Medical Image Analysis 93 (2024) 103097

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Immune subtyping of melanoma whole slide images using multiple instance learning

Lucy Godson [a,*], Navid Alemi [a], Jérémie Nsengimana [e], Graham P. Cook [c], Emily L. Clarke [b,d], Darren Treanor [b,d,g,h], D. Timothy Bishop [c], Julia Newton-Bishop [d], Ali Gooya [f], Derek Magee [a]

[a] School of Computing, University of Leeds, Woodhouse, Leeds, LS2 9JT, United Kingdom
[b] Department of Histopathology, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom
[c] Leeds Institute of Medical Research, University of Leeds School of Medicine, St. James's University Hospital, Leeds, United Kingdom
[d] Division of Pathology and Data Analytics, Leeds Institute of Cancer and Pathology, University of Leeds, Beckett Street, Leeds, LS9 7TF, United Kingdom
[e] Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom
[f] School of Computing, University of Glasgow, Glasgow, G12 8QQ, United Kingdom
[g] Department of Clinical Pathology and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
[h] Centre for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Multiple instance learning
Self-supervised learning
Computational pathology
Melanoma
Histopathology
Image classification

ABSTRACT

Determining early-stage prognostic markers and stratifying patients for effective treatment are two key challenges for improving outcomes for melanoma patients. Previous studies have used tumour transcriptome data to stratify patients into immune subgroups, which were associated with differential melanoma specific survival and potential predictive biomarkers. However, acquiring transcriptome data is a time-consuming and costly process. Moreover, it is not routinely used in the current clinical workflow. Here, we attempt to overcome this by developing deep learning models to classify gigapixel haematoxylin and eosin (H&E) stained pathology slides, which are well established in clinical workflows, into these immune subgroups. We systematically assess six different multiple instance learning (MIL) frameworks, using five different image resolutions and three different feature extraction methods. We show that pathology-specific self-supervised models using 10x resolution patches generate superior representations for the classification of immune subtypes. In addition, 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. Furthermore, we show that these models are able to stratify patients into 'high' and 'low immune' subgroups with significantly different melanoma specific survival outcomes (log rank test, $P < 0.005$). We anticipate that MIL methods will allow us to find new biomarkers of high importance, act as a tool for clinicians to infer the immune landscape of tumours and stratify patients, without needing to carry out additional expensive genetic tests.

1. Introduction

Immunotherapy has revolutionised the treatment of melanoma patients (Robert et al., 2015; Ugurel et al., 2016; Wolchok et al., 2021), however, a decade on from the first immune checkpoint inhibitor being approved for treatment of advanced melanoma (Huang and Zappasodi, 2022), there are still patients who do not derive long-lasting benefits or respond to immunotherapeutic treatment. Therefore, determining early-stage prognostic biomarkers, understanding disease progression and stratifying patients accordingly for effective treatment of melanoma, have all emerged as increasingly important challenges to address.

In 2019, Poźniak et al. (2019), used tumour transcriptomic data to stratify melanoma patients into immune-related subsets, based on

Abbreviations: AUC, Area Under the receiver operating characteristic Curve; CLAM, Clustering-constrained Attention Multiple instance learning; CNN, Convolutional Neural Network; DSMIL, Dual Stream Multiple-instance learning; H&E, Haematoxylin and Eosin; LMC, Leeds Melanoma Cohort; MIL, Multiple Instance Learning; MSS, Melanoma specific survival; ResNet, Residual Network; SimCLR, Simple contrastive learning for learning visual features; SSL, self-supervised learning; TCGA, The Cancer Genome Atlas; TILs, tumour infiltrating lymphocytes; TransMIL, Transformer based Multiple-instance learning; WSI, Whole Slide Image

* Corresponding author.
E-mail address: sclg@leeds.ac.uk (L. Godson).

https://doi.org/10.1016/j.media.2024.103097
Received 16 December 2022; Received in revised form 15 January 2024; Accepted 25 January 2024
Available online 1 February 2024

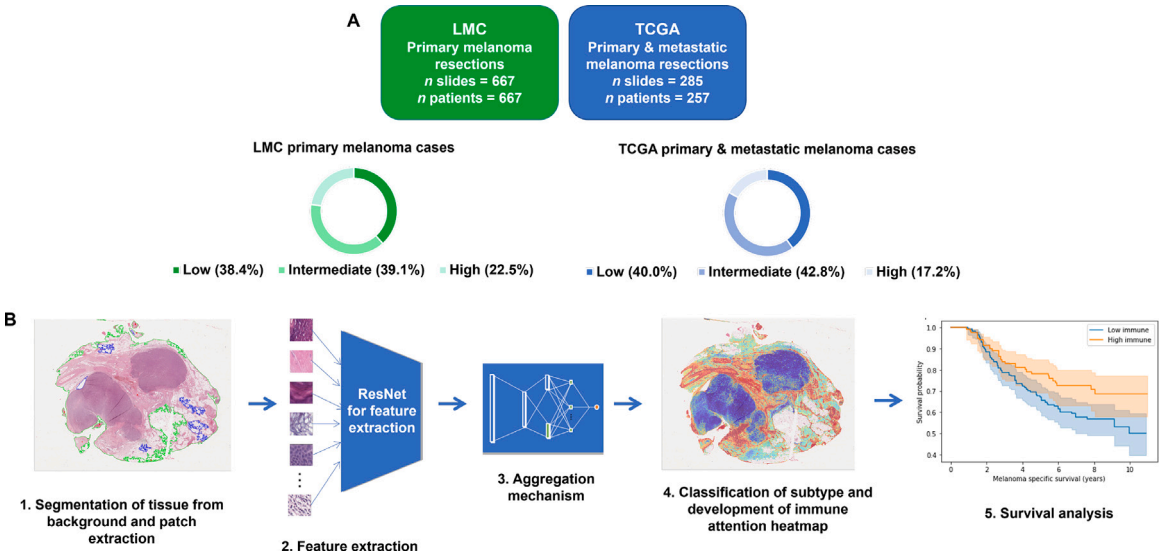

Fig. 1. Datasets and experimental framework. A. Patient cohorts: two independent datasets that were used for training and testing of models and the distribution of immune subtype labels within these datasets. B. The images were processed, then three different feature extraction methods were used to generate 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.

the inferred abundance of 27 immune cells as described by Angelova et al. (2015). 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 (Acharya et al., 2020; Kjeldsen et al., 2021; Rotte, 2019). They were also able to replicate these findings, in an independent transcriptomic dataset of melanoma tumours, from the cancer genome atlas (TCGA).

