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

Lucy Godson[1(✉)], Navid Alemi[1], Jérémie Nsengimana[5], Graham P. Cook[3], Emily L. Clarke[2,4], Darren Treanor[2,4], D. Timothy Bishop[3], Julia Newton-Bishop[3], and Derek Magee[1]

[1] School of Computing, University of Leeds, Woodhouse, Leeds, LS2 9JT, UK
`sclg@leeds.ac.uk`

[2] Division of Pathology and Data Analytics, Leeds Institute of Cancer and Pathology, University of Leeds, Beckett Street, Leeds, LS9 7TF, UK

[3] Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, LS2 9JT, UK

[4] Department of Histopathology, Leeds Teaching Hospitals Trust, Leeds, LS2 9JT, UK

[5] Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Abstract. Stratifying melanoma patients into immune subgroups is important for understanding patient outcomes and treatment options. Current weakly supervised classification methods often involve dividing digitised whole slide images into patches, which leads to the loss of important contextual diagnostic information. Here, we propose using graph attention neural networks, which utilise graph representations of whole slide images, to introduce context to classifications. In addition, we present a novel hierarchical graph approach, which leverages histopathological features from multiple resolutions to improve on state-of-the-art (SOTA) multiple instance learning (MIL) methods. We achieve a mean test area under the curve metric of 0.80 for classifying low and high immune melanoma subtypes, using multi-level and 20x patch graph representations of whole slide images, compared to 0.77 when using SOTA MIL methods. Our experimental results comprehensively show how our whole slide image graph representation is a valuable improvement on the MIL paradigm and could help to determine early-stage prognostic markers and stratify melanoma patients for effective treatments. Code is available at `https://github.com/lucyOCg/graph_mil_project/`.

Keywords: Computational pathology · graph neural networks · melanoma.

1 Introduction

Melanoma is the most aggressive form of skin cancer [1], however, immunotherapy has revolutionised the treatment of patients [8]. Yet, the most effective and well tolerated drug, PD-1 blockade only benefits around 35% of patients [23]. Increasing our understanding of the interaction between immune and tumour cells and being able to identify disease subtypes is vital for stratifying patients into effective treatment groups and improving outcomes. Through consensus clustering of patient transcriptomes, previous studies have found distinct immunological subgroups within a population ascertained cohort (the Leeds Melanoma Cohort [LMC]), with differing clinical outcomes and potential treatment targets [20,22].

While tumour transcriptome data analysis is not routinely carried out for melanoma patients, haematoxylin and eosin (H&E) histopathology slides are well established in the diagnostic workflow of patients. Moreover, convolutional neural networks (CNNs), have been shown to identify molecular immune cell signatures from morphological patterns in digitised WSIs [26]. Analysis of these whole slide images (WSIs) can be challenging, as they are multi-resolution and multi-gigabyte, so current classification methods such as multiple instance learning (MIL) divide these images into patches, which are processed individually, but this leads to a loss of contextual information.

Here we propose 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 cell-graphs [16,18,21,28], developing a multi-level approach, which exploits the inherent hierarchical relationships between features extracted from patches at different resolutions. We believe that the addition of graph features increases the contextual information learned by the models for classifying melanomas into immune subgroups and demonstrate how this leads to improved performance over state-of-the-art MIL methods for our novel application.

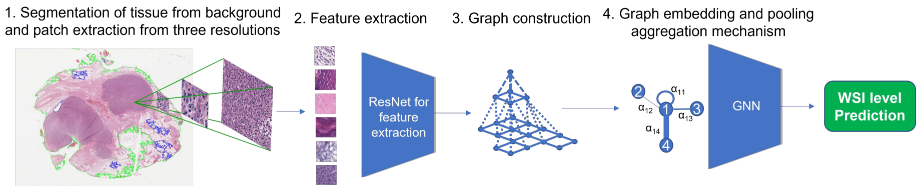

1. Segmentation of tissue from background and patch extraction from three resolutions
2. Feature extraction
3. Graph construction
4. Graph embedding and pooling aggregation mechanism
ResNet for feature extraction
2
1
3
4
α_{11}
α_{12}
α_{13}
α_{14}
GNN
WSI level Prediction

Fig. 1. Experimental framework for classifying the WSIs, using graph representations.

