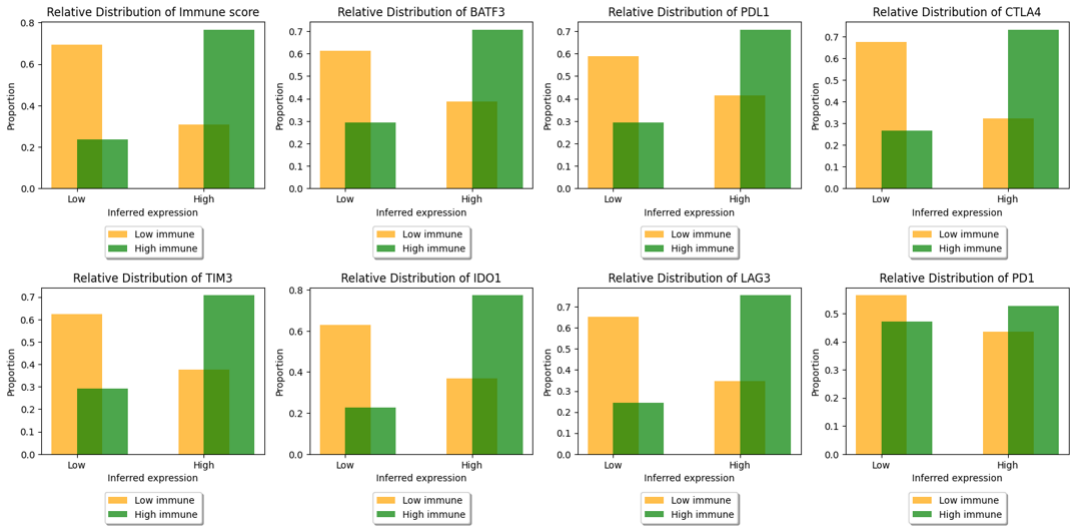


Figure 6.9: Relative distributions of potential treatments targets and immune biomarkers for patients in the high and low immune classes as predicted by the GNN attention 40x (3 GATv2) model. GATv2, Graph Attention Network version 2; GNN, Graph Neural Network.

low immune patient tumours (4%). Furthermore, the predicted high immune tumours have a lower proportion (7%) of tumours classified as absent compared to the predicted low immune patient tumours (16%). The differences in the brisk vs. absent classification proportions in the predicted immune subgroups are even greater than those reported for the immune subgroups found in the paper published by Pozniak *et al.* (2019) [1] (Chapter 3, Table 3.1). For the ground truth immune subgroups the proportion of brisk TILs in the 'high immune' group is 13% and 12% in the 'low immune' group, whereas the proportion of patients with tumours classified as absent is 10% in 'high immune' and 14% in the 'low immune' subgroup. Moreover, a large proportion of the LMC cases have not been classified for TILs and as TIL classification is a subjective measurement carried out by a pathologist, it is possible that the difference in absent vs. brisk TILs may be even greater than reported here. Nevertheless, the increases in proportion of brisk TILs in the 40x attention GNN model high immune subgroup compared to the ground truth 'high immune' patient group, supports our hypothesis from Chapter 4 and Chapter 5 that the attention GNN model is classifying patients based on a simplified version of the Pozniak *et al.* (2019) expression signature, related to the presence or absence of brisk TILs. This also might be the reason why these new predicted immune subgroups remains more prognostic, as the presence of brisk TILs have been shown to be prognostic [224].

Table 6.12: Summary data of the clinical and clinicohistological features of the high and low risk LMC patients from the 10x SurvGNN model. AJCC; American Joint Committee on Cancer; LMC, Leeds Melanoma Cohort; SurvGNN, Survival Graph Neural Network; TIL, Tumour Infiltrating Lymphocyte.

	Low risk	High risk
Number of tumours	465	203
Melanoma death (%)	21	46
Age at diagnosis (median, years)	57	61
Sex (% males)	44	52
Breslow thickness (median, mm)	2.1	3.0
Ulcerated (%)	28.0	46.8
Mitotic rate (median count, mm^2)	4	5
AJCC stage (%)	Low risk	High risk
I	37	18
II	50	54
III	13	25
Unknown	1	3
TIL classification (%)	Low risk	High risk
Brisk	11	9
Non-brisk	49	47
Absent	11	17
Unclassified	29	17
Immune subgroup [1] (%)	Low risk	High risk
Low	36	45
Intermediate	40	37
High	24	18
Histological subtype (%)	Low risk	High risk
Superficial spreading	53	38
Nodular	30	38
Lentigo maligna	1	1
Acral lentiginous	3	5
Other	5	9
Unclassified	7	9

6.4 Discussion

Overall our results show that deep learning models are promising tools for for the development of new risk groups and prognostic biomarkers associated with melanoma specific survival. While the risk groups derived from the models, are not sufficient to replace the AJCC 7^{th} edition staging groups, we show that using 10x WSI high vs. low risk groups (Table 6.4) adds prognostic value for predicting MSS for the patients from LMC dataset (Table 6.11). Moreover, when combining these SurvGNN risk groups, with the immune subgroups classified by the 40x GNN attention model (from Chapter 5) and other clinical variables such as AJCC staging, age at diagnosis and tumour site in a multivariate CPH model, we show that this model fits the LMC data the best with the lowest AIC value and has the joint best discriminative power for survival

Table 6.13: Summary data of the clinical and clinicohistological features of the high immune and low immune LMC patients (including the intermediate subset of patients) classified by the 40x attention GNN model described in Chapter 5. AJCC; American Joint Committee on Cancer; GNN, Graph Neural Network; LMC, Leeds Melanoma Cohort; TIL, Tumour Infiltrating Lymphocyte

	Predicted low immune	Predicted high immune
Number of tumours	395	273
Melanoma death (%)	36	18
Age at diagnosis (median, years)	58	58
Sex (% males)	44	51
Breslow thickness (median, mm)	2.5	2.2
Ulcerated (%)	37.5	28.2
Mitotic rate (median count, mm^2)	< 1	< 1
AJCC stage (%)	Predicted low immune	Predicted high immune
I	28	35
II	52	50
III	18	14
Unknown	2	1
TIL classification (%)	Predicted low immune	Predicted high immune
Brisk	4	19
Non-brisk	49	47
Absent	16	7
Unclassified	30	27
Immune subgroup [1] (%)	Predicted low immune	Predicted high immune
Low	56	14
Intermediate	39	38
High	5	48
Histological subtype (%)	Predicted low immune	Predicted high immune
Superficial spreading	47	50
Nodular	32	34
Lentigo maligna	1	2
Acral lentiginous	4	3
Other	8	4
Unclassified	9	6

times (joint highest C-index with the 20x 3 risk group model and 10x small model).

We also demonstrate that the immune subgroups classified by the 40x GNN attention models combined with other covariates render the immune subgroupings derived by Pozniak *et al.* no longer significant [1]. Showing that along with the other clinical variables, the genetics derived immune subgroups generated by the GNN attention model are sufficient to replace the subgroups generated by Pozniak *et al.* (2019), with the model that used the GNN attention subgroups instead also generating a lower AIC and high C-index (Fig. 6.7). Meanwhile, we add to the evidence from our previous analysis of high attention regions of the GNN attention 40x model, which suggested that the model may be classifying high and low immune subgroups

based on the presence of TILs (Chapter 5, Fig. 5.6), by examining the TIL classifications of these predicted subgroups (Table 6.13). The original immune signatures by Pozniak *et al.* (2019) were based on the inferred abundance of 27 immune cells [1], which is difficult to capture using H&E staining - as we are unable to distinguish the types of immune cells clearly without using more advanced staining techniques, such as immunohistochemistry staining. Furthermore, the GNN attention 40x model categorises subgroups in a binary manner, disregarding the 'intermediate' immune subset among patients. Instead, it assigns these patients to either the 'high' or 'low' immune subsets. This simplification of the immune signature from Pozniak *et al.* (2019) [1] might maintain significance in the CPH by eliminating 'noise' generated by the spatial bias of the ground truth labels. For instance, during the analysis of specific misclassified WSIs labeled as 'low immune,' it was observed that regions with high attention patches contained lymphocytic infiltration, despite a 'low immune' ground truth label indicating a low abundance or absence of immune cells (shown in Chapter 5, Fig. 5.6). The reason for these unexpected 'high immune' regions in 'low immune' labelled images could be due to the immune subgroup labels being generated from a small 0.6mm needle biopsy region, where the genetic data was extracted from. Whereas the classifications from the 40x GNN attention models are derived from the entire tissue from the WSI image and are based on the WSI graph representations, so include greater context compared to the transcriptome generated labels. Nevertheless, as the immune labels were required to train this model, they still give valuable information and are indicative of potential treatment options for patients. While we see from analysis that the two subgroups show different presence of potential treatment targets, as shown in Fig 6.9 we would need to carry out further analysis to have better indication.

Works by Saltz *et al.* (2018) and Zormpas-Petridis *et al.* (2019) have shown that multiple resolutions generate better prognostic features for predicting the survival of melanoma patients, for our analysis the novel multi-level representations did not improve results over single resolution graph representations [171, 173]. One reason for this may be that these previous works explored handcrafted features through cell detection, to generate TIL scores [171] or ratios [173] and were 'supervised' in the sense of detecting specific tissues or cellular features at different resolutions. Whereas our approach relies on patch based features, rather than specific biological entities. Moreover, as discussed in the previous Chapter 5, our use of a graph mean pooling mechanism to generate patient-level risk scores, may mean that noisy instances from higher resolutions with more patch nodes may reduce the prognostic signal of informative patches from lower resolutions, preventing the multi-level graphs from generating prognostic representations. In addition if different regions or biological entities at different resolutions are prognostic, it may be better to use an attention mechanism and pooling at each resolution separately, to capture the most informative node features in each level of the WSI graph representation.

Nevertheless, our step-wise multivariate CPH model (Fig. 6.7) showed that there appears to be strong prognostic features at both 10x and 40x resolutions, with the risk groups derived from 10x WSI graph representations and immune subgroups derived from 40x WSI graph

representations remaining significant. Our previous high attention patch analysis from the graph attention models (Chapter 5, Fig. 5.6) and analysis of the relative distributions of potential treatments targets, immune biomarkers and TIL classifications (Fig. 6.9 and Table 6.13) suggest that the GNN attention 40x model immune subgroups are based on the differing presence of immune cells. On the other hand, it is not clear what or if there are biological correlates underpinning the different survival outcomes of the SurvGNN 10x (small) high and low risk subgroups. Through examination of the distribution of clinical and histopathological features of the two risk groups, we found there were differing levels of ulcerated tumours between the two groups (Table 6.12). This is similar to findings from Couetil *et al.* (2023), who used machine learning methods to identify high risk patients from melanoma WSIs [253]. Their study found that tumours which contained dense regions of Small-Hyperchromatic cells had worse prognosis [253]. These cells coincided with tumors that had more ulceration, necrosis and pockets of dense inflammatory cell infiltration, suggesting that the increased ulceration in our high risk group could be related to other prognostic features. Moreover, the small differences in relative distributions of the immune biomarkers in Fig. 6.8, suggests that the difference in survival in risk groups is unlikely to be caused by differences in immune expression. Nevertheless, further analysis of our results is required to determine if there are morphological signatures underlying these subgroups, ruling out any potential confounding factors that could be leading to difference in survival distributions.

While incorporating the attention mechanism improved C-index and increased the HR size for the models using larger graph inputs (40x and multi-level), the risk group stratification of the 10x SurvGNN model with reduced hidden layer sizes was superior to the attention-based SurvGNN models. Moreover, inspection of the high attention feature patches of the low risk patient group, revealed that these regions often contained minimal prognostic information (Fig. 6.5A). This may suggest that attention models are overfitting to noise within the images. Previous studies have mitigated this issue by employing pre-trained tumour detection networks to select only patches from tumour [253] or high cell density [176, 251] regions as inputs. However, training such models typically requires pathologist annotations, which are time-consuming, particularly in understaffed pathology departments [18], and may overlook prognostic features outside the tumour area. An alternative approach to enhance feature diversity for lower-risk patients could be using dropout within the attention layer of the network. Despite this, the high attention regions in Fig. 6.5B, suggests the potential of these networks to identify new biomarkers such as tumour stroma ratio, although further work is necessary to improve the performance of these models and validate these findings.

Whilst the results we present are promising and some of the first on a large melanoma dataset with a large early-stage subset of patients, our use of only the LMC dataset is a large limitation of this work. Although we show that the risk groups generated from the SurvGNN model risk scores have predictive power, it's possible for that the models we present are able to perform well on the LMC dataset but not able to generalise well to new, unseen data. Therefore in order

to validate the risk groups we found, we would need to test the cut-points we determined on a large independent melanoma dataset to confirm the prognostic value of the risk groups. This is a difficult task, as data collected after 2012 is likely to include patients who have been treated with immunotherapy [163], making it difficult to determine the biological effects of the tumour on survival separate from the treatment intervention. In order to fully determine the biological entities or immune response that is contributing to improved survival, there is a need to collect historic datasets that are not impacted by immunotherapy treatments. In addition we would need to explore how to quantify and assess the causality of the SurvGNN model biomarkers, for example finding suitable ways to quantify the tumour stroma ratio shown in Fig 6.5 to see if does differ between high and low risk patients. In addition, for this work, we only used a subset of the available LMC images (n=667) that had the immune subgroup ground-truth labels [1]. In the future we hope to use all available 1115 images which have corresponding survival data for training and use the subset which have immune subgroup labels for the testing, so we are still able to compare our risk groups to the immune subgroups [1], but have a larger training dataset.

6.4.1 Conclusion

In summary, our results show how risk groups generated using WSI graph representations can add prognostic value beyond the established clinical variables associated with melanoma survival. Moreover, we show that the immune subgroups generated by the GNN attention models from the previous Chapter are complementary to our SurvGNN risk groups presented in this chapter and are sufficient to represent the immune labels from Pozniak *et al.* (2019) [1]. In future studies we hope to validate our findings on a large independent dataset and determine the biological correlates of the SurvGNN risk groups.

Chapter 7

Conclusions and future work

In this chapter we will summarise the contributions of this thesis and potential limitations that should be investigated further in future work.