Currently, the identification of prognostic molecular biomarkers for cancer 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 (The Royal College of Pathologists, 2018). The integration of deep learning and image analysis presents a cost-effective opportunity to automate the detection of prognostic biomarkers within clinical workflows (Kacew et al., 2021). 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. Several studies have successfully demonstrated the use of deep learning techniques with H&E-stained slides to predict genomic subtypes and the expression of specific genes. For example, deep learning pipelines that utilise WSIs have been implemented to classify and predict mutations for lung (Coudray et al., 2018; Yu et al., 2020) and breast cancer subtypes (Couture et al., 2018; Rawat et al., 2020; Qu et al., 2021; Lu et al., 2021b; Campanella et al., 2022), predict tumor mutational burden in bladder cancer patients (Xu et al., 2019) and detect microsatellite instability in colorectal, gastric, endometrial and ovarian cancers (Hildebrand et al., 2021; Alam et al., 2022; Park et al., 2022; Guo et al., 2023). Additionally, Sirinukunwattana et al. (2021) showcased how deep learning applied to WSIs could be used to predict image-based consensus molecular subtypes in colorectal cancer, exhibiting enhanced prognostic capabilities compared to current grading systems.

Despite these advancements, there are a limited number of studies focusing on classifying melanoma tumors based on image-based molecular subtypes beyond the examination of mutational burden and individual mutations (Kim et al., 2020, 2022; Noorbakhsh et al., 2020). In this study our objective is to classify melanoma H&E slides into immune subtypes (Poźniak et al., 2019), that have added prognostic ability compared to the current melanoma staging system, using deep learning.

1.1. Related works

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 that inherit the bag or slide-level label (Dietterich et al., 1997), have been implemented to effectively analyse WSIs (Campanella et al., 2018, 2019; Lu et al., 2020; Courtiol et al., 2019; Lu et al., 2021a; Schirris et al., 2021; Valieris et al., 2020). 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 (Ilse et al., 2018), 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, in 2018 (Ilse et al., 2018) added a trainable attention-based pooling operator and in 2020 (Lu et al., 2020) further added to this work by implementing a clustering mechanism to further separate instances for cancer subtype classification. Li et al. (2021a) 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), 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 (Li et al., 2021b; Zhang et al., 2022; Chen et al., 2022; Tourniaire et al., 2023; Lin et al., 2023).

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 (Awan et al., 2018; Courtiol et al., 2019; Kather et al., 2019; Lu et al., 2021a; Ghaffari Laleh et al., 2022), which is comprised of millions of annotated natural images, from thousands of categories, such as food, locations, animals and people (Deng et al., 2009). These CNN feature extractors, pretrained using ImageNet (Deng et al., 2009) 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 (Awan et al., 2018; Courtiol et al., 2019; Kather et al., 2019; Lu et al., 2021a; Ghaffari Laleh et al., 2022), the differences between the image domains, can lead to a reduction in accuracy for certain pathology specific tasks. For example, Yu et al. (2020), 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) 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 (Coudray et al., 2018; Campanella et al., 2019; Fu et al., 2020), 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 unlabelled patches from WSIs. Abbasi-Sureshjani et al. (2021) compared the performance of models using features from a CNN pretrained using a Bootstrap your own latent (BYOL) (Grill et al., 2020) 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) 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), demonstrated how using a histopathology-specific feature extractor, pretrained using simple contrastive learning for learning visual features (SimCLR Chen et al., 2020 [an SSL approach]), improved classification of MSI and homologous recombination deficiency in breast and colorectal cancer datasets.

1.2. 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 Poźniak et al. (2019) (see Fig. 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.

2. Methods

2.1. Segmentation and feature extraction

The H&E tissue in the all WSIs was segmented from the background using the protocol designed by Lu et al. (2020). This required 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 pixel × 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 (He et al., 2016), pretrained using supervised learning on the ImageNet dataset (Deng et al., 2009), a ResNet18 CNN, which was also pretrained using supervised learning on ImageNet and a ResNet18 pretrained using SSL on histopathology images (Ciga et al., 2021). 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 (Liu et al., 2018) 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 (Ciga et al., 2021) from https://github.com/ozanciga/self-supervised-histopathology (last accessed: November 2022), which has been trained using 57 multi-organ, multi-resolution (10x, 20x, 40x and 100x) histopathology datasets.

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.

2.2.1. Max pooling MIL

The baseline methodology we used was a max pooling MIL method, referred to as a MIL model:

$$\forall_{m=1,\ldots,M} : z_m = \max_{k=1,\ldots,K}(\mathbf{h}_{km}),$$

where z_m is the slide-level representation for the *mth* slide from the M WSIs. $\mathbf{h}_{km}$ is the *kth* patch instance embedding from the *mth* slide and is the low-dimensional patch representation with the highest probability for a certain class, that provides the overall WSI label.

2.2.2. Attention MIL

Both of the attention-based MIL pooling functions with and without gating (Ilse et al., 2018), 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 *kth* 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. The weights for the attention mechanism without gating are formulated by:

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

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.

2.2.3. Gated attention MIL

Moreover, we implemented a gated attention mechanism, developed by Dauphin et al. (2017) to introduce sigmoid non-linearity for learning more complex relationships:

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

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.

2.2.4. Clustering-constrained-attention MIL

Furthermore, we evaluated clustering-constrained-attention MIL (CLAM) developed by Lu et al. (2020). 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 (Lu et al., 2020). 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 (Lapin et al., 2016).

2.2.5. Transformer based MIL

Additionally, we implemented transformer based MIL (TransMIL), which is based on a correlated framework by Shao et al. (2021). Self attention and multi-head attention are leveraged to incorporate spatial information into slide-level classifications.

2.2.6. Dual-stream MIL

The final MIL model we implemented was Dual stream MIL (DSMIL) (Li et al., 2021a), which is another variation of MIL framework, which models bag representations and instance-level representations 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 (Li et al., 2021a) 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 code used to implement the CLAM, Attention, Gated-attention and max-MIL models was adapted from the following code repository: https://github.com/mahmoodlab/CLAMGitHubrepository (last accessed: November 2022). In addition, the code used to train and assess 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).