2 Related works

WSIs are multi-resolution gigapixel images, which means they can be computationally and memory intensive to process, especially on GPUs. Moreover, most WSIs will only have slide-level labels, as pixel-wise labelling can be time consuming for a pathologist. MIL frameworks, where an image is treated as a bag with instances which inherit the bag or slide-level label [9], have been applied with high accuracy in computational pathology tasks for classifying WSIs [4,17,25]. In this method, 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 pooling functions, such as maximum or mean operators [15], can be applied to the features to estimate the slide-level label classification.

While weakly supervised methods like MIL address problems such as the lack of annotation within WSIs and the aggregation of patch instances within an image, they also lead to a loss of contextual information. Each patch-instance or patch-feature is treated individually as it is passed through a network, with the focus being on the local visual and morphological patterns within the patch region. This means we lose global information about the tumour architecture, which can be important when evaluating the role of immune cells within the tumour microenvironment (TME), as they can differ depending on their spatial arrangement, locations and interactions with other cell types within the TME.

One way to resolve this, is through representing histopathological features in a graph structure and using a graph pooling mechanism. Recent works show how graph neural networks (GNN) can be a powerful tool for WSI analysis and subtyping [5,18,21]. However, these methods require sampling techniques to select patches due to memory constraints and therefore lose the overall WSI structure. Moreover, previous works by [16,18,21,28], who use hierarchical cell graphs show how the combination of multi-level features can improve model performance by providing additional cellular context.

As early as 2011, [24] showed how multi-scale WSI analysis could be useful in segmentation tasks, demonstrating that lower resolution levels could be used for removing the background from a WSI and how higher resolution levels could be used to detect tumour and classify mitotic figures within the tissue. A more recent study by [13], showed how GNNs using multi-scale WSI graph representations using embeddings from “Thumbnail”, ×5, ×10 scales as nodes, could be used for grading and subtyping esophageal cancer and kidney cancer TCGA datasets. Here we build on this idea, but also look at how one-hot encodings for node and edge features can be used to generate novel patch-based graph representations, which have a global hierarchical “multi-level" structure.

3 Dataset

The dataset used for our work was from the LMC [19]. This is a population ascertained cohort, including 667 digitised WSIs of primary melanoma tumours. The labels for the images were delineated by clustering transcriptomes, based on the inferred abundance of 27 immune cell types [22]. The three subgroups are the “high immune” class (22.5%), which corresponds to 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 has the least inferred immune cell infiltrate in the tumour and worst survival response of patients. We worked under the assumption that each group with a distinct immune genetic signature would have a distinct his-

Table 1. Summary data of the clinical and clinicohistological features of the three LMC immune subgroups found in the paper published by [22].

	Low	Intermediate	High
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	Intermediate	High
I	30	36	39
II	50	50	47
III	18	13	14
Unknown	2	2	1
TIL classification (%)	Low	Intermediate	High
Brisk	12	8	13
Non-brisk	44	50	49
Absent	14	11	10
Unclassified	30	31	28

tological pattern. Initial experiments were carried out 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.

All slides come from Formalin-Fixed Paraffin-Embedded (FFPE) blocks and were scanned in batches using a Leica Biosystems Aperio Digital Pathology Slide Scanner, at 0.25 micrometers-per-pixel (m.p.p.). The tumour transcriptomic data that was used to develop the immune subgroup labels was produced from the archived FFPE tumour blocks, using Illumina array DASL HT12.4 and normalised using standard methods as described in the study by [20].