7.1 Summary and thesis contributions

The work in this thesis aimed to investigate the ability of deep learning frameworks for stratifying melanoma patients into existing or new prognostic subgroups. Our studies mainly focused on patients from the LMC dataset, which is one of the largest population-based melanoma datasets containing imaging data for melanoma primaries with long follow-up of patients [189], as we described in Chapter 3. Chapter 4 and Chapter 5 examined whether image-based tools could be used to classify patients into immune subgroups [1] which were determined using tumour tissue derived transcriptomic data. These subgroups have differing survival outcomes and different potential treatment targets, so could be used to identify at risk patients and aid in selecting effective stratified treatments [1]. While MIL methods were able to identify patients from the 'high' and 'low immune' classes with a mean AUC of 0.80, the 'intermediate immune' class was more difficult to classify. We postulate that 'intermediate immune' class is challenging to classify because of its heterogeneity. In Chapter 5, we found that using GNN models, with graph-based WSI representations further improved classification of three immune subgroups, including the 'intermediate immune' subset of patients. With the attention GNN model classifying WSIs from 'high' immune vs 'low immune' status patients with the highest mean AUC of 0.81 demonstrating that image-based deep learning pipelines show potential for subtyping patients based on immune expression signatures. Finally, in Chapter 6, we examined whether neural networks trained using MSS data could be utilised to stratify melanoma patients into risk groups. Here we found that risk groups generated using graph WSI representations of the LMC patients, added significant prognostic value for predicting MSS when included with established prognostic factors in multi-variate analyses. Below is a more detailed description of

our thesis contributions.

7.1.1 The importance of resolution, feature extraction method attention for classifying immune subtypes

In Chapter 4, our main aims were to examine how image resolution, feature extraction method and type of MIL model affected our task of classifying WSIs into immune subgroups. We found that models trained with 10x patches performed better than models trained with 40x, 20x, 5x or 2.5x resolution patches, which we hypothesised was due to 10x patches having a greater field of view including more immune microenvironment and spatial context. The pathology specific feature extraction method [125], generated the best performance in most circumstances and more recent works following our own work [114, 267] have further highlighted how SSL domain-specific methods consistently out-perform supervised pre-training on ImageNet. Finally, we found that although attention-based methods outperformed max-pool MIL, there was no clear superior model out of the five attention-based methods we used for our task. Yet we found that by using multiple metrics including HRs can help in discriminating a model's abilities, assessing the prognostic ability of models as well as the classification performance.

7.1.2 Context improves immune subtype classifications of melanoma whole slide images

In Chapter 5, our aim was to experiment with graph representations of WSIs to examine how additional context affected the classification of melanoma WSIs into the immune subgroups. We found that GNNs trained with the graph representations improved performance beyond the MIL methods used in Chapter 4. In addition we again demonstrated the importance of attention mechanisms, as the best AUC performance for classifying 'high' and 'low immune' classes, was generated by a model which utilised 40x WSI representations with an attention mechanism. Here we also showed how three or more GATv2 layers leads to superior GNN model performance.

7.1.3 Melanoma whole slide images contain prognostic information beyond the established prognostic factors

In Chapter 6 we examined how neural networks can be applied to whole slide images to stratify patients into *de novo* risk groups based on their expected melanoma specific survival using the well-curated LMC dataset. Consistent with the findings of Chapter 4, we found that models trained with 10x graph representations of WSIs, generated risk groups with the greatest prognostic ability. We also experimented with reducing the size of the SurvGNN models and found that reduction in hidden layer size can improve model performance, by reducing model complexity and overfitting. Moreover by jointly modelling the subgroups from the 40x attention

GNN we presented in Chapter 5 (which were from models developed to discriminate between the low and high immune subgroups), with the SurvGNN 10x model *de novo* risk groups, we showed that they improve survival prediction of patients from the LMC datasets, over using the immune subgroups derived by Pozniak *et al.* (2019) [1].

Overall this thesis shows that context, whether that be from patches at a 10x resolution or through graph representations, generates better model performance for classifying melanoma WSIs. In addition our experiments highlight the strength of attention-based models, but our SurvGNN model experiments also highlight that while GNN and attenion-based models perform well, we still need to work on effective multi-resolution attention-based aggregation methods and find ways to mitigate for model complexity and overfitting.

7.2 Limitations and future directions

7.2.1 Immediate limitations and future directions specific to this work

While our results for classifying the immune subgroups were promising, a challenge for this work was the use of transcriptomic derived labels as ground truth labels, were from small sections (0.6mm needle biopsy) of the tissue shown in the H&E stained slides. This meant that only a small subset of patches, from all the patches we used as an input for our models came from the tumour region that generated the slide-level label. In Chapter 4, we demonstrated that models were more likely to misclassify cases where tissue was over-stained or lymph nodes where there was an abundance of immune cells throughout the tissue not just localised in the tumour. This indicates that the model may be performing shortcut learning by conducting the simplified task of classifying images based on the presence of any immune cells or dark overly-stained regions of tissue, rather than classifying images based on the immune cell subtypes in the region, the score of which was used to generate the immune labels [1]. In addition, due to contamination when sampling the transcriptomic data used to generate those labels, there may be some labelling errors as described in Chapter 4. Together, this shortcut learning combined with noisy labels could mean there is a performance barrier, with classification accuracy for 'high' and 'low immune' subtypes remaining at a mean AUC of around 0.80. One method we could explore in the future to combat mislabelling errors due to contamination could be label denoising methods, as described in the Prostate cANcer graDe Assessment (PANDA) challenge [268], where teams were working to classify WSIs based on subjective pathologist gradings. We also believe that further refinement of our graph-based methods which are more 'context aware' could improve erroneous classification based on scattered immune cells outside of the tumour region.

In Chapter 4 and 5, improving model classification performance for the 'intermediate subgroup' was a significant challenge. Although the models using graph representations of the WSIs, showed improved ability for classifying the 'intermediate immune' subgroup (mean AUC of 0.63 [95% CI 0.61 to 0.65]) compared to the best MIL model (mean AUC of 0.61 [95% CI

0.57–0.65]), we could not achieve an AUC which would be clinical useful. In the future we hope to explore features with higher acuity, for example exploring how foundation models trained with SSL and large diverse histological datasets (as described in more detail in the following section 7.2.2) could generate 'intermediate immune' features which are more discernible from the other classes. In addition, as discussed in Chapter 4, the immune subgroups are simplification of a what is a continuum of patients, therefore a weakly supervised regression [110] method may avoid models performing shortcut learning and improve modelling of the immune subtypes, especially due to overlap in the 'low' and 'intermediate' subgroups. Future studies could focus on whether an ordinal loss function, which could rank the subtypes, penalising misclassification between 'high' and 'intermediate' or 'low' and 'intermediate' less could improve determination of the 'intermediate immune' subgroup.

Other works have also shown that multi-resolution graph models outperform single resolution representations [133, 177, 269, 270, 271]. In the results we present, multi-resolution graphs generate similar performance to single resolution, so we intend to explore how to improve on this. In 2024 Mirabadi *et al.* constructed graph representations of ovarian and bladder cancer WSIs, where each resolution (5x, 10x, 20x) of the multi-resolution graphs had an equal number of nodes. They then used a three layer GNN and averaged nodes across three magnifications to get a graph or slide-level readout, finding their multi-resolution graph representations generated superior performance over single resolution graphs for tumour subtyping [271]. As the graph representations we experimented with in Chapter 5 contained a greater amount of 40x resolution nodes, the final readout from out models may be biased towards 40x nodes not benefiting from all three resolutions. Moreover Hou *et al.* (2022) presented an "Iterative Hierarchical Pooling module", which pools the nodes of individual magnifications after each "resolution-aware attention convolution" layer, before aggregating the readouts of each convolution layer together [133]. For our work we utilised a mean pooling mechanism and attention mechanism for all nodes, so developing a separate resolution pooling aggregation mechanism may improve our results for multi-resolution graphs avoiding noise from features at different scales. In addition as we noted in the conclusion of Chapter 5, there is a growing trend for combining elements of vision transformer networks with GNN backbones to process graph-based WSI representations can improve model performance, leveraging both local and global context to model classifications [237, 238]. This could be an important potential direction, and is in agreement with our results from Chapter 4 and Chapter 5 which showed that both TransMIL (transformer-based method) and the graph-based models generated high performance for immune subtype classifications.

The results from Chapter 6 suggest that the survival models we implemented derive additional prognostic value from the WSIs, separate to the 7th edition AJCC staging system [38] and the Pozniak et al. (2019) immune subgroups [1]. Further research could explore refinement of the imaging biomarkers derived when stratifying patients using the survival models and testing the robustness of our cut-off points using large independent datasets. This could involve experimenting with more sophisticated methods to derive cut-off points [272, 273, 274].

7.2.2 Limitations and future directions for wider digital pathology research applied to melanoma

The experiments presented within this thesis were formulated and carried out between October 2020 to October 2023 (when funding for this PhD project ended). During and after this period deep learning methodologies have continued to develop and advance at a rapid rate. Notably, in the field of computational pathology there has been the rise of pathology specific 'foundation models' - large vision models usually trained using SSL on large histopathology datasets for application on different downstream tasks. While this thesis touches on SSL pathology-specific models, with our use of the Ciga *et al.* (2021) [125] SSL ResNet18 feature extraction model in Chapter 4, it is worth discussing the advancements in models in this field beyond our work. Recent studies have applied pathology-specific SSL to train transformer-based frameworks, such as Hierarchical Image Pyramid Transformer (HIPT) [207], TransPath [275] and iBOT-Path [276] models. Moreover, networks are now trained with image-text contrastive learning, such as BioMedCLIP [277] and PLIP [278] and Vorontsov *et al.* (2024) developed a foundation model trained on 1.5 million internal WSIs [279], by leveraging SSL with the DINOv2 algorithm [280]. A difficulty for us, was the restricted time limit for carrying out our experiments, because although testing newer SSL vision transformer feature extraction architectures might help us to achieve SOTA performance, feature extraction is one of the earlier and more time consuming steps in our model pipelines. So testing newer feature extraction methods, while extremely valuable, would have taken time from carrying out other essential experiments so we determined that this might not be the best use of our limited time, however, we anticipate that future research will implement these promising newer methods. Moreover it is worth noting that our findings that a pathology-specific feature extraction method is an improvement on pathology agnostic feature extraction method (pre-training using ImageNet), is still likely to stand and has also been shown in a recent comprehensive SSL benchmarking study by Kang *et al.* (2023) [114].

Another limitation of the pipelines presented in this thesis is that they do not use image augmentation. Image augmentation is a technique where images can be altered using transformations [281] such as rotations, translations, flipping or cropping [282, 283] to increase the diversity and size of a dataset. This can help to mitigate model overfitting and improve the generalisability and robustness of a network. Here, data augmentation could be useful the survival analysis pipeline, we presented in Chapter 6, where model complexity appeared to cause overfitting (hence smaller models performing better). Lack of augmentation is not only an issue for the works presented in this thesis, it was also noted in a 2023 review of 54 studies applying deep learning to melanoma histopathology images, with only 50% of the studies included in the review citing the use of data augmentation [169]. One reason for the scarcity of studies carrying out image augmentation could be that many of these papers use features generated from pre-trained CNNs, consequently if image augmentation is carried out in the traditional pixel-space of images, feature extraction would need to occur every time a new augmentation is

required which is time consuming and computationally demanding. One potential method for generating augmented training examples is *mixup*, which generates augmented features through the weighted averaging of two random feature pairs from the training set [284, 285]. However, carrying out data augmentation in the embedding space is not straightforward as transforming an embedding may generate unrealistic or non-salient representations and traditional image augmentation strategies may work for natural images, but histological images have very different features. Zaffar *et al.* (2022), attempted to combat this by proposing *EmbAugmenter*, a generative adversarial network to synthesise realistic augmentations in the embedding space, improving accuracy for the grading of prostate biopsies [286]. In the future, we hope to apply these embedding augmentation strategies and also test if there are ways to generate salient features, by carrying out domain-specific augmentations of the embedding space to reduce overfitting of our models. Moreover, subsequent research might consider graph augmentation techniques such as node and edge drop-out to determine if this also improves model performance.

In a recent publication, Reinke *et al.* (2023) highlighted that a significant barrier to the adoption of AI algorithms in clinical settings is the improper application or misinterpretation of validation metrics in image analysis [287]. Although the AI algorithms and pipelines presented in this thesis could not be implemented in a clinical setting, before there are improvements to the classification performance and further validation on large external datasets, we did aim to test our models using multiple metrics to better understand the clinical utility of the models we developed for our specific tasks. However, there are still many metrics and analyses that were out of the scope of this thesis, but could be explored in future work such as cost-benefit analysis to understand which misclassifications could lead to better or worse clinical utility, or using expected cost to account for the ordinal grading of the immune subgroups.

Another future research direction worth considering is the use of deep learning methods to integrate imaging, genetic and clinical data to see if a 'precision medicine approach', which reflects the genetic diversity of patients and encompasses histological features which are overlooked in the current AJCC staging system, can generate a superior prognostic tool. Peng *et al.* (2020) used a combination of WSI features with clinicopathologic variables and mutation status of BRAF and NRAS genes to predict recurrence-free survival of later stage melanoma patients, using the publicly available TCGA dataset [195]. Subsequently Li *et al.* (2023), explored combining features associated with prognosis-related genes from TCGA cases, with features from WSIs extracted using CellProfiler software as inputs for their pathomic binary prediction model [288]. Their results were promising, with their pathomic model achieving an accuracy of 0.97 for predicting patient survival status (whether a patient was dead or alive), but this did not beat the model using image features alone which achieved an accuracy of 0.99 [288]. The lack of increased performance with genetic features, suggests there could be a correlation between genetic and imaging features, and understanding what molecular data can be recovered from WSIs is an area of research that needs to be more understood before we can fully exploit multimodal inputs. Both of these studies focus of later AJCC stage patients, which only account for

around 20% of melanoma patients, therefore, there is still a necessity to determine prognostic markers in a large cohort of early-stage patients, as this has the potential to improve patient triaging and transform melanoma clinical care.

7.3 Conclusion

In this chapter we have summarised the main contributions of this thesis, demonstrating that attention, context, resolution and feature extraction techniques all affect model performance and should be considered when developing WSI analysis pipelines. We also add to the evidence that deep learning when applied to digital pathology images can be used to derive image based signatures, that have prognostic value beyond the current staging tools used in the treatment planning of melanoma patients. Finally we examine the limitations of this work, positioning it within the existing research landscape and outlining potential avenues for future developments.