2.2.7. 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, Adam optimiser with a learning rate of $2 \times 10e^{-4}$, and weight decay of $1\times 10e^{-5}$ were applied. For the TransMIL model, following the author's implementation (Shao et al., 2021), we utilised a Lookahead optimiser (Zhang et al., 2019) with rectified Adam (RAdam) optimiser (Liu et al., 2021) with a 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 (Xiong et al., 2021) 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 Poźniak et al. (2019) 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 1 & Table 2). 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.

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.

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 (Kaplan and Meier, 1958) 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.

A univariate Cox proportional-hazards model (Cox, 1972; Breslow, 1975) 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) staging version 7 (Balch et al., 2009), age, sex and melanoma site (limbs vs. rest) as possible confounding factors. The multivariate Cox proportional-hazards model can be represented as:

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

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.

Table 1
LMC dataset splits, showing the number of WSIs labelled with the three different immune subgroups found by Poźniak et al. (2019), 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 2
TCGA dataset splits, showing the number of digitised WSIs labelled with the three different immune subgroups found by Poźniak et al. (2019), 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

3. Datasets

The primary dataset we used for our work was from the Leeds Melanoma Cohort (LMC) (Newton-Bishop et al., 2015, 2009). This is a population ascertained cohort, including 667 digitised WSIs of melanoma primaries (667 slides, 667 patients). The labels for the images were delineated by clustering the transcriptomes, based on the inferred abundance of 27 immune cell types. 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 Nsengimana et al. (2018) (see Table 1).

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 (Heath et al., 2021). The WSIs (285 slides, 257 patients) are from both primary (177 slides, 176 patients) and metastatic melanomas (107 slides, 81 patients) from regional cutaneous or subcutaneous tissue (including satellite and in-transit metastases), regional lymph nodes and distant metastases. All lesions included in our comparison against the LMC dataset were from metastatic patients, as metadata from the primary melanoma cases suggested they had worse survival compared to metastatic patients, with patients being biased towards late diagnosis. The immune labels for the TCGA images were formulated, using RNAseq expression data from these cases and assigning them to the subtype cluster centroid which had the strongest Spearman's correlation (Poźniak et al., 2019). The WSIs and corresponding RNAseq data are available from the NIH genomic data commons (https://portal.gdc.cancer.gov [date last accessed: November 2022]). We selected H&E stained FFPE diagnostic slides (eliminating frozen sections) and images which were scanned in at 0.25 micrometers-per-pixel (m.p.p.), to see if our results were replicable in a second dataset (see Table 2).

4. Results

4.1. Predicting high, intermediate and low immune subtypes

Initial experiments were carried out by training the models using the three subgroups determined by Poźniak et al. (2019). 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

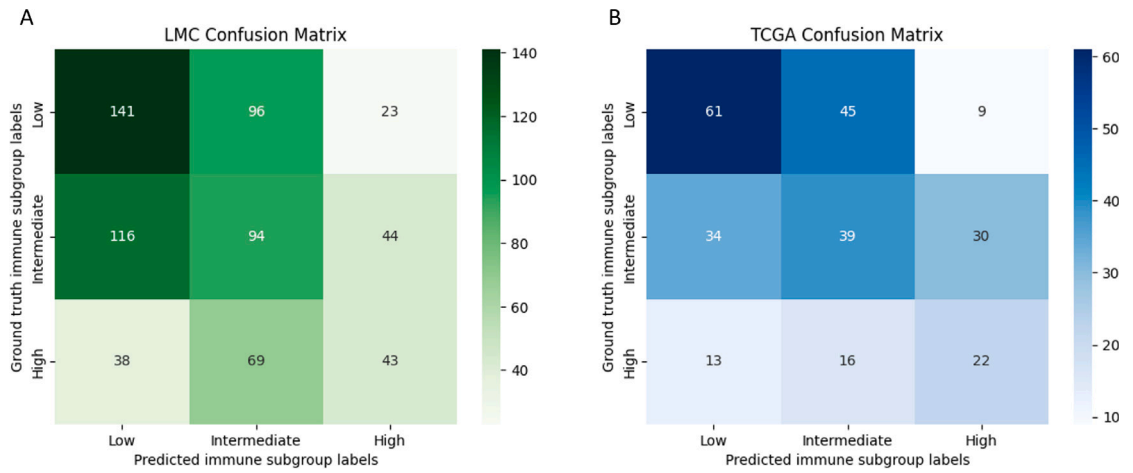

Fig. 2. Confusion matrices for the model predictions of the 'low', 'intermediate' and 'high' cases on the 10-fold test sets. A. Predictions from the best performing model for classifying the three subtypes using the LMC dataset. B. Predictions from the best performing model for classifying the three subtypes using the TCGA dataset. LMC, Leeds Melanoma Cohort; TCGA, The Cancer Genome Atlas.