We show the clinical and prognostic features for the different subgroups in Table 1. Breslow thickness describes how deep the tumour has grown into of the skin (epidermis) and is measured in millimetres (mm). Depth has been shown to have a continuous association with patient prognosis [2]. Microscopic ulceration (shown as "Ulcerated" in Table 2) is characterised by a lack of intact epidermis and have reactive changes within the skin. Ulceration has also been shown to be a strong independent predictor of melanoma death [12]. Studies have also demonstrated that brisk TILs, where there is robust infiltration of TILs throughout the entire tumour or surrounding the tumour base, are associated with improved patient prognosis [11]. However, due to inter-observer variation and lack of standardisation in grading they are currently not included in the current AJCC staging systems. Where the TIL classification is stated as "Unclassified", it indicates that the pathologists assessing the tumour are undecided between brisk and non-brisk, but TILs are definitely present.

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, while the LMC 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 [1]. As a result of this, 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. In addition because the study was carried out before the new AJCC staging 8th edition was established, in Table 1 we show the 7th edition staging for patients.

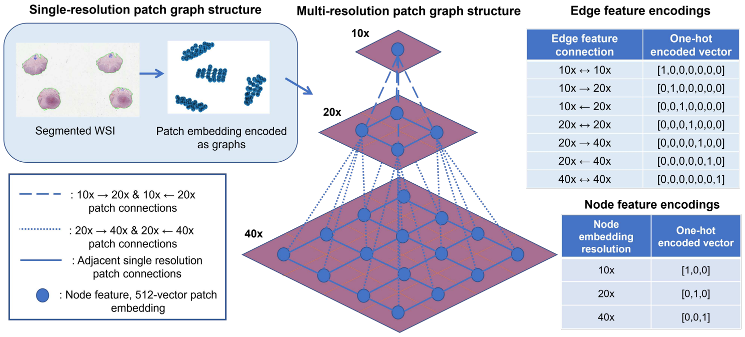
4 Methods

4.1 Segmentation and feature extraction

The H&E stained tissue in the WSI was segmented from the background using thresholding and morphological operations described by [17]. The segmented tissue was then split into 256 pixel × 256 pixel non-overlapping patches at three different resolutions (10x, 20x and 40x). A ResNet18, which had been pretrained by self-supervised learning on 57 multi-organ, multi-resolution histopathology datasets by [6], was used to extract 512-dimensional feature embeddings from the patches.

4.2 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 patches that are to the left, right, above, or below each other (diagonal edges 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 4 potential corresponding 20x patch nodes in the higher resolution level. Then defining edge connections between the 20x patch nodes to the 4 potential corresponding 40x patch nodes to create hierarchical multi-level 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]). 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. 2).


Single-resolution patch graph structure
Segmented WSI
Patch embedding encoded as graphs
Multi-resolution patch graph structure
10x
20x
40x
: 10x → 20x & 10x ← 20x patch connections
: 20x → 40x & 20x ← 40x patch connections
: Adjacent single resolution patch connections
: Node feature, 512-vector patch embedding

Edge feature encodings

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

Node feature encodings

Node embedding resolution	One-hot encoded vector
10x	[1,0,0]
20x	[0,1,0]
40x	[0,0,1]

Fig. 2. Single and multi-resolution patch graph representation construction and edge feature encodings used for multi-resolution graphs. Self-loops are not visualised.

4.3 Model architecture

When implementing the GNNs, we used GATv2 graph attention layers with one attention head [3]. Edge indices, node and edge features were passed through the GATv2 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:

$$e(\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{e}_{ij}) = LeakyReLU((\boldsymbol{a} \odot \boldsymbol{W})^T [\boldsymbol{h}_i || \boldsymbol{h}_j || \boldsymbol{e}_{ij}]) \tag{1}$$

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:

$$softmax_j(\boldsymbol{a}_{ij}) = \frac{exp(e(\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{e}_{ij}))}{\sum_{k \in \mathcal{N}_i} exp(e(\boldsymbol{h}_i, \boldsymbol{h}_k, \boldsymbol{e}_{ik}))} \tag{2}$$