Bibliography

[1] Joanna Poźniak, Jérémie Nsengimana, Jonathan P. Laye, Sally J. O'Shea, Joey Mark S. Diaz, Alastair P. Droop, Anastasia Filia, Mark Harland, John R. Davies, Tracey Mell, Juliette A. Randerson-Moor, Sathya Muralidhar, Sabrina A. Hogan, Sandra Nicole Freiberger, Mitchell P. Levesque, Graham P. Cook, D. Timothy Bishop, and Julia Newton-Bishop. Genetic and Environmental Determinants of Immune Response to Cutaneous Melanoma. *Cancer Research*, 79(10):2684–2696, May 2019.

[2] NHS. Melanoma skin cancer, Section: conditions nhs, October Last accessed: 2023-06-18.

[3] Cancer Research UK. Melanoma skin cancer statistics, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer, Last accessed: 2024-07-04.

[4] Caroline Robert, Boguslawa Karaszewska, Jacob Schachter, Piotr Rutkowski, Andrzej Mackiewicz, Daniil Stroiakovski, Michael Lichinitser, Reinhard Dummer, Florent Grange, Laurent Mortier, Vanna Chiarion-Sileni, Kamil Drucis, Ivana Krajsova, Axel Hauschild, Paul Lorigan, Pascal Wolter, Georgina V. Long, Keith Flaherty, Paul Nathan, Antoni Ribas, Anne-Marie Martin, Peng Sun, Wendy Crist, Jeff Legos, Stephen D. Rubin, Shonda M. Little, and Dirk Schadendorf. Improved overall survival in melanoma with combined dabrafenib and trametinib. *The New England Journal of Medicine*, 372(1):30–39, January 2015.

[5] Selma Ugurel, Joachim Röhmel, Paolo A. Ascierto, Keith T. Flaherty, Jean Jacques Grob, Axel Hauschild, James Larkin, Georgina V. Long, Paul Lorigan, Grant A. McArthur, Antoni Ribas, Caroline Robert, Dirk Schadendorf, and Claus Garbe. Survival of patients with advanced metastatic melanoma: The impact of novel therapies. *European Journal of Cancer (Oxford, England: 1990)*, 53:125–134, January 2016.

[6] Jedd D. Wolchok, Vanna Chiarion-Sileni, Rene Gonzalez, Jean-Jacques Grob, Piotr Rutkowski, Christopher D. Lao, C. Lance Cowey, Dirk Schadendorf, John Wagstaff, Reinhard Dummer, Pier Francesco Ferrucci, Michael Smylie, Marcus O. Butler, Andrew Hill, Ivan Márquez-Rodas, John B. A. G. Haanen, Massimo Guidoboni, Michele Maio,

Patrick Schöffski, Matteo S. Carlino, Céleste Lebbé, Grant McArthur, Paolo A. Ascierto, Gregory A. Daniels, Georgina V. Long, Tuba Bas, Corey Ritchings, James Larkin, and F. Stephen Hodi. Long-Term Outcomes With Nivolumab Plus Ipilimumab or Nivolumab Alone Versus Ipilimumab in Patients With Advanced Melanoma. *Journal of Clinical Oncology*, page JCO.21.02229, November 2021. Publisher: Wolters Kluwer.

[7] Alexander C. Huang and Roberta Zappasodi. A decade of checkpoint blockade immunotherapy in melanoma: understanding the molecular basis for immune sensitivity and resistance. *Nature Immunology*, 23(5):660–670, May 2022.

[8] Lena A. von Schuckmann, Maria Celia B. Hughes, Reza Ghiasvand, Maryrose Malt, Jolieke C. van der Pols, Vanessa L. Beesley, Kiarash Khosrotehrani, B. Mark Smithers, and Adele C. Green. Risk of Melanoma Recurrence After Diagnosis of a High-Risk Primary Tumor. *JAMA dermatology*, 155(6):688–693, June 2019.

[9] Göran Jönsson, Christian Busch, Stian Knappskog, Jürgen Geisler, Hrvoje Miletic, Markus Ringnér, Johan R. Lillehaug, Åke Borg, and Per Eystein Lønning. Gene Expression Profiling–Based Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome. *Clinical Cancer Research*, 16(13):3356–3367, July 2010. Publisher: American Association for Cancer Research Section: Human Cancer Biology.

[10] Katja Harbst, Johan Staaf, Martin Lauss, Anna Karlsson, Anna Måsbäck, Iva Johansson, Pär-Ola Bendahl, Johan Vallon-Christersson, Therese Törngren, Henrik Ekedahl, Jürgen Geisler, Mattias Höglund, Markus Ringnér, Lotta Lundgren, Karin Jirström, Håkan Olsson, Christian Ingvar, Åke Borg, Hensin Tsao, and Göran Jönsson. Molecular Profiling Reveals Low- and High-Grade Forms of Primary Melanoma. *Clinical Cancer Research*, 18(15):4026–4036, August 2012. Publisher: American Association for Cancer Research Section: Human Cancer Biology.

[11] Jérémie Nsengimana, Jon Laye, Anastasia Filia, Christy Walker, Rosalyn Jewell, Joost J Van den Oord, Pascal Wolter, Poulam Patel, Antje Sucker, Dirk Schadendorf, Göran B Jönsson, D. Timothy Bishop, and Julia Newton-Bishop. Independent replication of a melanoma subtype gene signature and evaluation of its prognostic value and biological correlates in a population cohort. *Oncotarget*, 6(13):11683–11693, March 2015.

[12] Pedram Gerami, Robert W. Cook, Jeff Wilkinson, Maria C. Russell, Navneet Dhillon, Rodabe N. Amaria, Rene Gonzalez, Stephen Lyle, Clare E. Johnson, Kristen M. Oelschlager, Gilchrist L. Jackson, Anthony J. Greisinger, Derek Maetzold, Keith A. Delman, David H. Lawson, and John F. Stone. Development of a Prognostic Genetic Signature to Predict the Metastatic Risk Associated with Cutaneous Melanoma. *Clinical Cancer Research*, 21(1):175–183, January 2015. Publisher: American Association for Cancer Research Section: Biology of Human Tumors.

[13] Laura K. Ferris, Aaron S. Farberg, Brooke Middlebrook, Clare E. Johnson, Natalie Lassen, Kristen M. Oelschlager, Derek J. Maetzold, Robert W. Cook, Darrell S. Rigel, and Pedram Gerami. Identification of high-risk cutaneous melanoma tumors is improved when combining the online American Joint Committee on Cancer Individualized Melanoma Patient Outcome Prediction Tool with a 31-gene expression profile–based classification. *Journal of the American Academy of Dermatology*, 76(5):818–825.e3, May 2017. Publisher: Elsevier.

[14] Mihaela Angelova, Pornpimol Charoentong, Hubert Hackl, Maria L. Fischer, Rene Snajder, Anne M. Krogsdam, Maximilian J. Waldner, Gabriela Bindea, Bernhard Mlecnik, Jerome Galon, and Zlatko Trajanoski. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biology*, 16(1):64, March 2015.

[15] Nandini Acharya, Catherine Sabatos-Peyton, and Ana Carrizosa Anderson. Tim-3 finds its place in the cancer immunotherapy landscape. *Journal for Immunotherapy of Cancer*, 8(1):e000911, June 2020.

[16] Julie Westerlin Kjeldsen, Cathrine Lund Lorentzen, Evelina Martinenaite, Eva Ellebaek, Marco Donia, Rikke Boedker Holmstroem, Tobias Wirenfeldt Klausen, Cecilie Oelvang Madsen, Shamaila Munir Ahmed, Stine Emilie Weis-Banke, Morten Orebo Holmström, Helle Westergren Hendel, Eva Ehrnrooth, Mai-Britt Zocca, Ayako Wakatsuki Pedersen, Mads Hald Andersen, and Inge Marie Svane. A phase 1/2 trial of an immune-modulatory vaccine against IDO/PD-L1 in combination with nivolumab in metastatic melanoma. *Nature Medicine*, 27(12):2212–2223, December 2021. Number: 12 Publisher: Nature Publishing Group.

[17] Anand Rotte. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. *Journal of Experimental & Clinical Cancer Research*, 38(1):255, June 2019.

[18] The Royal College of Pathologists. Meeting pathology demand: Histopathology workforce census, 2018.

[19] Alec J. Kacew, Garth W. Strohbehn, Loren Saulsberry, Neda Laiteerapong, Nicole A. Cipriani, Jakob N. Kather, and Alexander T. Pearson. Artificial Intelligence Can Cut Costs While Maintaining Accuracy in Colorectal Cancer Genotyping. *Frontiers in Oncology*, 11, 2021.

[20] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, October 2018. Number: 10 Publisher: Nature Publishing Group.

[21] Kun-Hsing Yu, Feiran Wang, Gerald J Berry, Christopher Ré, Russ B Altman, Michael Snyder, and Isaac S Kohane. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. *Journal of the American Medical Informatics Association : JAMIA*, 27(5):757–769, May 2020.

[22] Gabriele Campanella, David Ho, Ida Häggström, Anton S. Becker, Jason Chang, Chad Vanderbilt, and Thomas J. Fuchs. H&E-based Computational Biomarker Enables Universal EGFR Screening for Lung Adenocarcinoma, June 2022. arXiv:2206.10573 [cs, q-bio].

[23] Heather D. Couture, Lindsay A. Williams, Joseph Geradts, Sarah J. Nyante, Ebonee N. Butler, J. S. Marron, Charles M. Perou, Melissa A. Troester, and Marc Niethammer. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ breast cancer*, 4:30, 2018.

[24] Wenqi Lu, Michael Toss, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. SlideGraph+: Whole Slide Image Level Graphs to Predict HER2Status in Breast Cancer, October 2021. arXiv:2110.06042 [cs].

[25] Hui Qu, Mu Zhou, Zhennan Yan, He Wang, Vinod K. Rustgi, Shaoting Zhang, Olivier Gevaert, and Dimitris N. Metaxas. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *npj Precision Oncology*, 5(1):1–11, September 2021. Number: 1 Publisher: Nature Publishing Group.

[26] Rishi R. Rawat, Itzel Ortega, Preeyam Roy, Fei Sha, Darryl Shibata, Daniel Ruderman, and David B. Agus. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. *Scientific Reports*, 10(1):7275, April 2020. Number: 1 Publisher: Nature Publishing Group.

[27] Hongming Xu, Sunho Park, Sung Hak Lee, and Tae Hyun Hwang. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients, February 2019. Pages: 554527 Section: New Results.

[28] Mohammad Rizwan Alam, Jamshid Abdul-Ghafar, Kwangil Yim, Nishant Thakur, Sung Hak Lee, Hyun-Jong Jang, Chan Kwon Jung, and Yosep Chong. Recent Applications of Artificial Intelligence from Histopathologic Image-Based Prediction of Microsatellite Instability in Solid Cancers: A Systematic Review. *Cancers*, 14(11):2590, May 2022.

[29] Jack Breen, Katie Allen, Kieran Zucker, Pratik Adusumilli, Andrew Scarsbrook, Geoff Hall, Nicolas M. Orsi, and Nishant Ravikumar. Artificial intelligence in ovarian cancer histopathology: a systematic review. *npj Precision Oncology*, 7(1):1–14, August 2023. Publisher: Nature Publishing Group.

[30] Lindsey A. Hildebrand, Colin J. Pierce, Michael Dennis, Munizay Paracha, and Asaf Maoz. Artificial Intelligence for Histology-Based Detection of Microsatellite Instability

and Prediction of Response to Immunotherapy in Colorectal Cancer. *Cancers*, 13(3):391, January 2021.

[31] Bangwei Guo, Xingyu Li, Miaomiao Yang, Jitendra Jonnagaddala, Hong Zhang, and Xu Steven Xu. Predicting microsatellite instability and key biomarkers in colorectal cancer from H&E-stained images: achieving state-of-the-art predictive performance with fewer data using Swin Transformer. *The Journal of Pathology: Clinical Research*, 9(3):223–235, February 2023.

[32] Ji Hyun Park, Eun Young Kim, Claudio Luchini, Albino Eccher, Kalthoum Tizaoui, Jae Il Shin, and Beom Jin Lim. Artificial Intelligence for Predicting Microsatellite Instability Based on Tumor Histomorphology: A Systematic Review. *International Journal of Molecular Sciences*, 23(5):2462, February 2022.

[33] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S. Hveem, Knut Liestøl, John Maddison, Hanne A. Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, Inger Nina Farstad, Enric Domingo, David N. Church, Arild Nesbakken, Neil A. Shepherd, Ian Tomlinson, Rachel Kerr, Marco Novelli, David J. Kerr, and Håvard E. Danielsen. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet (London, England)*, 395(10221):350–360, February 2020.

[34] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, and Sunghoon Kwon. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, August 2022.

[35] Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction using Patch-based Graph Convolutional Networks, July 2021. arXiv:2107.13048 [cs, eess, q-bio].

[36] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, and Jianhua Yao. Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4836–4845, Seattle, WA, USA, June 2020. IEEE.

[37] Richard A. Scolyer, Robert V. Rawson, Jeffrey E. Gershenwald, Peter M. Ferguson, and Victor G. Prieto. Melanoma pathology reporting and staging. *Modern Pathology*, 33:15–24, January 2020.

[38] Charles M. Balch, Jeffrey E. Gershenwald, Seng-jaw Soong, John F. Thompson, Michael B. Atkins, David R. Byrd, Antonio C. Buzaid, Alistair J. Cochran, Daniel G.

Coit, Shouluan Ding, Alexander M. Eggermont, Keith T. Flaherty, Phyllis A. Gimotty, John M. Kirkwood, Kelly M. McMasters, Martin C. Mihm, Donald L. Morton, Merrick I. Ross, Arthur J. Sober, and Vernon K. Sondak. Final Version of 2009 AJCC Melanoma Staging and Classification. *Journal of Clinical Oncology*, 27(36):6199–6206, December 2009.

[39] Jeffrey E. Gershenwald and Richard A. Scolyer. Melanoma Staging: American Joint Committee on Cancer (AJCC) 8th Edition and Beyond. *Annals of Surgical Oncology*, 25(8):2105–2110, August 2018.