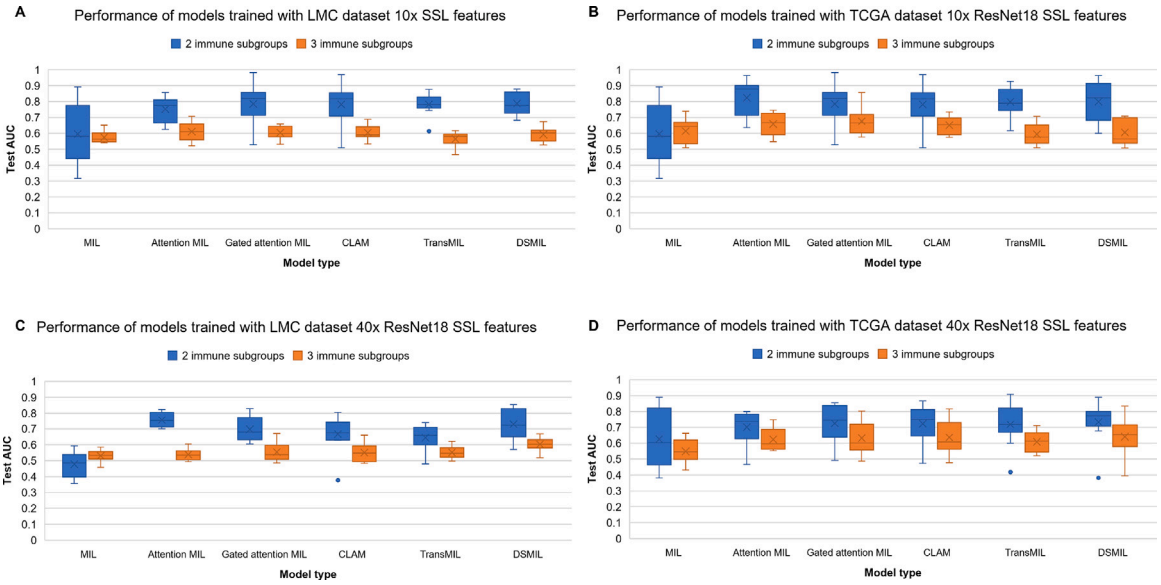

Fig. 3. Performance comparison of the six different types of MIL models when classifying the melanoma immune subtypes in the two datasets. Box-plots showing the 10-fold test AUCs for each model type, when carrying out the 2 immune subtyping task ('high immune' and 'low immune' [in blue]) and three immune subtyping task ('high', 'intermediate' and 'low immune' [in orange]). The boxes show the quartile values and the whiskers extend to data points within 1.5x 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.

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) (Figs. 2A & 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) (Figs. 2B & 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. 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 (Fig. 2. Due to these errors, we simplified the task to a binary problem of classifying the 'high immune' against 'low immune' subtypes (Fig. 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.

4.2. Comparison of MIL models

We evaluated the performance of six different MIL models on the classification tasks (Fig. 3). In general, the results indicate that the

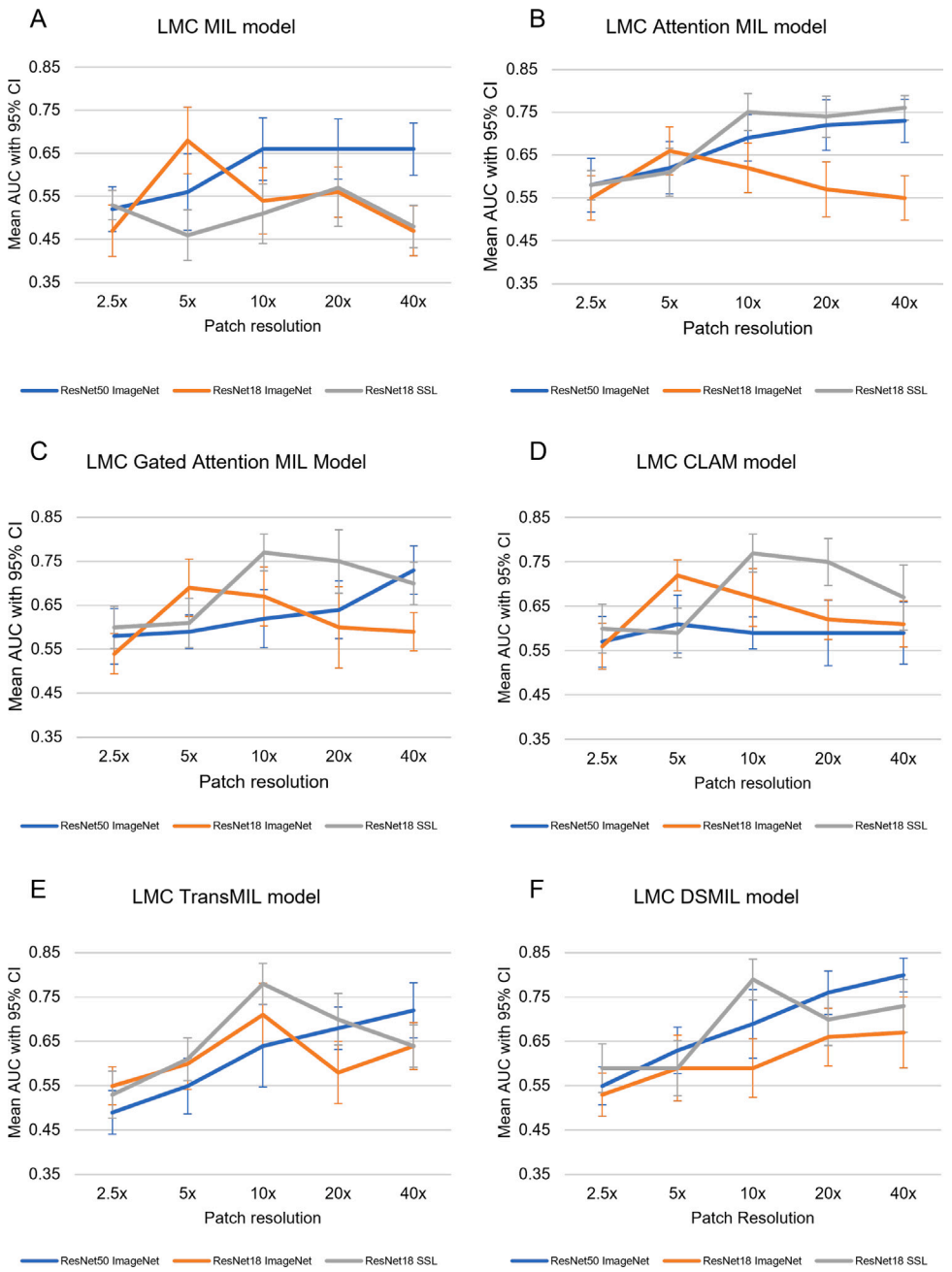

Fig. 4. Model performance when altering the feature extraction method and resolution of input patches for models trained with the LMC dataset. Mean test AUC with 95% CI of MIL (A), attention MIL (B), gated attention MIL (C), CLAM (D) and TransMIL (E) and DSMIL (F) models for classifying 'high' and 'low 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.

standard max pooling MIL model gives the poorest performance for our subtyping tasks (Fig. 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 (Figs. 4 and 5), as well as substantial interquartile ranges (Fig. 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% 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 (Figs. 4A & 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 comparison to the other MIL methods presented.