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 represented by:

$$\boldsymbol{h_i'} = \sigma\left(\sum_{j\in\mathcal{N}_i} e(\boldsymbol{h}_i, \boldsymbol{h}_i, \boldsymbol{e}_{ij}) \cdot (\mathbf{W} \odot \mathbf{a})h_j\right) \tag{3}$$

Where $\boldsymbol{h_i'}$ is the updated representation of node $\boldsymbol{i}$ after the GATv2 layer operation. σ represents the Exponential Linear Unit (ELU) activation function [7], which was followed by a global mean pooling operation. The number of GATv2 layers was varied to assess their impact on performance. 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 to give the classification output (the Network architecture is shown in Supplementary material Figure 1). We compared the GNNs, to four weakly supervised MIL methods, including a max-pooling MIL model, a gated-attention MIL model (gated AttMIL) [15], an attention MIL (AttMIL) [15] and a clustering-constrained attention MIL model (CLAM) [17].

4.4 Model training

The networks were trained using a cross-entropy loss function, comparing the ground truth immune subtype slide label with the predicted slide-level label. A learning rate of 0.0002 and weight decay of $1 \times 10e^{-5}$ were applied. Dropout with a probability of 0.5 was used after each global mean pooling layer. During training, to mitigate class imbalances, 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 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 were trained for a minimum of 50 epochs, with early stopping if the validation loss did not improve for 20 epochs continuously.

Graph models were implemented using PyTorch 1.13.1, PyTorch geometric version 2.2.0 with CUDA 11.6, 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 with CUDA version 11.0, using one Tesla V100 32G NVLink 2.0 GPU from the Bede N8 HPC facility.

5 Results

For classifying the three immune subtypes, the combination of graph and visual features led to improved performance, when using single resolution and multi-resolution graphs with one-hot encoded node embeddings as input (Table 2). With the best performing network 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.

Table 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. "One-hot node" indicates multi-resolution graphs with one-hot encodings added to the node patch embeddings.

Resolution	Max-pooling MIL	Gated AttMIL	AttMIL	CLAM	Proposed GNN
10x	0.57±0.022	0.60±0.026	0.61±0.039	0.60±0.028	**0.63±0.023**
20x	0.55±0.026	0.57±0.035	0.59±0.027	0.57±0.034	0.62±0.021
40x	0.48±0.023	0.55±0.036	0.54±0.023	0.55±0.035	0.61±0.027
10x+20x+40x	-	-	-	-	0.61±0.025
One-hot node	-	-	-	-	0.62±0.017

When examining model performance for classifying "high" and "low" immune subtypes, we found using models trained with both 10x and 20x single resolution graphs and multi-resolution graphs outperformed current SOTA MIL models (Table 3). 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-resolution WSI graph representations which included one-hot encodings to represent different resolution patches in the embeddings.

Table 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. "one-hot node" indicates multi-resolution graphs with one-hot encodings added to the node patch embeddings. The 10x and 20x use three GATv2 layers, the 10x+20x+40x use two GATv2 layers and the GNNs with "one-hot node" and 40x embeddings use four GATv2 layers.

Resolution	Max-pooling MIL	Gated AttMIL	AttMIL	CLAM	Proposed GNN
10x	0.51±0.069	0.77±0.042	0.75±0.043	0.73±0.055	0.78±0.033
20x	0.57±0.090	0.75±0.072	0.74±0.048	0.75±0.053	**0.80±0.052**
40x	0.48±0.049	0.70±0.048	0.76±0.029	0.67±0.073	0.77±0.056
10x+20x+40x	-	-	-	-	0.78±0.042
One-hot node	-	-	-	-	**0.80±0.048**

Moreover, we tested how increasing the number of GATv2 layers affected model performance. We also explored using only the 10x and 20x patch and graph features together, to test whether 40x features were leading to a decrease in performance (Table 4). We found that overall, three GATv2 layers produced the best performance for the single and two resolution GNN models. However, when implementing four GATv2 layers, the models trained with multi-resolution one-hot encoded node embeddings achieved the equally highest mean test AUC score of 0.80 (95% CI 0.75 to 0.85). We examined the effect of adding five GATv2 layers for this input, but found this did not further increase performance, as the AUC was 0.79 (95% CI 0.71 to 0.87) (Supplementary material Table 1).