[40] A. Mitra, C. Conway, C. Walker, M. Cook, B. Powell, S. Lobo, M. Chan, M. Kissin, G. Layer, J. Smallwood, C. Ottensmeier, P. Stanley, H. Peach, H. Chong, F. Elliott, M. M. Iles, J. Nsengimana, J. H. Barrett, D. T. Bishop, and J. A. Newton-Bishop. Melanoma sentinel node biopsy and prediction models for relapse and overall survival. *British Journal of Cancer*, 103(8):1229–1236, October 2010.

[41] Laura Taylor, Kyle Hood, Lisa Reisch, Joann Elmore, Michael Piepkorn, Raymond Barnhill, Stevan Knezevich, Andrea Radick, and David Elder. Influence of variability in assessment of Breslow thickness, mitotic rate and ulceration among US pathologists interpreting invasive melanoma, for the purpose of AJCC staging. *Journal of cutaneous pathology*, 45(8):588–596, August 2018.

[42] Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Newton-Bishop Julia, and Ali Gooya. Weakly-supervised learning for image-based classification of primary melanomas into genomic immune subgroups. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 423–440. PMLR, December 2022. ISSN: 2640-3498.

[43] Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Julia Newton-Bishop, Ali Gooya, and Derek Magee. Immune subtyping of melanoma whole slide images using multiple instance learning. *Medical Image Analysis*, 93:103097, April 2024.

[44] Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P. Cook, Emily L. Clarke, Darren Treanor, D. Timothy Bishop, Julia Newton-Bishop, and Derek Magee. Multi-level Graph Representations of Melanoma Whole Slide Images for Identifying Immune Subgroups. In Seyed-Ahmad Ahmadi and Sérgio Pereira, editors, *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology*, pages 85–96, Cham, 2024. Springer Nature Switzerland.

[45] Serigne N. Lo, Richard A. Scolyer, and John F. Thompson. Long-Term Survival of Patients with Thin (T1) Cutaneous Melanomas: A Breslow Thickness Cut Point of 0.8 mm

Separates Higher-Risk and Lower-Risk Tumors. *Annals of Surgical Oncology*, 25(4):894–902, April 2018.

[46] Noemi Andor, Trevor A. Graham, Marnix Jansen, Li C. Xia, C. Athena Aktipis, Claudia Petritsch, Hanlee P. Ji, and Carlo C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, January 2016. Number: 1 Publisher: Nature Publishing Group.

[47] Christopher E. M. Griffiths, Jonathan Barker, Tanya O. Bleiker, Robert Chalmers, and Daniel Creamer. *Rook's Textbook of Dermatology*. John Wiley & Sons, February 2016. Google-Books-ID: EyypCwAAQBAJ.

[48] Lyn McDivitt Duncan. The classification of cutaneous melanoma. *Hematology/Oncology Clinics of North America*, 23(3):501–513, ix, June 2009.

[49] Raul Ossio, Rodrigo Roldán-Marín, Héctor Martínez-Said, David J. Adams, and Carla Daniela Robles-Espinoza. Melanoma: a global perspective. *Nature Reviews. Cancer*, 17(7):393–394, July 2017.

[50] Mary-Ann El Sharouni, Paul Johannes van Diest, Arjen Joost Witkamp, Vigfús Sigurdsson, and Carla Henrica van Gils. Subtyping Cutaneous Melanoma Matters. *JNCI Cancer Spectrum*, 4(6):pkaa097, October 2020.

[51] Michael Xiong, Ahmad Charifa, and Chih Shan J. Chen. Lentigo Maligna Melanoma. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023.

[52] Gretchen M. Alicea and Vito W. Rebecca. Un-Fair Skin: racial disparities in acral melanoma research. *Nature reviews. Cancer*, 22(3):127–128, March 2022.

[53] Wesley O. Greaves, Shalini Verma, Keyur P. Patel, Michael A. Davies, Bedia A. Barkoh, John M. Galbincea, Hui Yao, Alexander J. Lazar, Kenneth D. Aldape, L. Jeffrey Medeiros, and Rajyalakshmi Luthra. Frequency and spectrum of BRAF mutations in a retrospective, single-institution study of 1112 cases of melanoma. *The Journal of molecular diagnostics: JMD*, 15(2):220–226, March 2013.

[54] Michael Lattanzi, Yesung Lee, Danny Simpson, Una Moran, Farbod Darvishian, Randie H Kim, Eva Hernando, David Polsky, Doug Hanniford, Richard Shapiro, Russell Berman, Anna C Pavlick, Melissa A Wilson, Tomas Kirchhoff, Jeffrey S Weber, Judy Zhong, and Iman Osman. Primary Melanoma Histologic Subtype: Impact on Survival and Response to Therapy. *JNCI Journal of the National Cancer Institute*, 111(2):180–188, June 2018.

[55] Genomic Classification of Cutaneous Melanoma. *Cell*, 161(7):1681–1696, June 2015.

[56] Emily L. Clarke, Ryckie G. Wade, Derek Magee, Julia Newton-Bishop, and Darren Treanor. Image analysis of cutaneous melanoma histology: a systematic review and meta-analysis. *Scientific Reports*, 13(1):4774, March 2023. Number: 1 Publisher: Nature Publishing Group.

[57] Ji Young Seo, Soon Young Kim, Jeong Hee Kang, Young Ok Kim, and Bang Hur. Balloon Cell Malignant Melanoma: A case report. *Journal of Pathology and Translational Medicine*, 33(7):537–539. Publisher: The Korean Society of Pathologists and the Korean Society for Cytopathology.

[58] Stephanie E. Weissinger, Philipp Keil, David N. Silvers, Beate M. Klaus, Peter Möller, Basil A. Horst, and Jochen K. Lennerz. A diagnostic algorithm to distinguish desmoplastic from spindle cell melanoma. *Modern Pathology*, 27(4):524–534, April 2014. Number: 4 Publisher: Nature Publishing Group.

[59] Joann G. Elmore, Raymond L. Barnhill, David E. Elder, Gary M. Longton, Margaret S. Pepe, Lisa M. Reisch, Patricia A. Carney, Linda J. Titus, Heidi D. Nelson, Tracy Onega, Anna N. A. Tosteson, Martin A. Weinstock, Stevan R. Knezevich, and Michael W. Piepkorn. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ*, 357:j2813, June 2017. Publisher: British Medical Journal Publishing Group Section: Research.

[60] Andrea Ladányi. Prognostic and predictive significance of immune cells infiltrating cutaneous melanoma. *Pigment Cell & Melanoma Research*, 28(5):490–500, September 2015.

[61] Tobias Schatton, Richard A. Scolyer, John F. Thompson, and Martin C. Mihm. Tumor-Infiltrating Lymphocytes and Their Significance in Melanoma Prognosis. In Magdalena Thurin and Francesco M. Marincola, editors, *Molecular Diagnostics for Melanoma: Methods and Protocols*, Methods in Molecular Biology, pages 287–324. Humana Press, Totowa, NJ, 2014.

[62] Qiaofen Fu, Nan Chen, Chunlei Ge, Ruilei Li, Zhen Li, Baozhen Zeng, Chunyan Li, Ying Wang, Yuanbo Xue, Xin Song, Heng Li, and Gaofeng Li. Prognostic value of tumor-infiltrating lymphocytes in melanoma: a systematic review and meta-analysis. *Oncoimmunology*, 8(7):1593806, April 2019.

[63] Hashem O. Alsaab, Samaresh Sau, Rami Alzhrani, Katyayani Tatiparti, Ketki Bhise, Sushil K. Kashaw, and Arun K. Iyer. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Frontiers in Pharmacology*, 8:561, August 2017.

[64] Roy S. Herbst, Jean-Charles Soria, Marcin Kowanetz, Gregg D. Fine, Omid Hamid, Michael S. Gordon, Jeffery A. Sosman, David F. McDermott, John D. Powderly, Scott N.

Gettinger, Holbrook E. K. Kohrt, Leora Horn, Donald P. Lawrence, Sandra Rost, Maya Leabman, Yuanyuan Xiao, Ahmad Mokatrin, Hartmut Koeppen, Priti S. Hegde, Ira Mellman, Daniel S. Chen, and F. Stephen Hodi. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*, 515(7528):563–567, November 2014. Number: 7528 Publisher: Nature Publishing Group.

[65] Nicholas L Syn, Michele W L Teng, Tony S K Mok, and Ross A Soo. De-novo and acquired resistance to immune checkpoint targeting. *The Lancet Oncology*, 18(12):e731–e741, December 2017.

[66] Macmillan Cancer information and support. Staging of melanoma - TNM staging and Breslow thickness, 2024. Last ccessed: 2024-01-16.

[67] Faruk Tas and Kayhan Erturk. Histological lymphovascular invasion is associated with nodal involvement, recurrence, and survival in patients with cutaneous malignant melanoma. *International Journal of Dermatology*, 56(2):166–170, February 2017.

[68] W. Phillip Law, Nolette Pereira, and Kasturi Vaska. Perineural spread of recurrent cutaneous melanoma along cervical nerves into the spinal cord. *BJR Case Reports*, 3(2), January 2017.

[69] C. M. Balch, A. C. Buzaid, S. J. Soong, M. B. Atkins, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Houghton, J. M. Kirkwood, K. M. McMasters, M. F. Mihm, D. L. Morton, D. S. Reintgen, M. I. Ross, A. Sober, J. A. Thompson, and J. F. Thompson. Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 19(16):3635–3648, August 2001.

[70] Dennis L. Rousseau, Merrick I. Ross, Marcella M. Johnson, Victor G. Prieto, Jeffrey E. Lee, Paul F. Mansfield, and Jeffrey E. Gershenwald. Revised American Joint Committee on Cancer staging criteria accurately predict sentinel lymph node positivity in clinically node-negative melanoma patients. *Annals of Surgical Oncology*, 10(5):569–574, June 2003.

[71] P. J. Heenan, L. R. Matz, J. B. Blackwell, G. R. Kelsall, A. Singh, R. E. ten Seldam, and C. D. Holman. Inter-observer variation between pathologists in the classification of cutaneous malignant melanoma in western Australia. *Histopathology*, 8(5):717–729, September 1984.

[72] Rajmohan Murali, Michael T. Hughes, Patrick Fitzgerald, John F. Thompson, and Richard A. Scolyer. Interobserver variation in the histopathologic reporting of key prognostic parameters, particularly clark level, affects pathologic staging of primary cutaneous melanoma. *Annals of Surgery*, 249(4):641–647, April 2009.

[73] P. S. Colloby, K. P. West, and A. Fletcher. Observer variation in the measurement of Breslow depth and Clark's level in thin cutaneous malignant melanoma. *The Journal of Pathology*, 163(3):245–250, March 1991.

[74] M. Prade, H. Sancho-Garnier, J. P. Cesarini, and A. Cochran. Difficulties encountered in the application of Clark classification and the Breslow thickness measurement in cutaneous malignant melanoma. *International Journal of Cancer*, 26(2):159–163, August 1980.

[75] Samit Patrawala, Alexander Maley, Caitlin Greskovich, Lauren Stuart, Douglas Parker, Robert Swerlick, and Benjamin Stoff. Discordance of histopathologic parameters in cutaneous melanoma: Clinical implications. *Journal of the American Academy of Dermatology*, 74(1):75–80, January 2016.

[76] A. Spatz, M. G. Cook, D. E. Elder, M. Piepkorn, D. J. Ruiter, and R. L. Barnhill. Interobserver reproducibility of ulceration assessment in primary cutaneous melanomas. *European Journal of Cancer (Oxford, England: 1990)*, 39(13):1861–1865, September 2003.

[77] C. M. Balch, S. J. Soong, J. E. Gershenwald, J. F. Thompson, D. S. Reintgen, N. Cascinelli, M. Urist, K. M. McMasters, M. I. Ross, J. M. Kirkwood, M. B. Atkins, J. A. Thompson, D. G. Coit, D. Byrd, R. Desmond, Y. Zhang, P. Y. Liu, G. H. Lyman, and A. Morabito. Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 19(16):3622–3634, August 2001.

[78] V. J. McGovern, H. M. Shaw, G. W. Milton, and W. H. McCarthy. Ulceration and prognosis in cutaneous malignant melanoma. *Histopathology*, 6(4):399–407, July 1982.

[79] Rosalyn Jewell, Faye Elliott, Jonathan Laye, Jérémie Nsengimana, John Davies, Christy Walker, Caroline Conway, Angana Mitra, Mark Harland, Martin G. Cook, Andy Boon, Sarah Storr, Sabreena Safuan, Stewart G. Martin, Karin Jirström, Håkan Olsson, Christian Ingvar, Martin Lauss, Tim Bishop, Göran Jönsson, and Julia Newton-Bishop. The clinicopathological and gene expression patterns associated with ulceration of primary melanoma. *Pigment Cell & Melanoma Research*, 28(1):94–104, January 2015.

[80] Alexander M. M. Eggermont, Stefan Suciu, Mario Santinami, Alessandro Testori, Wim H. J. Kruit, Jeremy Marsden, Cornelis J. A. Punt, François Salès, Martin Gore, Rona Mackie, Zvonko Kusic, Reinhard Dummer, Axel Hauschild, Elena Musat, Alain Spatz, Ulrich Keilholz, and EORTC Melanoma Group. Adjuvant therapy with pegylated interferon alfa-2b versus observation alone in resected stage III melanoma: final results of EORTC 18991, a randomised phase III trial. *Lancet (London, England)*, 372(9633):117–126, July 2008.

[81] Alexander M. M. Eggermont, Stefan Suciu, Alessandro Testori, Mario Santinami, Wim H. J. Kruit, Jeremy Marsden, Cornelis J. A. Punt, François Salès, Reinhard Dummer, Caroline Robert, Dirk Schadendorf, Poulam M. Patel, Gaetan de Schaetzen, Alan Spatz, and Ulrich Keilholz. Long-term results of the randomized phase III trial EORTC 18991 of adjuvant therapy with pegylated interferon alfa-2b versus observation in resected stage III melanoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 30(31):3810–3818, November 2012.

[82] Manuela F. Azzola, Helen M. Shaw, John F. Thompson, Seng-jaw Soong, Richard A. Scolyer, Geoffrey F. Watson, Marjorie H. Colman, and Yuting Zhang. Tumor mitotic rate is a more powerful prognostic indicator than ulceration in patients with primary cutaneous melanoma. *Cancer*, 97(6):1488–1498, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.11196.