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. 4F). However, we found that the TransMIL and DSMIL models, trained and tested with 10x ResNet18 ImageNet SSL features from the LMC had similarly high performance,

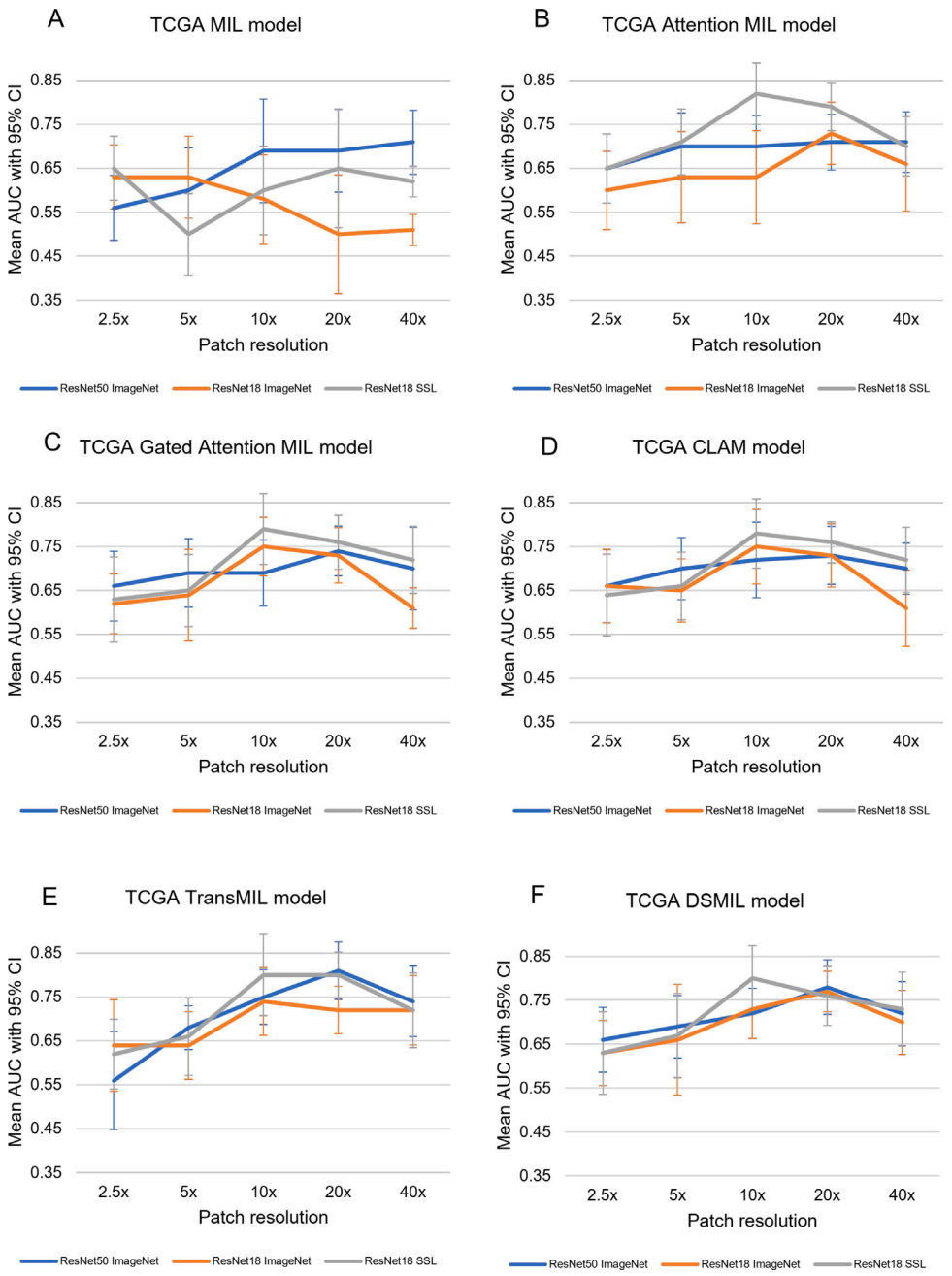

Fig. 5. Model performance when altering the feature extraction method and resolution of input patches for models trained with the TCGA dataset. Mean test AUC with 95% CI of MIL (A), attention MIL (B), gated attention MIL (C), CLAM (D), TransMIL (E) and DSMIL (F) models for classifying 'high' and 'low immune' subtypes within the TCGA metastatic 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 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. 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. 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. 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 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.

4.3. Comparison of feature extraction strategies

When deploying MIL models, feature extraction is a key component within the pipeline. Often off-the-shelf CNNs pre-trained with

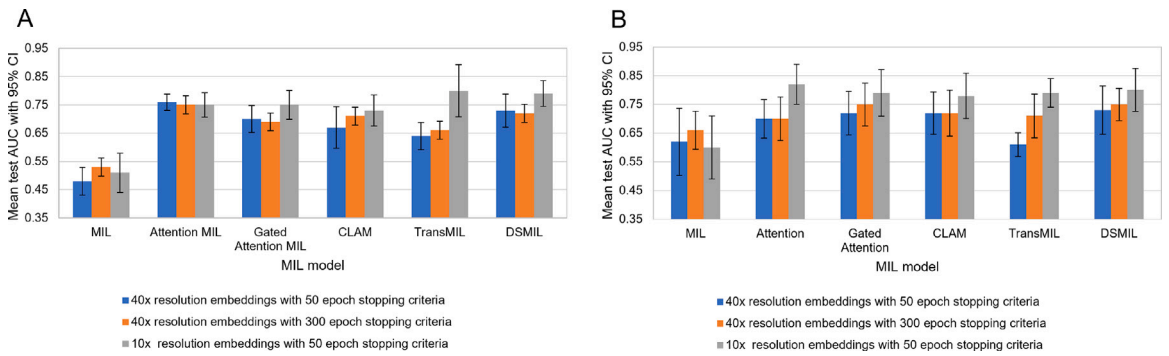

Fig. 6. Mean 10-fold test AUC performance with 95% CI of MIL models classifying 'high' and 'low immune' subtypes with different epoch stopping lengths during training. A. Results from models trained with ResNet18 SSL feature extractor embeddings from the LMC dataset. B. Results from models trained with 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.

the ImageNet dataset are utilised for this step, therefore we wanted to examine how using different feature extraction strategies 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 (Figs. 4B-E and 5B-D). In addition, the SSL ResNet18 pre-trained with pathology specific images improved performance for the CLAM (Fig. 4D) and attention MIL (Fig. 4B) models trained and tested with 40x resolution LMC WSIs and the CLAM (Fig. 5C) and gated attention MIL (Fig. 5D) models trained and tested with the TCGA dataset. In addition, there was a competing trend when using ResNet50 ImageNet features, with five of the MIL models (excluding the CLAM model) trained with the LMC dataset having increased performance as the resolution of input patches was 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. 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 (Figs. 4A-D & 5A).