Table 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-resolution graphs with one-hot encodings added to the node patch embeddings.

Resolution	1 GATv2 layer	2 GATv2 layers	3 GATv2 layers	4 GATv2 layers
10x	0.76±0.050	0.76±0.041	0.78±0.033	0.78±0.038
20x	0.73±0.061	0.78±0.045	**0.80±0.052**	0.78±0.068
40x	0.68±0.061	0.75±0.050	0.75±0.084	0.77±0.056
10x+20x	0.73±0.056	0.76±0.042	0.79±0.033	0.77±0.044
10x+20x+40x	**0.77±0.091**	**0.78±0.042**	0.77±0.088	0.77±0.088
One-hot node	0.73±0.065	0.75±0.077	0.78±0.076	**0.80±0.048**

Notably, for the three and two immune subtype classification tasks, the types of graph representations that generated the best performance were not consistent (Table 3 & 4). For the three immune subtype classification task, the 10x patch graph (test mean AUC of 0.63 [95% CI CI 0.61 to 0.65]), followed by the 20x and one-hot node encoded graph representations generated the best test mean AUC performances of 0.62 (Table 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 to 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 highest AUC performance (Table 4). We believe this may be the case, as for this task the presence of immune cells in the TME is important for classification of the "high immune" class, therefore, cellular and spatial detail is important. As graph representations provide spatial context and higher resolution patches provide cellular detail, a trade-off between enough cellular detail and less uninformative patch node embeddings, which introduce noise due to classifications, is required. Hence, in Table 4 for single-resolution graphs, it appears that 20x patch graphs generate the best performance as 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 graph representations, while having the disadvantage of more noisy and heterogeneous nodes, also have additional 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).

6 Discussion and Conclusion

Recent studies [20,22] have shown that melanoma patients can be stratified into subgroups, with added prognostic value compared to AJCC melanoma staging systems [12]. However, these studies are carried out using transcriptomic data, which can be expensive and time consuming to analyse. In this paper we show that routinely used H&E images can be used to develop models that classify patients into these immune subgroups. We show how GNNs with graph representations of WSIs can improve performance over current state-of-the-art MIL methods for classifying melanoma WSIs into immune subgroups. While performance does not increase beyond a mean test AUC of 0.63, for classifying the three immune subtypes, we show that GNN models lead to improved performance. However, in order to improve performance further, we will need to decipher tumour heterogeneity and complexity within the "intermediate" subgroup. To tackle this problem, we may need to look at further dividing this subgroup, as a previous study by [20] found two distinct subgroups which overlap with this "intermediate" group, or look at different techniques to learn more discriminant representations of the images.

Moreover, for the task of classifying "high immune" and "low immune" subtypes we show that 20x graph representations and the one-hot node encoded multi-level graph representation generate the superior performance, with a mean test AUC performance of 0.80. In agreement with findings by [28], our study also demonstrates that increasing the number of GATv2 message passing layers in models appears to enhance information transfer through the network, leading to increased performance when using one-hot encoded node embeddings (Table 4). We also see this trend with 40x resolution graphs, suggesting these larger graph structures, containing 40x node embeddings, benefit more from increased message passing layers for information transfer.