[83] Mohammed Kashani-Sabet, James R. Miller III, Serigne Lo, Mehdi Nosrati, Jonathan R. Stretch, Kerwin F. Shannon, Andrew J. Spillane, Robyn P. M. Saw, James E. Cleaver, Kevin B. Kim, Stanley P. Leong, John F. Thompson, and Richard A. Scolyer. Reappraisal of the prognostic significance of mitotic rate supports its reincorporation into the melanoma staging system. *Cancer*, 126(21):4717–4725, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.33088.

[84] Jérémie Nsengimana, Jon Laye, Anastasia Filia, Sally O'Shea, Sathya Muralidhar, Joanna Pożniak, Alastair Droop, May Chan, Christy Walker, Louise Parkinson, Joanne Gascoyne, Tracey Mell, Minttu Polso, Rosalyn Jewell, Juliette Randerson-Moor, Graham P. Cook, D. Timothy Bishop, and Julia Newton-Bishop. -Catenin–mediated immune evasion pathway frequently operates in primary cutaneous melanomas, May 2018. Publisher: American Society for Clinical Investigation.

[85] Geoffrey T. Gibney, Louis M. Weiner, and Michael B. Atkins. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *The Lancet. Oncology*, 17(12):e542–e551, December 2016.

[86] Rui-Ru Ji, Scott D. Chasalow, Lisu Wang, Omid Hamid, Henrik Schmidt, John Cogswell, Suresh Alaparthy, David Berman, Maria Jure-Kunkel, Nathan O. Siemers, Jeffrey R. Jackson, and Vafa Shahabi. An immune-active tumor microenvironment favors clinical response to ipilimumab. *Cancer immunology, immunotherapy: CII*, 61(7):1019–1031, July 2012.

[87] James Larkin, Vanna Chiarion-Sileni, Rene Gonzalez, Jean Jacques Grob, C. Lance Cowey, Christopher D. Lao, Dirk Schadendorf, Reinhard Dummer, Michael Smylie, Piotr Rutkowski, Pier Francesco Ferrucci, Andrew Hill, John Wagstaff, Matteo S. Carlino, John B. Haanen, Michele Maio, Ivan Marquez-Rodas, Grant A. McArthur, Paolo A.

Ascierto, Georgina V. Long, Margaret K. Callahan, Michael A. Postow, Kenneth Grossmann, Mario Sznol, Brigitte Dreno, Lars Bastholt, Arvin Yang, Linda M. Rollin, Christine Horak, F. Stephen Hodi, and Jedd D. Wolchok. Combined Nivolumab and Ipilimumab or Monotherapy in Previously Untreated Melanoma. *The New England journal of medicine*, 373(1):23–34, July 2015.

[88] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, Anne Berger, Patrick Bruneval, Wolf Herman Fridman, Christoph Becker, Franck Pagès, Michael R. Speicher, Zlatko Trajanoski, and Jérôme Galon. Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity*, 39(4):782–795, October 2013. Publisher: Elsevier.

[89] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.

[90] B. Widrow and M. E. Hoff. Adaptive Switching Circuits. *1960 IRE WESCON Convention Record*, 1960.

[91] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington DC., 1962.

[92] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. Number: 6088 Publisher: Nature Publishing Group.

[93] Chris Bishop. *Pattern Recognition and Machine Learning*. New York :Springer, 2006.

[94] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

[95] Paul J. Werbos. *The Roots of Backpropagation : From Ordered Derivatives to Neural Networks and Political Forecasting*. New York: John Wiley & Sons, 1994.

[96] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160:106–154, 1962.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[98] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[99] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An All-In-One Convolutional Neural Network for Face Analysis. *arXiv:1611.00851 [cs]*, November 2016. arXiv: 1611.00851.

[100] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. DeepTest: automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, pages 303–314, Gothenburg Sweden, May 2018. ACM.

[101] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, January 2021.

[102] Emily L Clarke and Darren Treanor. Colour in digital pathology: a review. *Histopathology*, 70(2):153–163, 2017.

[103] Miao Cui and David Y. Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, April 2021.

[104] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, February 2021. Number: 4 Publisher: Nature Publishing Group.

[105] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Lecture Notes in Computer Science, pages 411–418, Berlin, Heidelberg, 2013. Springer.

[106] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, January 1997.

[107] Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, pages 570–576. MIT Press, 1998.

[108] Narmin Ghaffari Laleh, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu, Christian Trautwein, Rupert Langer, Bastian Dislich, Roman D. Buelow, Heike Irmgard Grabsch, Hermann Brenner, Jenny Chang-Claude, Elizabeth Alwers, Titus J. Brinker, Firas Khader, Daniel Truhn, Nadine T. Gaisa, Peter Boor, Michael Hoffmeister, Volkmar Schulz, and Jakob Nikolas Kather. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical Image Analysis*, 79:102474, July 2022.

[109] Ming Y. Lu, Melissa Zhao, Maha Shady, Jana Lipkova, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Deep Learning-based Computational Pathology Predicts Origins for Cancers of Unknown Primary. *Nature*, 594(7861):106–110, June 2021. arXiv:2006.13932 [cs, q-bio].

[110] Omar S. M. El Nahhas, Chiara M. L. Loeffler, Zunamys I. Carrero, Marko van Treeck, Fiona R. Kolbinger, Katherine J. Hewitt, Hannah S. Muti, Mara Graziani, Qinghe Zeng, Julien Calderaro, Nadina Ortiz-Brüchle, Tanwei Yuan, Michael Hoffmeister, Hermann Brenner, Alexander Brobeil, Jorge S. Reis-Filho, and Jakob Nikolas Kather. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides, April 2023. arXiv:2304.05153 [cs].

[111] Ruqayya Awan, Navid Alemi Koohbanani, Muhammad Shaban, Anna Lisowska, and Nasir Rajpoot. Context-Aware Learning using Transferable Features for Classification of Breast Cancer Histology Images, March 2018. arXiv:1803.00386 [cs].

[112] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G. Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, October 2019. Number: 10 Publisher: Nature Publishing Group.

[113] Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, July 2019. Number: 7 Publisher: Nature Publishing Group.

[114] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets, April 2023. arXiv:2212.04690 [cs].

[115] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, March 2020. arXiv:1911.05722 [cs].

[116] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

[117] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[118] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning, September 2020. arXiv:2006.07733 [cs, stat].

[119] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. pages 15750–15758, 2021.

[120] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12310–12320. PMLR, July 2021. ISSN: 2640-3498.

[121] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11218, pages 139–156. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[122] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[123] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, Montreal, QC, Canada, October 2021. IEEE.

[124] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, Montreal, QC, Canada, October 2021. IEEE.

[125] Ozan Ciga, Tony Xu, and Anne L. Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, September 2022. arXiv: 2011.13971.

[126] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. arXiv:2002.05709 [cs, stat].

[127] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. ISSN: 1063-6919.

[128] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. *arXiv:1802.04712 [cs, stat]*, June 2018. arXiv: 1802.04712.

[129] Andrew H. Song, Guillaume Jaume, Drew F. K. Williamson, Ming Y. Lu, Anurag Vaidya, Tiffany R. Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, December 2023. Number: 12 Publisher: Nature Publishing Group.

[130] Eftychia Chatziioannou, Jana Roßner, Thazin New Aung, David L. Rimm, Heike Niessner, Ulrike Keim, Lina Maria Serna-Higuita, Irina Bonzheim, Luis Kuhn Cuellar, Dana Westphal, Julian Steininger, Friedegund Meier, Oltin Tiberiu Pop, Stephan Forchhammer, Lukas Flatz, Thomas Eigentler, Claus Garbe, Martin Röcken, Teresa Amaral, and Tobias Sinnberg. Deep learning-based scoring of tumour-infiltrating lymphocytes is prognostic in primary melanoma and predictive to PD-1 checkpoint inhibition in melanoma metastases. *eBioMedicine*, 93, July 2023. Publisher: Elsevier.

[131] David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. A Survey on Graph-Based Deep Learning for Computational Histopathology, September 2021. arXiv:2107.00272 [cs, q-bio].

[132] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing Cellular Topology in Multi-Gigapixel Pathology Images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1049–1058, Seattle, WA, USA, June 2020. IEEE.

[133] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):933–941, June 2022. Section: AAAI Technical Track on Computer Vision I.

[134] Pushpak Pati, Guillaume Jaume, Lauren Alisha Fernandes, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Daniel Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Orcun Goksel, Jean-Philippe Thiran, Maria Frucci, and Maria Gabrani. HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification. In Carole H. Sudre, Hamid Fehri, Tal Arbel, Christian F. Baumgartner, Adrian Dalca, Ryutaro Tanno, Koen Van Leemput, William M. Wells, Aristeidis Sotiras, Bartlomiej Papiez, Enzo Ferrante, and Sarah Parisot, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, volume 12443, pages 208–219. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[135] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Jean-Philippe Thiran, Maria Frucci, Orcun Goksel, and Maria Gabrani. Hierarchical graph representations in digital pathology. *Medical Image Analysis*, 75:102264, January 2022.

[136] Pushpak Pati, Guillaume Jaume, Zeineb Ayadi, Kevin Thandiackal, Behzad Bozorgtabar, Maria Gabrani, and Orcun Goksel. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Medical Image Analysis*, 89:102915, October 2023.

[137] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. pages 0–0, 2019.

[138] Joe Sims, Heike I. Grabsch, and Derek Magee. Using Hierarchically Connected Nodes and Multiple GNN Message Passing Steps to Increase the Contextual Information in Cell-Graph Classification. In Luigi Manfredi, Seyed-Ahmad Ahmadi, Michael Bronstein, Anees Kazi, Davide Lomanto, Alwyn Mathew, Ludovic Magerand, Kamilia Mullakaeva, Bartlomiej Papiez, Russell H. Taylor, and Emanuele Trucco, editors, *Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis*, Lecture Notes in Computer Science, pages 99–107, Cham, 2022. Springer Nature Switzerland.

[139] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes in Computer Science, pages 755–764, Cham, 2023. Springer Nature Switzerland.

[140] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs, September 2018. arXiv:1706.02216 [cs, stat].

[141] Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks?, January 2022. arXiv:2105.14491 [cs] version: 3.

[142] Jure Leskovec. Representation Learning on Networks. Last ccessed: 2024-03-26.

[143] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression, 2019.

[144] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x.

[145] Charles M. Balch, Seng-jaw Soong, Jeffrey E. Gershenwald, John F. Thompson, Daniel G. Coit, Michael B. Atkins, Shouluan Ding, Alistair J. Cochran, Alexander M.M. Eggermont, Keith T. Flaherty, Phyllis A. Gimotty, Timothy Johnson, John M. Kirkwood, Stanley P. Leong, Kelly M. McMasters, Martin C. Mihm, Donald L. Morton, Merrick I. Ross, and Vernon K. Sondak. Age as a Prognostic Factor in Patients with Localized Melanoma and Regional Metastases. *Annals of surgical oncology*, 20(12):3961–3968, November 2013.

[146] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1):24, December 2018. arXiv:1606.00931 [cs, stat].

[147] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, January 1995.

[148] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, December 2016.

[149] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. WSISA: Making Survival Prediction from Whole Slide Histopathological Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6863, July 2017. ISSN: 1063-6919.

[150] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, October 2020.

[151] Muyi Sun, Wei Zhou, Xingqun Qi, Guanhong Zhang, Leonard Girnita, Stefan Seregard, Hans E. Grossniklaus, Zeyi Yao, Xiaoguang Zhou, and Gustav Stålhammar. Prediction of BAP1 Expression in Uveal Melanoma Using Densely-Connected Deep Classification Networks. *Cancers*, 11(10):1579, October 2019. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[152] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. pages 4700–4708, 2017.

[153] Randie H. Kim, Sofia Nomikou, Nicolas Coudray, George Jour, Zarmeena Dawood, Runyu Hong, Eduardo Esteva, Theodore Sakellaropoulos, Douglas Donnelly, Una Moran, Aristides Hatzimemos, Jeffrey S. Weber, Narges Razavian, Iannis Aifantis, David Fenyo, Matija Snuderl, Richard Shapiro, Russell S. Berman, Iman Osman, and Aristotelis Tsirigos. Deep Learning and Pathomics Analyses Reveal Cell Nuclei as Important Features for Mutation Prediction of BRAF-Mutated Melanomas. *Journal of Investigative Dermatology*, 142(6):1650–1658.e6, June 2022.

[154] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA, June 2015. IEEE.

[155] Hongrun Zhang, Helen Kalirai, Amelia Acha-Sagredo, Xiaoyun Yang, Yalin Zheng, and Sarah E. Coupland. Piloting a Deep Learning Model for Predicting Nuclear BAP1 Immunohistochemical Expression of Uveal Melanoma from Hematoxylin-and-Eosin Sections. *Translational Vision Science & Technology*, 9(2):50, September 2020.

[156] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].

[157] Chiara Maria Lavinia Loeffler, Nadine T. Gaisa, Hannah Sophie Muti, Marko van Treeck, Amelie Echle, Narmin Ghaffari Laleh, Christian Trautwein, Lara R. Heij, Heike I. Grabsch, Nadina Ortiz Bruechle, and Jakob Nikolas Kather. Predicting Mutational Status of Driver and Suppressor Genes Directly from Histopathology With Deep Learning: A Systematic Study Across 23 Solid Tumor Types. *Frontiers in Genetics*, 12:806386, 2021.

[158] Daisuke Komura, Akihiro Kawabe, Keisuke Fukuta, Kyohei Sano, Toshikazu Umezaki, Hirotomo Koda, Ryohei Suzuki, Ken Tominaga, Mieko Ochi, Hiroki Konishi, Fumiya Masakado, Noriyuki Saito, Yasuyoshi Sato, Takumi Onoyama, Shu Nishida, Genta Furuya, Hiroto Katoh, Hiroharu Yamashita, Kazuhiro Kakimi, Yasuyuki Seto, Tetsuo Ushiku, Masashi Fukayama, and Shumpei Ishikawa. Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports*, 38(9), March 2022. Publisher: Elsevier.

[159] Yu Fu, Alexander W. Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R. Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, August 2020. Number: 8 Publisher: Nature Publishing Group.

[160] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):16878, December 2017. Number: 1 Publisher: Nature Publishing Group.

[161] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, Salt Lake City, UT, June 2018. IEEE.

[162] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition, April 2018. arXiv:1707.07012 [cs, stat] version: 4.

[163] Richard A. Diaz, Raisa Sidhu, Janet Robertson, and Jane Adam. NICE guidance on ipilimumab for previously treated advanced melanoma. *The Lancet Oncology*, 14(1):23–24, January 2013. Publisher: Elsevier.