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 (Figs. 4 and 5). Furthermore, as mentioned previously, different feature extraction methods, appeared to perform the better at different resolutions. This is exemplified in Fig. 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. 4A-D). Whereas the optimum resolution for the TCGA dataset WSIs is more variable for the ImageNet pretrained ResNet18 (Fig. 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. 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, 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 with LMC data, with all models performing worse when increasing the number of epochs a model was trained for (Fig. 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.

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. 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. 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 3). 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.

4.5. Immune attention heatmaps for interpretability

Furthermore, we developed immune attention heatmaps, for three of the MIL models with highest mean AUC performance, showing where

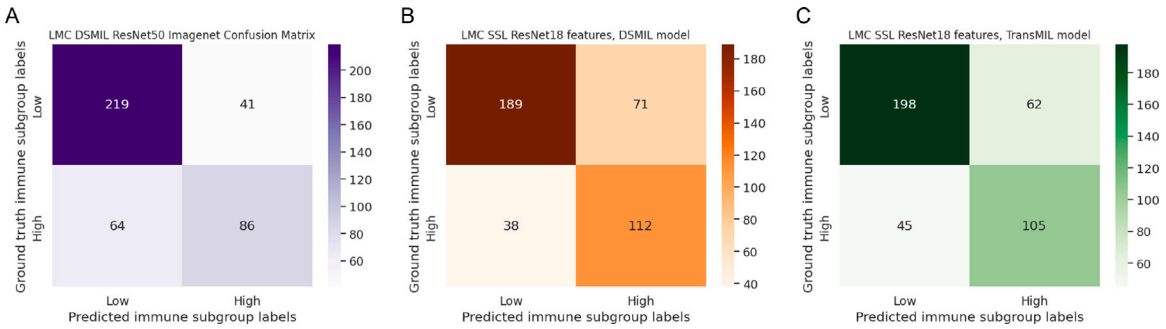

Fig. 7. Confusion matrices for the model predictions of the 'low', and 'high' cases on the 10-fold test sets. A. Predictions from the LMC DSMIL model trained with 40x resolution ResNet50 ImageNet features. B. Predictions from the LMC DSMIL model trained with 10x resolution ResNet18 SSL features. C. Predictions 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 3
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)

the patches with the highest and lowest attention weights were located. This was to gain a better understanding of model performance. In Fig. 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 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. 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 (Figs. 9 and 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. 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 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 examined cases that were misclassified to further our understanding of the potential limitations of the MIL models and the experimental design of this study. In Fig. 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 (Clarke and Treanor, 2017). 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 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. 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 (Nsengimana et al., 2018).

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. 12) 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. 12B), 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 'high immune' subgroup as predicted by all models apart from the attention MIL model, had a significantly lower hazard of melanoma death compared to the 'low immune' patient subgroups (Table 4). Moreover when adjusting for clinical predictors (age, sex, tumour site, AJCC

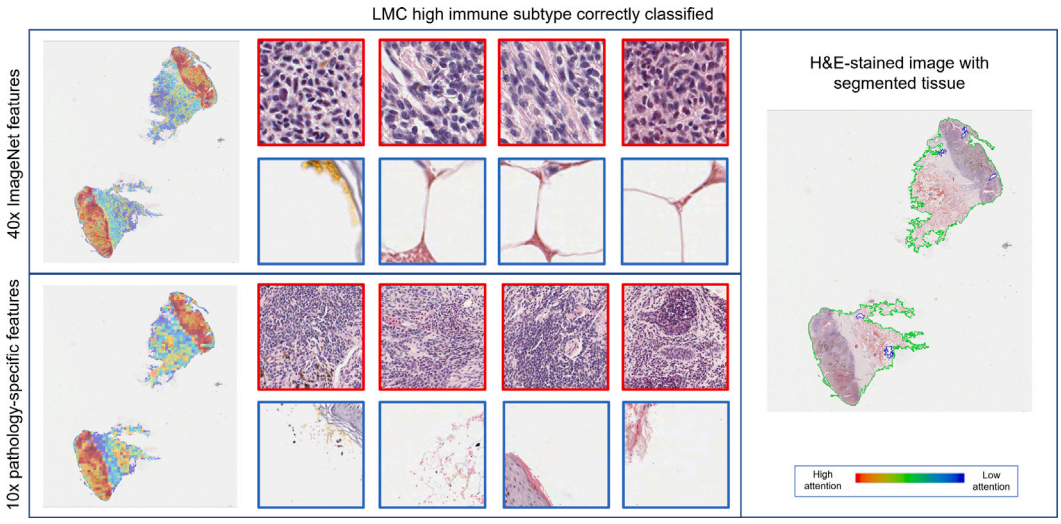

Fig. 8. Comparison of correctly classified LMC 'high immune' tumours using different input features. Immune attention heatmaps with the original H&E stained WSI and four patches that contributed the most (red) to the 'high immune' subtype prediction and four patches that contributed the least (blue). 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.