We have also identified a limitation in the performance of our multi-level graph approach due to the graph mean pooling mechanism, where all node embeddings are averaged together and equally contribute to the final slide-level prediction. The graph representations we have constructed utilise every patch extracted from the segmented image tissue. However, it is important to acknowledge that these representations may include patches that do not contribute meaningful information and can consequently result in misclassifications. This issue becomes particularly challenging for our task due to high levels of tumoural heterogeneity and because the ground truth labels are derived from small 0.6-mm regions of the tumours. As a result, the multi-level graph representations are more likely to contain noisy instances due to the increased number of patch inputs from all three resolutions, further exacerbating the potential for misclassifications. To address this concern, we propose the incorporation of a learned attention mechanism, which can emphasise the contributions of the most informative node embeddings and enhance both the performance and interpretability of the model. A study by [14], demonstrated how applying an attention mechanism to highlight significant nodes can improve hierarchical representations for graph classification. This may help to avoid errors due to node heterogeneity and bias due to imbalance in the number of nodes within different resolution levels.

Recent studies have also shown how elements of both graph and vision transformer models can be combined to classify WSIs, retaining positional encodings with detailed patch level features. In 2022, [29] demonstrated how 20x patches from WSIs could be used to generate WSI representations, creating 8-adjacency subgraphs of patches that are adjacent to one another. These representations were then used as inputs for their GPT model, which combined GCN layers with Transformer layers to classifying lung cancer subtypes. Here they found their model outperformed TransMIL [27] and AttPool [14] baseline models. Following this, in 2023, [10] developed a model, which utilised two independent "Efficient Graph-based Transformer" branches, which processed both low-resolution and high-resolution patch embeddings. They used an attention matrix of self-attention to adaptively generate the adjacency matrix which is used to learn the whole slide image graph representation during training. [10], also leveraged a multi-scale fusion model, which uses cross attention to share information between the different scale branches, to exploit detail at multiple resolutions. We believe that implementing transformer layers in our own GNNs, may be another way of improving model performance, by introducing self-attention and positional encodings that may help reduce errors caused by non-informative patch embeddings.

Overall, we present a comprehensive study highlighting the superiority of graph-based methods over MIL in the novel task of classifying melanoma WSIs into immune subgroups. These findings strongly suggest that graph-based techniques could be applied to a wide variety of other problems where MIL is regarded as the gold standard. Furthermore, we showcase the clinical utility of graph-based methods in stratifying patients into prognostic immune groups, suggesting these methods are superior when modelling spatial relationships within the TME.

Acknowledgements This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]; Cancer Research UK [C588/A19167, C8216/A6129, and C588/A10721 and NIH CA83115]; and the Medical Research Council [MR/S001530/1]. This work also made use of the WSIs which were digitised by the National Pathology Imaging Co-operative, NPIC (Project no. 104687) which is supported by a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI). This work made use of 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] made use of time on Tier 2 HPC facility JADE2, funded by EPSRC (EP/T022205/1). We would

like to also thank the LMC patients for their involvement and generosity in providing data for this study.