[164] Caroline Robert, Antoni Ribas, Jacob Schachter, Ana Arance, Jean-Jacques Grob, Laurent Mortier, Adil Daud, Matteo S. Carlino, Catriona M. McNeil, Michal Lotem, James M. G. Larkin, Paul Lorigan, Bart Neyns, Christian U. Blank, Teresa M. Petrella, Omid Hamid, Shu-Chih Su, Clemens Krepler, Nageatte Ibrahim, and Georgina V. Long. Pembrolizumab versus ipilimumab in advanced melanoma (KEYNOTE-006): post-hoc 5-year results from an open-label, multicentre, randomised, controlled, phase 3 study. *The Lancet Oncology*, 20(9):1239–1251, September 2019. Publisher: Elsevier.

[165] Caroline Robert, Jacob Schachter, Georgina V. Long, Ana Arance, Jean Jacques Grob, Laurent Mortier, Adil Daud, Matteo S. Carlino, Catriona McNeil, Michal Lotem, James Larkin, Paul Lorigan, Bart Neyns, Christian U. Blank, Omid Hamid, Christine Mateus, Ronnie Shapira-Frommer, Michele Kosh, Honghong Zhou, Nageatte Ibrahim, Scot Ebbinghaus, and Antoni Ribas. Pembrolizumab versus Ipilimumab in Advanced Melanoma. *New England Journal of Medicine*, 372(26):2521–2532, June 2015. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1503093.

[166] Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A. J. Sommer, Peter Bankhead, Loes F. S. Kooreman, Jefree J. Schulte, Nicole A. Cipriani, Roman D. Buelow, Peter Boor, Nadi-na Ortiz-Brüchle, Andrew M. Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A. van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T. Pearson, and Tom Luedde. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, August 2020.

[167] James A. Diao, Jason K. Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N. Mitchell, Benjamin Glass, Sara Hoffman, Sudha K. Rao, Chirag Maheshwari, Abhik Lahiri, Aaditya Prakash, Ryan McLoughlin, Jennifer K. Kerner, Murray B. Resnick, Michael C. Montalto, Aditya Khosla, Ilan N. Wapinski, Andrew H. Beck, Hunter L. Elliott, and Amaro Taylor-Weiner. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature Communications*, 12(1):1613, March 2021.

[168] Vésteinn Thorsson, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F. Gao, Christopher L. Plaisier, James A.

Eddy, Elad Ziv, Aedin C. Culhane, Evan O. Paull, I. K. Ashok Sivakumar, Andrew J. Gentles, Raunaq Malhotra, Farshad Farshidfar, Antonio Colaprico, Joel S. Parker, Lisle E. Mose, Nam Sy Vo, Jianfang Liu, Yuexin Liu, Janet Rader, Varsha Dhankani, Sheila M. Reynolds, Reanne Bowlby, Andrea Califano, Andrew D. Cherniack, Dimitris Anastassiou, Davide Bedognetti, Younes Mokrab, Aaron M. Newman, Arvind Rao, Ken Chen, Alexander Krasnitz, Hai Hu, Tathiane M. Malta, Houtan Noushmehr, Chandra Sekhar Pedamallu, Susan Bullman, Akinyemi I. Ojesina, Andrew Lamb, Wanding Zhou, Hui Shen, Toni K. Choueiri, John N. Weinstein, Justin Guinney, Joel Saltz, Robert A. Holt, Charles S. Rabkin, Cancer Genome Atlas Research Network, Alexander J. Lazar, Jonathan S. Serody, Elizabeth G. Demicco, Mary L. Disis, Benjamin G. Vincent, and Ilya Shmulevich. The Immune Landscape of Cancer. *Immunity*, 48(4):812–830.e14, April 2018.

[169] Daniel Sauter, Georg Lodde, Felix Nensa, Dirk Schadendorf, Elisabeth Livingstone, and Markus Kukuk. Deep learning in computational dermatopathology of melanoma: A technical systematic literature review. *Computers in Biology and Medicine*, 163:107083, September 2023.

[170] Bettina Sobottka, Marta Nowak, Anja Laura Frei, Martina Haberecker, Samuel Merki, Mitchell P. Levesque, Reinhard Dummer, Holger Moch, Viktor Hendrik Koelzer, Rudolf Aebersold, Melike Ak, Faisal S. Al-Quaddoomi, Jonas Albinus, Ilaria Alborelli, Sonali Andani, Per-Olof Attinger, Marina Bacac, Daniel Baumhoer, Beatrice Beck-Schimmer, Niko Beerenwinkel, Christian Beisel, Lara Bernasconi, Anne Bertolini, Bernd Bodenmiller, Ximena Bonilla, Ruben Casanova, Stéphane Chevrier, Natalia Chicherova, Maya D'Costa, Esther Danenberg, Natalie Davidson, Monica-Andreea Drăganmoch, Stefanie Engler, Martin Erkens, Katja Eschbach, Cinzia Esposito, André Fedier, Pedro Ferreira, Joanna Ficek, Bruno Frey, Sandra Goetze, Linda Grob, Gabriele Gut, Detlef Günther, Martina Haberecker, Pirmin Haeuptle, Viola Heinzelmann-Schwarz, Sylvia Herter, Rene Holtackers, Tamara Huesser, Anja Irmisch, Francis Jacob, Andrea Jacobs, Tim M. Jaeger, Katharina Jahn, Alva R. James, Philip M. Jermann, André Kahles, Abdullah Kahraman, Werner Kuebler, Jack Kuipers, Christian P. Kunze, Christian Kurzeder, Kjong-Van Lehmann, Sebastian Lugert, Gerd Maass, Markus G. Manz, Philipp Markolin, Julien Mena, Ulrike Menzel, Julian M. Metzler, Nicola Miglino, Emanuela S. Milani, Simone Muenst, Riccardo Murri, Charlotte K. Y. Ng, Stefan Nicolet, Patrick G. A. Pedrioli, Lucas Pelkmans, Salvatore Piscuoglio, Michael Prummer, Mathilde Ritter, Christian Rommel, María L. Rosano-González, Gunnar Rätsch, Natascha Santacroce, Jacobo Sarabia del Castillo, Ramona Schlenker, Petra C. Schwalie, Severin Schwan, Tobias Schär, Gabriela Senti, Franziska Singer, Sujana Sivapatham, Berend Snijder, Vipin T. Sreedharan, Stefan Stark, Daniel J. Stekhoven, Alexandre P. A. Theocharides, Tinu M. Thomas, Markus Tolnay, Vinko Tosevski, Nora C. Toussaint, Mustafa A. Tuncel, Marina Tusup,

Audrey Van Drogen, Marcus Vetter, Tatjana Vlajnic, Sandra Weber, Walter P. Weber, Rebekka Wegmann, Michael Weller, Fabian Wendt, Norbert Wey, Andreas Wicki, Mattheus HE Wildschut, Bernd Wollscheid, Shuqing Yu, Johanna Ziegler, Marc Zimmermann, Martin Zoche, and Gregor Zuend. Establishing standardized immune phenotyping of metastatic melanoma by digital pathology. *Laboratory Investigation*, 101(12):1561–1570, December 2021. Publisher: Elsevier.

[171] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Tianhao Zhao, Rebecca Batiste, John Van Arnam, Cancer Genome Atlas Research Network, Ilya Shmulevich, Arvind U. K. Rao, Alexander J. Lazar, Ashish Sharma, and Vésteinn Thorsson. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Reports*, 23(1):181–193.e7, April 2018.

[172] Margaret Chou, Irineu Illa-Bochaca, Ben Minxi, Farbod Darvishian, Paul Johannet, Una Moran, Richard L. Shapiro, Russell S. Berman, Iman Osman, George Jour, and Hua Zhong. Optimization of an automated tumor-infiltrating lymphocyte algorithm for improved prognostication in primary melanoma. *Modern Pathology*, pages 1–10, October 2020. Publisher: Nature Publishing Group.

[173] Konstantinos Zormpas-Petridis, Henrik Failmezger, Shan E Ahmed Raza, Ioannis Roxanis, Yann Jamin, and Yinyin Yuan. Superpixel-Based Conditional Random Fields (SuperCRF): Incorporating Global and Local Context for Enhanced Deep Learning in Melanoma Histopathology. *Frontiers in Oncology*, 9, 2019.

[174] Nathalie Harder, Ralf Schönmeyer, Katharina Nekolla, Armin Meier, Nicolas Brieu, Carolina Vanegas, Gabriele Madonna, Mariaelena Capone, Gerardo Botti, Paolo A. Ascierto, and Günter Schmidt. Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Scientific Reports*, 9(1):7449, May 2019. Number: 1 Publisher: Nature Publishing Group.

[175] Henrik Failmezger, Sathya Muralidhar, Antonio Rullan, Carlos E. de Andrea, Erik Sahai, and Yinyin Yuan. Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology. *Cancer Research*, 80(5):1199–1209, March 2020.

[176] Prathamesh M. Kulkarni, Eric J. Robinson, Jaya Sarin Pradhan, Robyn D. Gartrell-Corrado, Bethany R. Rohr, Megan H. Trager, Larisa J. Geskin, Harriet M. Kluger, Pok Fai Wong, Balazs Acs, Emanuelle M. Rizk, Chen Yang, Manas Mondal, Michael R. Moore, Iman Osman, Robert Phelps, Basil A. Horst, Zhe S. Chen, Tammie Ferringer, David L. Rimm, Jing Wang, and Yvonne M. Saenger. Deep Learning Based on Standard H&E Images of Primary Melanoma Tumors Identifies Patients at Risk for Visceral Recurrence and Death. *Clinical Cancer Research*, 26(5):1126–1134, March 2020.

[177] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, 2021.

[178] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification, 2021.

[179] Richard J. Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.e6, August 2022. Publisher: Elsevier.

[180] Oliver Lester Saldanha, Chiara M. L. Loeffler, Jan Moritz Niehues, Marko van Treeck, Tobias P. Seraphin, Katherine Jane Hewitt, Didem Cifci, Gregory Patrick Veldhuizen, Siddhi Ramesh, Alexander T. Pearson, and Jakob Nikolas Kather. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *npj Precision Oncology*, 7(1):1–5, March 2023. Publisher: Nature Publishing Group.

[181] Tobias Paul Seraphin, Mark Luedde, Christoph Roderburg, Marko van Treeck, Pascal Scheider, Roman D. Buelow, Peter Boor, Sven H. Loosen, Zdenek Provaznik, Daniel Mendelsohn, Filip Berisha, Christina Magnussen, Dirk Westermann, Tom Luedde, Christoph Brochhausen, Samuel Sossalla, and Jakob Nikolas Kather. Prediction of heart transplant rejection from routine pathology slides with self-supervised deep learning. *European Heart Journal. Digital Health*, 4(3):265–274, May 2023.

[182] Jan Moritz Niehues, Philip Quirke, Nicholas P. West, Heike I. Grabsch, Marko Van Treeck, Yoni Schirris, Gregory P. Veldhuizen, Gordon G.A. Hutchins, Susan D. Richman, Sebastian Foersch, Titus J. Brinker, Junya Fukuoka, Andrey Bychkov, Wataru Uegami, Daniel Truhn, Hermann Brenner, Alexander Brobeil, Michael Hoffmeister, and Jakob Nikolas Kather. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. *Cell Reports Medicine*, 4(4):100980, April 2023.

[183] Filmon Yacob, Jan Siarov, Kajsa Villiamsson, Juulia T. Suvilehto, Lisa Sjöblom, Magnus Kjellberg, and Noora Neittaanmäki. Weakly supervised detection and classification of basal cell carcinoma using graph-transformer on whole slide images. *Scientific Reports*, 13(1):7555, May 2023. Number: 1 Publisher: Nature Publishing Group.

[184] T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks, 2023.

[185] Randie H. Kim, Sofia Nomikou, Nicolas Coudray, George Jour, Zarmeena Dawood, Runyu Hong, Eduardo Esteva, Theodore Sakellaropoulos, Douglas Donnelly, Una Moran, Aristides Hatzimemos, Jeffrey S. Weber, Narges Razavian, Ioannis Aifantis, David Fenyo,

Matija Snuderl, Richard Shapiro, Russell S. Berman, Iman Osman, and Aristotelis Tsirigos. A Deep Learning Approach for Rapid Mutational Screening in Melanoma. *bioRxiv*, page 610311, August 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[186] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, December 2022. Number: 12 Publisher: Nature Publishing Group.

[187] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of Patch Features Through Self-supervised Learning and Transformer for Survival Analysis on Whole Slide Images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 561–570, Cham, 2021. Springer International Publishing.

[188] Julia A. Newton-Bishop, Samantha Beswick, Juliette Randerson-Moor, Yu-Mei Chang, Paul Affleck, Faye Elliott, May Chan, Susan Leake, Birute Karpavicius, Sue Haynes, Kairen Kukalizch, Linda Whitaker, Sharon Jackson, Edwina Gerry, Clarissa Nolan, Chandra Bertram, Jerry Marsden, David E. Elder, Jennifer H. Barrett, and D. Timothy Bishop. Serum 25-Hydroxyvitamin D3 Levels Are Associated With Breslow Thickness at Presentation and Survival From Melanoma. *Journal of Clinical Oncology*, 27(32):5439–5444, November 2009. Publisher: Wolters Kluwer.

[189] Julia A. Newton-Bishop, John R. Davies, Faheem Latheef, Juliette Randerson-Moor, May Chan, Jo Gascoyne, Saila Waseem, Susan Haynes, Charles O'Donovan, and D. Timothy Bishop. 25-Hydroxyvitamin D2/D3 levels and factors associated with systemic inflammation and melanoma survival in the Leeds Melanoma Cohort. *International Journal of Cancer*, 136(12):2890–2899, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.29334.

[190] Christine Delon, Katrina F. Brown, Nick W. S. Payne, Yannis Kotrotsios, Sally Vernon, and Jon Shelton. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *British Journal of Cancer*, 126(12):1765–1773, June 2022. Number: 12 Publisher: Nature Publishing Group.

[191] Andrew Knight, Lilit Karapetyan, and John M. Kirkwood. Immunotherapy in Melanoma: Recent Advances and Future Directions. *Cancers*, 15(4):1106, February 2023.