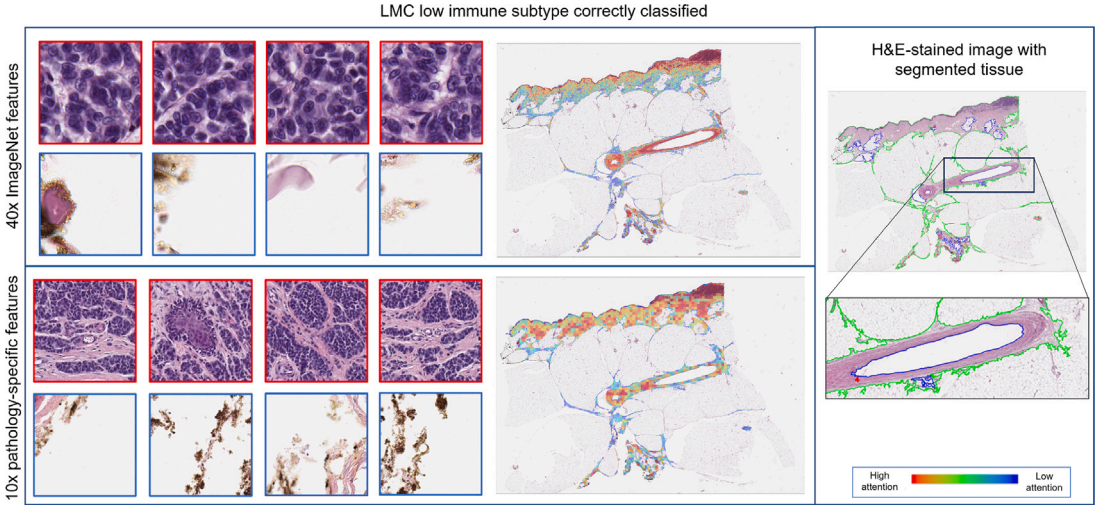

Fig. 9. Comparison of correctly classified LMC 'low immune' tumours using different input features. Immune attention heatmaps with the original H&E stained WSI and four patches that contributed the most (red) to the 'low immune' subtype prediction and four patches that contributed the least (blue). We 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.

stage), by using a multivariate model, we also observed a significantly lower hazard of melanoma death for patients in the 'high immune' subgroup compared to 'low immune' subgroup (Table 4), when using 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, suggesting that image features could have a stronger association with survival, than the immune subgroup features derived from the transcriptomic data.

5. Discussion

5.1. Overview of findings

Recent studies (Poźniak et al., 2019; Nsengimana et al., 2018) have shown that melanoma patients can be stratified into subgroups, with added prognostic value compared to the current melanoma staging system (Gershenwald and Scolyer, 2018). 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

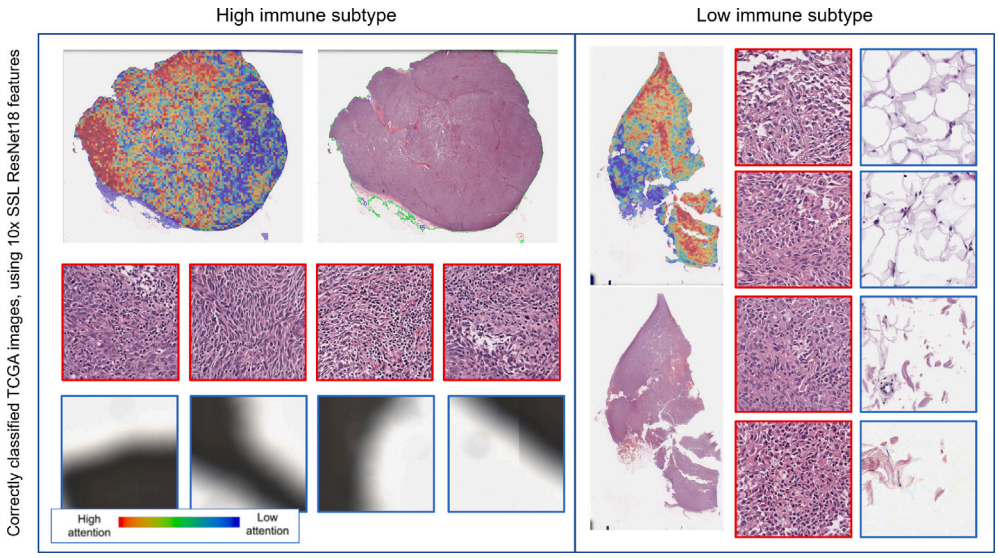

Fig. 10. Correctly classified 'high' and 'low immune' TCGA tumours. Immune attention heatmaps with the original H&E stained WSI and four patches (red) that contributed the most and least (blue) to the subtype predictions. Input features were generated using 10x feature embeddings from the SSL ResNet18 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.

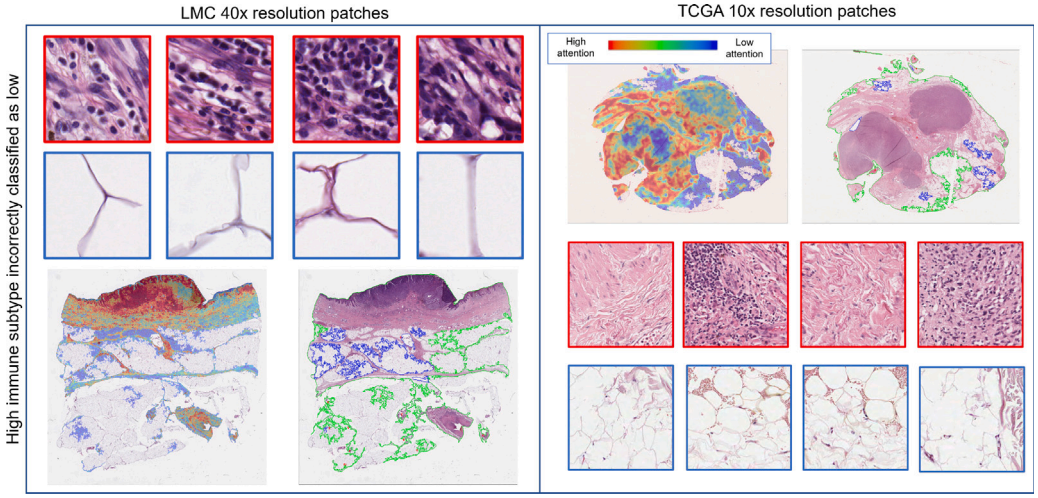

Fig. 11. Incorrectly classified 'high immune' LMC and TCGA tumours. Immune attention heatmaps with the original H&E stained WSI and four patches (red) that contributed the most and least (blue) to the incorrect 'low immune' subtype prediction. For the LMC WSI, the input features were generated using 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.

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. 12 & Table 4). We also show feature inputs are important for improving model performance, in terms of both patch resolution and feature extraction methodology (Figs. 4 and 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.