References

1. Melanoma skin cancer statistics | Cancer Research UK, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer
2. Balch, C.M., Soong, S.J., Gershenwald, J.E., Thompson, J.F., Reintgen, D.S., Cascinelli, N., Urist, M., McMasters, K.M., Ross, M.I., Kirkwood, J.M., Atkins, M.B., Thompson, J.A., Coit, D.G., Byrd, D., Desmond, R., Zhang, Y., Liu, P.Y., Lyman, G.H., Morabito, A.: 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 (Aug 2001). https://doi.org/10.1200/JCO.2001.19.16.3622
3. Brody, S., Alon, U., Yahav, E.: How Attentive are Graph Attention Networks? (Jan 2022), http://arxiv.org/abs/2105.14491, arXiv:2105.14491 [cs] version: 3
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (Aug 2019). https://doi.org/10.1038/s41591-019-0508-1, https://www.nature.com/articles/s41591-019-0508-1., number: 8 Publisher: Nature Publishing Group
5. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. arXiv:1912.08937 [cs, q-bio] (Sep 2020), http://arxiv.org/abs/1912.08937, arXiv: 1912.08937
6. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. arXiv:2011.13971 [cs, eess] (Sep 2021), http://arxiv.org/abs/2011.13971, arXiv: 2011.13971
7. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs) (Feb 2016). https://doi.org/10.48550/arXiv.1511.07289, http://arxiv.org/abs/1511.07289, arXiv:1511.07289 [cs] version: 5
8. Curti, B.D., Faries, M.B.: Recent Advances in the Treatment of Melanoma. New England Journal of Medicine (Jun 2021). https://doi.org/10.1056/NEJMra2034861, https://www.nejm.org/doi/10.1056/NEJMra2034861, publisher: Massachusetts Medical Society
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(1-2), 31–71 (Jan 1997). https://doi.org/10.1016/S0004-3702(96)00034-3, https://linkinghub.elsevier.com/retrieve/pii/S0004370296000343
10. Ding, S., Li, J., Wang, J., Ying, S., Shi, J.: Multi-scale Efficient Graph-Transformer for Whole Slide Image Classification (May 2023). https://doi.org/10.48550/arXiv.2305.15773, http://arxiv.org/abs/2305.15773, arXiv:2305.15773 [cs]
11. Fu, Q., Chen, N., Ge, C., Li, R., Li, Z., Zeng, B., Li, C., Wang, Y., Xue, Y., Song, X., Li, H., Li, G.: Prognostic value of tumor-infiltrating lymphocytes in melanoma: a systematic review and meta-analysis. Oncoimmunology

8(7), 1593806 (Apr 2019). https://doi.org/10.1080/2162402X.2019.1593806, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527267/

12. Gershenwald, J.E., Scolyer, R.A.: Melanoma Staging: American Joint Committee on Cancer (AJCC) 8th Edition and Beyond. Annals of Surgical Oncology **25**(8), 2105–2110 (Aug 2018). https://doi.org/10.1245/s10434-018-6513-7
13. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: H^2-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 (Jun 2022). https://doi.org/10.1609/aaai.v36i1.19976, https://ojs.aaai.org/index.php/AAAI/article/view/19976, number: 1
14. Huang, J., Li, Z., Li, N., Liu, S., Li, G.: AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Networks via Attention Mechanism. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6479–6488. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00658, https://ieeexplore.ieee.org/document/9009471/
15. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based Deep Multiple Instance Learning. arXiv:1802.04712 [cs, stat] (Jun 2018), http://arxiv.org/abs/1802.04712, arXiv: 1802.04712
16. Lee, Y., Park, J.H., Oh, S., Shin, K., Sun, J., Jung, M., Lee, C., Kim, H., Chung, J.H., Moon, K.C., Kwon, S.: Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nature Biomedical Engineering (Aug 2022). https://doi.org/10.1038/s41551-022-00923-0, https://www.nature.com/articles/s41551-022-00923-0
17. Lu, M.Y., Zhao, M., Shady, M., Lipkova, J., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Deep Learning-based Computational Pathology Predicts Origins for Cancers of Unknown Primary. Nature **594**(7861), 106–110 (Jun 2021). https://doi.org/10.1038/s41586-021-03512-4, http://arxiv.org/abs/2006.13932, arXiv:2006.13932 [cs, q-bio]
18. Lu, W., Toss, M., Rakha, E., Rajpoot, N., Minhas, F.: SlideGraph+: Whole Slide Image Level Graphs to Predict HER2Status in Breast Cancer (Oct 2021), http://arxiv.org/abs/2110.06042, arXiv:2110.06042 [cs]
19. Newton-Bishop, J.A., Beswick, S., Randerson-Moor, J., Chang, Y.M., Affleck, P., Elliott, F., Chan, M., Leake, S., Karpavicius, B., Haynes, S., Kukalizch, K., Whitaker, L., Jackson, S., Gerry, E., Nolan, C., Bertram, C., Marsden, J., Elder, D.E., Barrett, J.H., Bishop, D.T.: Serum 25-Hydroxyvitamin D3 Levels Are Associated With Breslow Thickness at Presentation and Survival From Melanoma. Journal of Clinical Oncology **27**(32), 5439–5444 (Nov 2009). https://doi.org/10.1200/JCO.2009.22.1135, https://ascopubs.org/doi/10.1200/JCO.2009.22.1135, publisher: Wolters Kluwer
20. Nsengimana, J., Laye, J., Filia, A., O'Shea, S., Muralidhar, S., Poźniak, J., Droop, A., Chan, M., Walker, C., Parkinson, L., Gascoyne, J., Mell, T., Polso, M., Jewell, R., Randerson-Moor, J., Cook, G.P., Bishop, D.T., Newton-Bishop, J.: beta-Catenin–mediated immune evasion pathway frequently operates in primary cutaneous melanomas (May 2018). https://doi.org/10.1172/JCI95351, https://www.jci.org/articles/view/95351/pdf, publisher: American Society for Clinical Investigation
21. Pati, P., Jaume, G., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A.M., Scognamiglio, G., Brancati, N., Fiche, M., Dubruc, E., Riccio, D., Di Bonito, M., De Pietro, G., Botti, G., Thiran, J.P., Frucci, M., Goksel, O., Gabrani, M.: Hierarchical graph representations in digital pathology. Medical Image Anal-