[192] M. R. Middleton, J. J. Grob, N. Aaronson, G. Fierlbeck, W. Tilgen, S. Seiter, M. Gore, S. Aamdal, J. Cebon, A. Coates, B. Dreno, M. Henz, D. Schadendorf, A. Kapp, J. Weiss, U. Fraass, P. Statkevich, M. Muller, and N. Thatcher. Randomized phase III study of

temozolomide versus dacarbazine in the treatment of patients with advanced metastatic malignant melanoma. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 18(1):158–166, January 2000.

[193] Agop Y. Bedikian, Michael Millward, Hubert Pehamberger, Robert Conry, Martin Gore, Uwe Trefzer, Anna C. Pavlick, Ronald DeConti, Evan M. Hersh, Peter Hersey, John M. Kirkwood, Frank G. Haluska, and Oblimersen Melanoma Study Group. Bcl-2 antisense (oblimersen sodium) plus dacarbazine in patients with advanced melanoma: the Oblimersen Melanoma Study Group. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 24(29):4738–4745, October 2006.

[194] Jedd D. Wolchok, Vanna Chiarion-Sileni, Rene Gonzalez, Jean-Jacques Grob, Piotr Rutkowski, Christopher D. Lao, Charles Lance Cowey, Dirk Schadendorf, John Wagstaff, Reinhard Dummer, Pier Francesco Ferrucci, Michael Smylie, Marcus O. Butler, Andrew Graham Hill, Ivan Marquez-Rodas, John B. A. G. Haanen, Tuba Bas, Wim van Dijck, James Larkin, and F. Stephen Hodi. CheckMate 067: 6.5-year outcomes in patients (pts) with advanced melanoma. *Journal of Clinical Oncology*, 39(15_suppl):9506–9506, May 2021. Publisher: Wolters Kluwer.

[195] Allison P. Heath, Vincent Ferretti, Stuti Agrawal, Maksim An, James C. Angelakos, Renuka Arya, Rosita Bajari, Bilal Baqar, Justin H. B. Barnowski, Jeffrey Burt, Ann Catton, Brandon F. Chan, Fay Chu, Kim Cullion, Tanja Davidsen, Phuong-My Do, Christian Dompierre, Martin L. Ferguson, Michael S. Fitzsimons, Michael Ford, Miyuki Fukuma, Sharon Gaheen, Gajanan L. Ganji, Tzintzuni I. Garcia, Sameera S. George, Daniela S. Gerhard, Francois Gerthoffert, Fauzi Gomez, Kang Han, Kyle M. Hernandez, Biju Issac, Richard Jackson, Mark A. Jensen, Sid Joshi, Ajinkya Kadam, Aishmit Khurana, Kyle M. J. Kim, Victoria E. Kraft, Shenglai Li, Tara M. Lichtenberg, Janice Lodato, Laxmi Lolla, Plamen Martinov, Jeffrey A. Mazzone, Daniel P. Miller, Ian Miller, Joshua S. Miller, Koji Miyauchi, Mark W. Murphy, Thomas Nullet, Rowland O. Ogwara, Francisco M. Ortuño, Jesús Pedrosa, Phuong L. Pham, Maxim Y. Popov, James J. Porter, Raymond Powell, Karl Rademacher, Colin P. Reid, Samantha Rich, Bessie Rogel, Himanso Sahni, Jeremiah H. Savage, Kyle A. Schmitt, Trevar J. Simmons, Joseph Sislow, Jonathan Spring, Lincoln Stein, Sean Sullivan, Yajing Tang, Mathangi Thiagarajan, Heather D. Troyer, Chang Wang, Zhining Wang, Bedford L. West, Alex Wilmer, Shane Wilson, Kaman Wu, William P. Wysocki, Linda Xiang, Joseph T. Yamada, Liming Yang, Christine Yu, Christina K. Yung, Jean Claude Zenklusen, Junjun Zhang, Zhenyu Zhang, Yuanheng Zhao, Ariz Zubair, Louis M. Staudt, and Robert L. Grossman. The NCI Genomic Data Commons. *Nature Genetics*, 53(3):257–262, March 2021. Number: 3 Publisher: Nature Publishing Group.

[196] Mirja Mittermaier, Marium M. Raza, and Joseph C. Kvedar. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine*, 6(1):1–3, June 2023. Number: 1 Publisher: Nature Publishing Group.

[197] David Wen, Saad M. Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K. Denniston, Xiaoxuan Liu, and Rubeta N. Matin. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, January 2022. Publisher: Elsevier.

[198] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton J. V. Campbell, Liron Pantanowitz, H. R. Tizhoosh, and Shahryar Rahnamayan. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic Pathology*, 18(1):67, May 2023.

[199] Shira Ronen, Rebecca C. Czaja, Natali Ronen, Cooley G. Pantazis, and Kenneth A. Iczkowski. Small Cell Variant of Metastatic Melanoma: A Mimicker of Lymphoblastic Leukemia/Lymphoma. *Dermatopathology*, 6(4):231–236, November 2019.

[200] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. *arXiv:2004.09666 [cs, eess, q-bio]*, May 2020. arXiv: 2004.09666.

[201] Gabriele Campanella, Vitor Werneck Krauss Silva, and Thomas J. Fuchs. Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology. *arXiv:1805.06983 [cs]*, September 2018. arXiv: 1805.06983.

[202] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, August 2019. Number: 8 Publisher: Nature Publishing Group.

[203] Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images, July 2021. arXiv:2107.09405 [cs, eess].

[204] Renan Valieris, Lucas Amaro, Cynthia Aparecida Bueno de Toledo Osório, Adriana Passos Bueno, Rafael Andres Rosales Mitrowsky, Dirce Maria Carraro, Diana Noronha Nunes, Emmanuel Dias-Neto, and Israel Tojal da Silva. Deep Learning Predicts Underlying Features on Pathology Images with Therapeutic Relevance for Breast and Gastric Cancer. *Cancers*, 12(12):3687, December 2020.

[205] Hang Li, Fan Yang, Yu Zhao, Xiaohan Xing, Jun Zhang, Mingxuan Gao, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Dt-mil: Deformable transformer for multi-instance learning on histopathological image. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 206–216, Cham, 2021. Springer International Publishing.

[206] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification, 2022.

[207] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning, 2022.

[208] Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. Ms-clam: Mixed supervision for the classification and localization of tumors in whole slide images. *Medical Image Analysis*, 85:102763, 2023.

[209] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images, 2023.

[210] Javad Noorbakhsh, Saman Farahmand, Ali Foroughi pour, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-ha, Kourosh Zarringhalam, and Jeffrey H. Chuang. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nature Communications*, 11(1):6367, December 2020.

[211] Samaneh Abbasi-Sureshjani, Anıl Yüce, Simon Schönenberger, Maris Skujevskis, Uwe Schalles, Fabien Gaire, and Konstanty Korski. Molecular Subtype Prediction for Breast Cancer Using H&E Specialized Backbone. In *Proceedings of the MICCAI Workshop on Computational Pathology*, pages 1–9. PMLR, September 2021. ISSN: 2640-3498.

[212] Charlie Saillard, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoit Schmauch, and Simon Jegou. Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. *arXiv:2109.05819 [cs, eess]*, September 2021. arXiv: 2109.05819.

[213] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation, 2018.

[214] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language Modeling with Gated Convolutional Networks. *arXiv:1612.08083 [cs]*, September 2017. arXiv: 1612.08083.

[215] Maksim Lapin, Matthias Hein, and Bernt Schiele. Loss functions for top-k error: Analysis and insights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1468–1477, 2016.

[216] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019.

[217] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2021.

[218] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention, 2021.

[219] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958.

[220] N. E. Breslow. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975. Publisher: [Wiley, International Statistical Institute (ISI)].

[221] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, February 1996.

[222] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in medicine*, 30(10):1105–1117, May 2011.

[223] W. H. Clark, D. E. Elder, D. Guerry, L. E. Braitman, B. J. Trock, D. Schultz, M. Synnestvedt, and A. C. Halpern. Model predicting survival in stage I melanoma based on tumor progression. *Journal of the National Cancer Institute*, 81(24):1893–1904, December 1989.

[224] Qian Sun, Hongyan Sun, Nan Wu, Lele Cong, and Xianling Cong. Prognostic Significance of Tumor-Infiltrating Lymphocyte Grade in Melanoma: A Meta-Analysis. *Dermatology (Basel, Switzerland)*, 236(6):481–492, 2020.

[225] Heather D. Couture. Deep Learning-Based Prediction of Molecular Tumor Biomarkers from H&E: A Practical Review. *Journal of Personalized Medicine*, 12(12):2022, December 2022. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[226] David Díaz-Vico and José R. Dorronsoro. Deep Least Squares Fisher Discriminant Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2752–2763, August 2020. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[227] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W. Dean Wallace, William Speier, and Corey W. Arnold. A Multi-resolution Model for Histopathology Image Classification and Localization with Multiple Instance Learning, November 2020. arXiv:2011.02679 [cs, eess].

[228] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *arXiv:1912.08937 [cs, q-bio]*, September 2020. arXiv: 1912.08937.

[229] Vincent Roullier, Olivier Lézoray, Vinh-Thong Ta, and Abderrahim Elmoataz. Multi-resolution graph-based analysis of histopathological whole slide images: application to mitotic cell extraction and visualization. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 35(7-8):603–615, 2011.

[230] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), February 2016. arXiv:1511.07289 [cs] version: 5.

[231] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[232] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[233] Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, April 2019. arXiv:1903.02428 [cs, stat].

[234] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. Cuda, release: 10.2.89, 2020.

[235] Jingjia Huang, Zhangheng Li, Nannan Li, Shan Liu, and Ge Li. AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Networks via Attention Mechanism. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6479–6488, Seoul, Korea (South), October 2019. IEEE.

[236] Ruining Deng, Can Cui, Lucas W. Remedios, Shunxing Bao, R. Michael Womick, Sophie Chiron, Jia Li, Joseph T. Roland, Ken S. Lau, Qi Liu, Keith T. Wilson, Yaohong Wang, Lori A. Coburn, Bennett A. Landman, and Yuankai Huo. Cross-scale multi-instance learning for pathological image diagnosis. *Medical Image Analysis*, 94:103124, May 2024.

[237] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A graph-transformer for whole slide image classification, May 2022. arXiv:2205.09671 [cs].

[238] Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multi-scale Efficient Graph-Transformer for Whole Slide Image Classification, May 2023. arXiv:2305.15773 [cs].

[239] Alexander Breslow. Thickness, Cross-Sectional Areas and Depth of Invasion in the Prognosis of Cutaneous Melanoma. *Annals of Surgery*, 172(5):902, November 1970.

[240] David C. Whiteman, Adele C. Green, and Catherine M. Olsen. The Growing Burden of Invasive Melanoma: Projections of Incidence Rates and Numbers of New Cases in Six Susceptible Populations through 2031. *The Journal of Investigative Dermatology*, 136(6):1161–1171, June 2016.

[241] C. A. Drukker, S. G. Elias, M. V. Nijenhuis, J. Wesseling, H. Bartelink, P. Elkhuizen, B. Fowble, P. W. Whitworth, R. R. Patel, F. A. de Snoo, L. J. van 't Veer, P. D. Beitsch, and E. J. Th Rutgers. Gene expression profiling to predict the risk of locoregional recurrence in breast cancer: a pooled analysis. *Breast Cancer Research and Treatment*, 148(3):599–613, December 2014.

[242] Jia-Xing Zhang, Wu Song, Zhen-Hua Chen, Jin-Huan Wei, Yi-Ji Liao, Jian Lei, Ming Hu, Geng-Zhen Chen, Bing Liao, Jian Lu, Hong-Wei Zhao, Wei Chen, Yu-Long He, Hui-Yun Wang, Dan Xie, and Jun-Hang Luo. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *The Lancet. Oncology*, 14(13):1295–1306, December 2013.

[243] Daniel S. Widmer, Keith S. Hoek, Phil F. Cheng, Ossia M. Eichhoff, Thomas Biedermann, Marieke I. G. Raaijmakers, Silvio Hemmi, Reinhard Dummer, and Mitchell P. Levesque. Hypoxia Contributes to Melanoma Heterogeneity by Triggering HIF1-Dependent Phenotype Switching. *Journal of Investigative Dermatology*, 133(10):2436–2443, October 2013.

[244] Tao Li, Peizhen Xie, Jie Liu, Mingliang Chen, Shuang Zhao, Wenjie Kang, Ke Zuo, and Fangfang Li. Automated Diagnosis and Localization of Melanoma from Skin Histopathology Slides Using Deep Learning: A Multicenter Study. *Journal of Healthcare Engineering*, 2021:5972962, 2021.

[245] Yongyang Bao, Jiayi Zhang, Xingyu Zhao, Henghua Zhou, Ying Chen, Junming Jian, Tianlei Shi, and Xin Gao. Deep learning-based fully automated diagnosis of melanocytic lesions by using whole slide images. *The Journal of Dermatological Treatment*, 33(5):2571–2577, August 2022.

[246] Mike van Zon, Nikolas Stathonikos, Willeke A.M. Blokx, Selim Komina, Sybren L.N Maas, Josien P.W. Pluim, Paul J. van Diest, and Mitko Veta. Segmentation and Classification of Melanoma and Nevus in Whole Slide Images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 263–266, April 2020. ISSN: 1945-8452.

[247] Achim Hekler, Jochen S. Utikal, Alexander H. Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Dirk Schadendorf, Wiebke Sondermann, Cindy Franklin, Felix Bestvater, Michael J. Flaig, Dieter Krahl, Christof Von Kalle, Stefan Fröhling, and Titus J. Brinker. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118:91–96, September 2019.

[248] Francesco De Logu, Filippo Ugolini, Vincenza Maio, Sara Simi, Antonio Cossu, Daniela Massi, Romina Nassini, and Marco Laurino. Recognition of Cutaneous Melanoma on Digitized Histopathological Slides via Artificial Intelligence Algorithm. *Frontiers in Oncology*, 10:1559, August 2020.

[249] Julianna D. Ianni, Rajath E. Soans, Sivaramakrishnan Sankarapandian, Ramachandra Vikas Chamarthi, Devi Ayyagari, Thomas G. Olsen, Michael J. Bonham, Coleman C. Stavish, Kiran Motaparthi, Clay J. Cockerell, Theresa A. Feeser, and Jason B. Lee. Tailored for Real-World: A Whole Slide Image Classification System Validated on Uncurated Multi-Site Data Emulating the Prospective Pathology Workload. *Scientific Reports*, 10(1):3217, February 2020. Number: 1 Publisher: Nature Publishing Group.

[250] Dong Zhang, Hongcheng Han, Shaoyi Du, Longfei Zhu, Jing Yang, Xijing Wang, Lin Wang, and Meifeng Xu. MPMR: Multi-Scale Feature and Probability Map for Melanoma Recognition. *Frontiers in Medicine*, 8, 2022.

[251] Titus J. Brinker, Lennard Kiehl, Max Schmitt, Tanja B. Jutzi, Eva I. Krieghoff-Henning, Dieter Krahl, Heinz Kutzner, Patrick Gholam, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, Achim Hekler, Stefan Fröhling, Jakob N. Kather, Sarah Haggenmüller, Christof von Kalle, Markus Heppt, Franz Hilke, Kamran Ghoreschi, Markus Tiemann, Ulrike Wehkamp, Axel Hauschild, Michael Weichenthal, and Jochen S. Utikal. Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. *European Journal of Cancer (Oxford, England: 1990)*, 154:227–234, September 2021.

[252] Maria Colomba Comes, Livia Fucci, Fabio Mele, Samantha Bove, Cristian Cristofaro, Ivana De Risi, Annarita Fanizzi, Martina Milella, Sabino Strippoli, Alfredo Zito, Michele Guida, and Raffaella Massafra. A deep learning model based on whole slide images to predict disease-free survival in cutaneous melanoma patients. *Scientific Reports*, 12(1):20366, November 2022. Number: 1 Publisher: Nature Publishing Group.

[253] Justin Couetil, Ziyu Liu, Kun Huang, Jie Zhang, and Ahmed K. Alomari. Predicting melanoma survival and metastasis with interpretable histopathological features and machine learning models. *Frontiers in Medicine*, 9, January 2023. Publisher: Frontiers.

[254] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph CNN for Survival Analysis on Whole Slide Pathological Images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Lecture Notes in Computer Science, pages 174–182, Cham, 2018. Springer International Publishing.

[255] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[256] Mei-Yin C. Polley and James J. Dignam. Statistical Considerations in the Evaluation of Continuous Biomarkers. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 62(5):605–611, May 2021.

[257] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835, June 1994.

[258] Glenda G. Callender, Michael E. Egger, Alison L. Burton, Charles R. Scoggins, Merrick I. Ross, Arnold J. Stromberg, Lee Hagendoorn, Robert C. G. Martin, and Kelly M. McMasters. Prognostic implications of anatomic location of primary cutaneous melanoma of 1 mm or thicker. *American Journal of Surgery*, 202(6):659–664; discussion 664–665, December 2011.

[259] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[260] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[261] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[262] Stefani Spranger, Daisy Dai, Brendan Horton, and Thomas F. Gajewski. Tumor-Residing Batf3 Dendritic Cells Are Required for Effector T Cell Trafficking and Adoptive T Cell Therapy. *Cancer Cell*, 31(5):711–723.e4, May 2017. Publisher: Elsevier.

[263] Evan J. Lipson and Charles G. Drake. Ipilimumab: an anti-CTLA-4 antibody for metastatic melanoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 17(22):6958–6962, November 2011.

[264] Ming Liu, Xu Wang, Lei Wang, Xiaodong Ma, Zhaojian Gong, Shanshan Zhang, and Yong Li. Targeting the IDO1 pathway in cancer: from bench to bedside. *Journal of Hematology & Oncology*, 11(1):100, August 2018.

[265] Jiayuan Wu, Caixia Liang, Manyu Chen, and Wenmei Su. Association between tumor-stroma ratio and prognosis in solid tumor patients: a systematic review and meta-analysis. *Oncotarget*, 7(42):68954–68965, September 2016.

[266] Sander Greenland and Neil Pearce. Statistical Foundations for Model-Based Adjustments. *Annual Review of Public Health*, 36(1):89–108, 2015. _eprint: https://doi.org/10.1146/annurev-publhealth-031914-122559.

[267] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. Publisher: Nature Publishing Group.

[268] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, and Martin Eklund. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, January 2022. Publisher: Nature Publishing Group.

[269] Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. DAS-MIL: Distilling Across Scales for MIL Classification of Histological WSIs. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 248–258, Cham, 2023. Springer Nature Switzerland.

[270] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3851–3860, Seattle, WA, USA, June 2020. IEEE.

[271] Ali Khajegili Mirabadi, Graham Archibald, Amirali Darbandsari, Alberto Contreras-Sanz, Ramin Ebrahim Nakhli, Maryam Asadi, Allen Zhang, C. Blake Gilks, Peter Black, Gang Wang, Hossein Farahani, and Ali Bashashati. GRASP: GRAph-Structured Pyramidal Whole Slide Image Representation, February 2024. arXiv:2402.03592 [cs].

[272] Ethar Alzaid, Muhammad Dawood, and Fayyaz Minhas. A transductive approach to survival ranking for cancer risk stratification. In David A. Knowles and Sara Mostafavi, editors, *Proceedings of the 18th Machine Learning in Computational Biology meeting*, volume 240 of *Proceedings of Machine Learning Research*, pages 101–109. PMLR, 30 Nov–01 Dec 2024.

[273] Christian Thiele and Gerrit Hirschfeld. cutpointr: Improved estimation and validation of optimal cutpoints in r. *Journal of Statistical Software*, 98(11):1–27, 2021.

[274] Mónica López-Ratón, María Xosé Rodríguez-Álvarez, Carmen Cadarso-Suárez, and Francisco Gude-Sampedro. Optimalcutpoints: An r package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36, 2014.

[275] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 186–195, Cham, 2021. Springer International Publishing.

[276] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling, September 2023. Pages: 2023.07.21.23292757.

[277] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, January 2024. arXiv:2303.00915 [cs].

[278] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29(9):2307–2316, September 2023. Number: 9 Publisher: Nature Publishing Group.

[279] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A Million-Slide Digital Pathology Foundation Model, January 2024. arXiv:2309.07778 [cs, eess, q-bio].

[280] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. arXiv:2304.07193 [cs].

[281] Patrice Y. Simard, Yann A. LeCun, John S. Denker, and Bernard Victorri. Transformation Invariance in Pattern Recognition – Tangent Distance and Tangent Propagation. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 235–269. Springer, Berlin, Heidelberg, 2012.

[282] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. Conference Name: Proceedings of the IEEE.

[283] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[284] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[285] Michael Gadermayr, Lukas Koller, Maximilian Tschuchnig, Lea Maria Stangassinger, Christina Kreutzer, Sebastien Couillard-Despres, Gertie Janneke Oostingh, and Anton Hittmair. Mixup-mil: A study on linear multilinear interpolation-based data augmentation for whole slide image classification, 2023.

[286] Imaad Zaffar, Guillaume Jaume, Nasir Rajpoot, and Faisal Mahmood. Embedding Space Augmentation for Weakly Supervised Learning in Whole-Slide Images, October 2022. arXiv:2210.17013 [cs].

[287] Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela

Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Florian Buettner, M. Jorge Cardoso, Veronika Cheplygina, Jianxu Chen, Evangelia Christodoulou, Beth A. Cimini, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, Patrick Godau, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Jens Kleesiek, Florian Kofler, Thijs Kooi, Annette Kopp-Schneider, Michal Kozubek, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Susanne M. Rafelski, Nasir Rajpoot, Mauricio Reyes, Michael A. Riegler, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, Ziv R. Yaniv, Paul F. Jäger, and Lena Maier-Hein. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21(2):182–194, February 2024. Publisher: Nature Publishing Group.

[288] Xiaoyuan Li, Xiaoqian Yu, Duanliang Tian, Yiran Liu, and Ding Li. Exploring and validating the prognostic value of pathomics signatures and genomics in patients with cutaneous melanoma based on bioinformatics and deep learning. *Medical Physics*, 50(11):7049–7059, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16748.

Appendices

Appendix A

Multi-level graph representation additional results

Further evidence that multi-level graph models learn high attention scores for different regions depending on the resolution. Fig. A.1, shows a correctly classified high immune case with the six top high attention patches. There is little overlap of 40x and 10x high attention patches. Moreover the high attention patches appear to contain tumour infiltrating lymphocytes, suggesting the model is classifying high immune cases based on the presence of TILs.

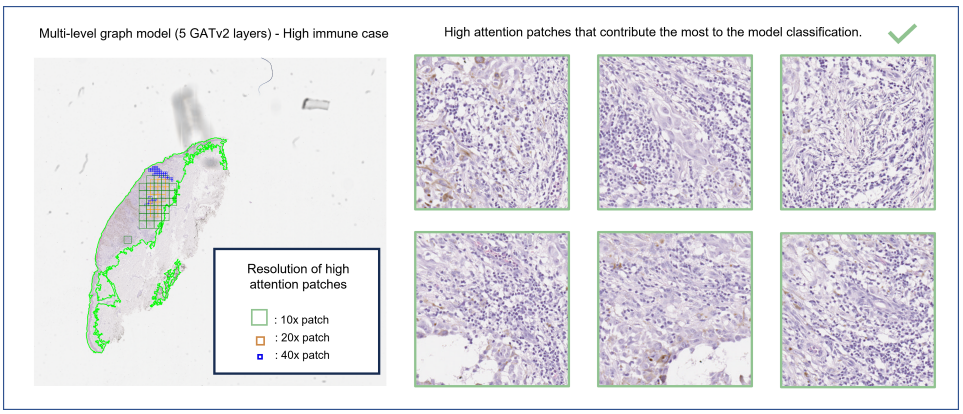


Figure A.1: Additional example of high attention regions generated from the best performing multi-level attention GNN. High immune H&E stained image with segmented tissue are shown annotated with the top 5% of patches which generated features with high attention scores. The original H&E stained WSI is shown next to the six patches which generated the highest attention scores. H&E, Haematoxylin and Eosin; GNN, Graph Neural Network; WSI, Whole Slide Image.

A. MULTI-LEVEL GRAPH REPRESENTATION ADDITIONAL RESULTS

Appendix B

Immune, clinical and clinicohistological features of the immune and risk subgroups

The immune subgroups from the study by Pozniak *et al.* (2019) are often referenced as having differing levels of treatment targets, in Fig. B.1 we show the expression of seven immune checkpoint and inhibitory molecules based on the transcriptomic data [1]. We also looked at a combined score called the 'immune score', which was a sum of all 27 immune cells used to generate the Pozniak *et al.* three immune subgroups [1]. The expression values for each molecule and the immune score were then dichotomised based on the median expression score into low and high inferred expression groups. As is expected the high immune subgroup has high proportions of the potential immune-related treatment targets, conversely the 'low immune' subset of patients has a high proportion of patients with low expression of these treatment targets.

In Table B.1, we show the clinicohistological features of the two risk groups generated from the 40x attention-based SurvGNN model. While this model did not generate risk groups with the greatest HR (1.79 [95% CI 1.39, 2.33]), the risk groups show similar proportions of patients with ulcerated tumours, as the risk groups from non-attention based model which generated risk groups with the highest HR (1.96 [95% CI 1.51, 2.56]). We believe this could indicate that the SurvGNN models are stratifying patient tumours, based on histological features that are associated with ulceration or more aggressive tumours, rather than immune cell expression.

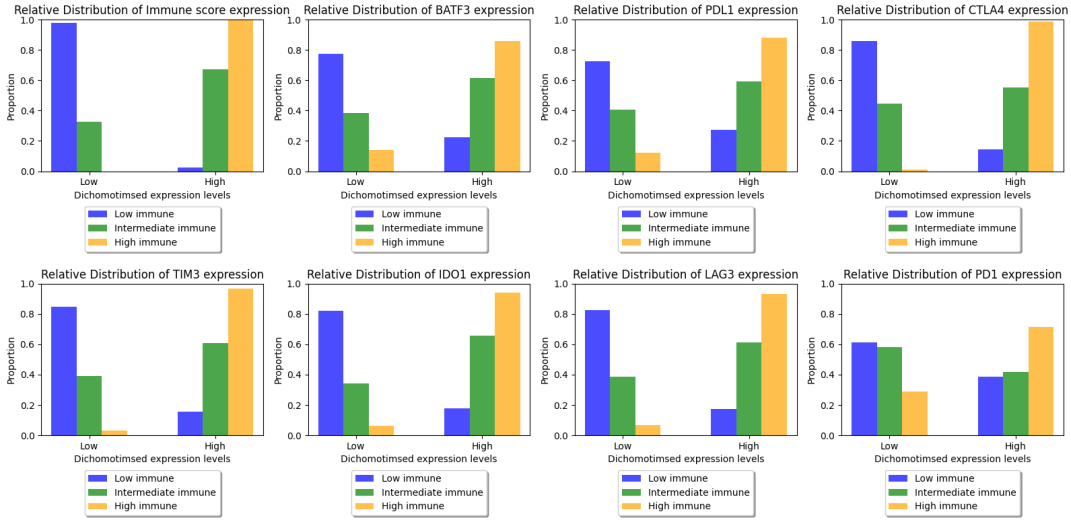


Figure B.1: Relative distributions of potential treatments targets and immune biomarkers for patients in the three immune subgroups derived by Pozniak *et al.* (2019) [1].

Table B.1: Summary data of the clinical and clinicohistological features of the high and low risk LMC patients from the attention-based 40x SurvGNN model. AJCC, American Joint Committee on Cancer; LMC, Leeds Melanoma Cohort; SurvGNN, Survival-based Graph Neural Network; TIL, Tumour Infiltrating Lymphocyte; WSI, Whole slide image.

	Low risk	High risk
Number of tumours	469	199
Melanoma death (%)	24	42
Age at diagnosis (median, years)	55	61
Sex (% males)	45	49
Breslow thickness (median, mm)	2.0	3.2
Ulcerated (%)	24.9	54.3
Mitotic rate (median count, mm^2)	< 1	< 1
AJCC stage (%)	Low risk	High risk
I	40	10
II	45	66
III	15	21
Unknown	0	4
TIL classification (%)	Low risk	High risk
Brisk	13	4
Non-brisk	48	51
Absent	11	16
Unclassified	28	30
Immune subgroup [1] (%)	Low risk	High risk
Low	36	45
Intermediate	41	34
High	23	10
Histological subtype (%)	Low risk	High risk
Superficial spreading	53	32
Nodular	28	42
Lentigo maligna	1	2
Acral lentiginous	3	5
Other	6	7
Unclassified	6	11