ysis **75**, 102264 (Jan 2022). https://doi.org/10.1016/j.media.2021.102264, https://www.sciencedirect.com/science/article/pii/S1361841521003091

22. Poźniak, J., Nsengimana, J., Laye, J.P., O'Shea, S.J., Diaz, J.M.S., Droop, A.P., Filia, A., Harland, M., Davies, J.R., Mell, T., Randerson-Moor, J.A., Muralidhar, S., Hogan, S.A., Freiberger, S.N., Levesque, M.P., Cook, G.P., Bishop, D.T., Newton-Bishop, J.: Genetic and Environmental Determinants of Immune Response to Cutaneous Melanoma. Cancer Research **79**(10), 2684–2696 (May 2019). https://doi.org/10.1158/0008-5472.CAN-18-2864, http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-18-2864
23. Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., Lichinitser, M., Dummer, R., Grange, F., Mortier, L., Chiarion-Sileni, V., Drucis, K., Krajsova, I., Hauschild, A., Lorigan, P., Wolter, P., Long, G.V., Flaherty, K., Nathan, P., Ribas, A., Martin, A.M., Sun, P., Crist, W., Legos, J., Rubin, S.D., Little, S.M., Schadendorf, D.: Improved overall survival in melanoma with combined dabrafenib and trametinib. The New England Journal of Medicine **372**(1), 30–39 (Jan 2015). https://doi.org/10.1056/NEJMoa1412690
24. Roullier, V., Lézoray, O., Ta, V.T., Elmoataz, A.: 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). https://doi.org/10.1016/j.compmedimag.2011.02.005
25. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images (Jul 2021), http://arxiv.org/abs/2107.09405, arXiv:2107.09405 [cs, eess]
26. Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., Clozel, T., Moarii, M., Courtiol, P., Wainrib, G.: A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nature Communications **11**(1), 3877 (Aug 2020). https://doi.org/10.1038/s41467-020-17678-4, https://www.nature.com/articles/s41467-020-17678-4, number: 1 Publisher: Nature Publishing Group
27. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification (2021)
28. Sims, J., Grabsch, H.I., Magee, D.: Using Hierarchically Connected Nodes and Multiple GNN Message Passing Steps to Increase the Contextual Information in Cell-Graph Classification. In: Manfredi, L., Ahmadi, S.A., Bronstein, M., Kazi, A., Lomanto, D., Mathew, A., Magerand, L., Mullakaeva, K., Papiez, B., Taylor, R.H., Trucco, E. (eds.) Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis. pp. 99–107. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-21083-9_10
29. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification (May 2022), http://arxiv.org/abs/2205.09671, arXiv:2205.09671 [cs]