5.2. Challenges using transcriptomic ground truth labels

We also 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), 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. 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 Figs. 8 and 10, the 'high immune' 10x attention patches contain both tumour cells and

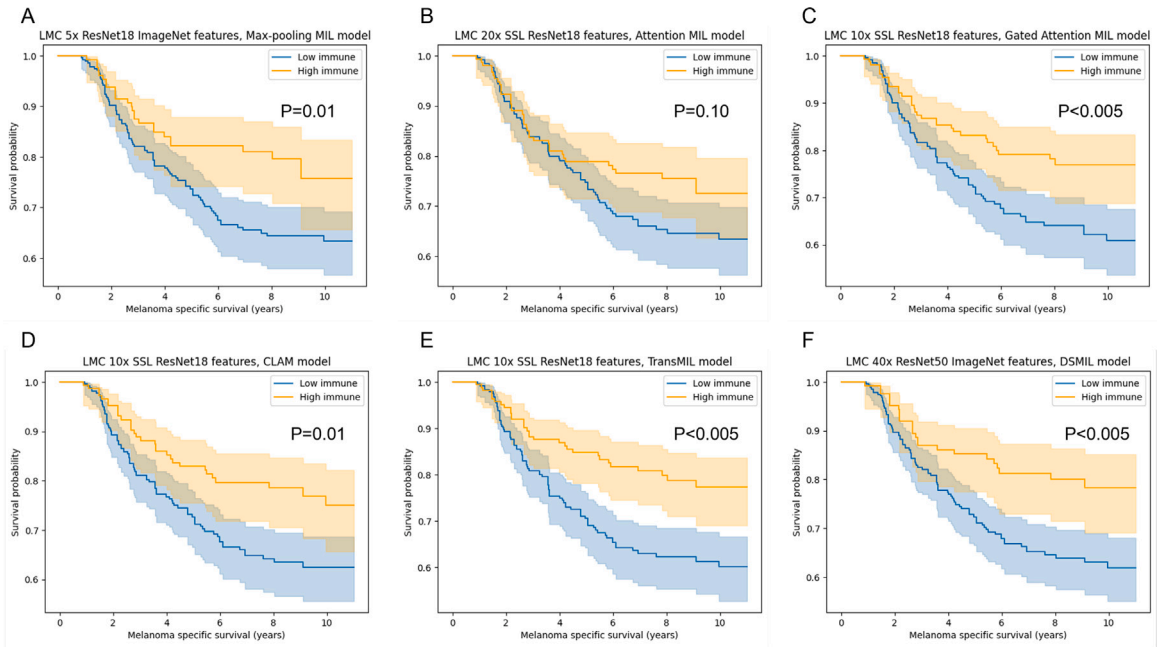

Fig. 12. Prognostic associations of the 'high' and 'low immune' subtypes from the LMC dataset (N = 230), using labels generated from (A) max pooling MIL model trained with ResNet18 ImageNet embeddings from 5x patches, (B) attention MIL model trained with SSL ResNet18 feature embeddings from 20x patches, (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.

Table 4
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. The first row gives 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**

* Indicates P values < 0.05.
** Indicates P values < 0.005.

immune cells. This indicates that the model has not solely focused on lymphocyte-rich regions outside of the tumour area, but may have utilised regions that contain tumor cells for context.

5.3. MIL models attend to regions with prognostic value

In 1989, Clark et al. (1989) 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) 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. 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. Nevertheless, in 2023, a study by Chatziioannou et al. (2023), 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 (Figs. 8 and 10), reinforcing this evidence that TILs have a prognostic signal.

In addition, in Table 4 and Fig. 12 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 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) suggests that using MIL models with all the patches from a WSI, may improve patient stratification into groups based on survival, due to reduced sensitivity to noise compared to the spatially biased ground truth labels.

5.4. The importance of feature inputs and model selection

A previous rigorous benchmarking study by Ghaffari Laleh et al. (2022) examined the difference in performance of different MIL models for different cancer subtyping tasks. However, our study examines the often overlooked effects of how input patch resolution and feature extraction method can influence performance for 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 (Couture, 2022). 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) 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 for MIL models trained using LMC ResNet50 ImageNet embeddings, which showed increased performance as resolution increased (as shown in Fig. 4). These findings also suggest that the modified ResNet50 feature extractor pretrained with ImageNet, can excel at capturing generic 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 (Figs. 8 and 9), where the 40x DSMIL model is shown attending to prognostic areas within the tumour. Yet, we also saw in Fig. 9 that the model learned high attention scores for uninformative regions, such as a blood vessels, suggesting this model is might be overfitting to noise caused by heavy staining or other artefacts in the image (as seen in Fig. 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. 11), or making classification errors due to non-salient features (Fig. 9). Additionally, when examining prognostic ability of the models, we also saw that 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). 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 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 (Saillard et al., 2021; Abbasi-Sureshjani et al., 2021; Yu et al., 2020; Noorbakhsh et al., 2020) 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 (Ghaffari Laleh et al., 2022). 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.

5.5. Limitations and future directions

Deciphering the 'intermediate immune' subgroup remains a challenge, due to the tumour heterogeneity and complexity within this group. To tackle this problem we may need to look at further dividing this subgroup, as a previous study by Nsengimana et al. (2018) found two distinct subgroups which overlap with this 'intermediate' group, or use techniques to learn more discriminant representations of the images, such as deep Fisher Discriminant analysis (Díaz-Vico and Dorronsoro, 2020). 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, it could be more useful to look at it as a regression-based problem as (Nahhas et al., 2023) reported when predicting 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. 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. (2021a), 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 (Li et al., 2020; Chen et al., 2022). 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) 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 ($N = 189$) (Poźniak et al., 2019), or due to the regions outside 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.

6. Conclusion

To the best of our knowledge our work is one of the first extensive studies that attempts to subtype melanoma histopathology images based on immune genetic signatures from transcriptomic data. 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.

CRediT authorship contribution statement

Lucy Godson: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. **Navid Alemi:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Jérémie Nsengimana:** Conceptualization, Data curation, Funding acquisition, Supervision, Writing – review & editing. **Graham P. Cook:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Emily L. Clarke:** Data curation, Formal analysis, Funding acquisition, Writing – review & editing. **Darren Treanor:** Data curation, Funding acquisition, Resources. **D. Timothy Bishop:** Data curation, Funding acquisition. **Julia Newton-Bishop:** Conceptualization, Data curation, Funding acquisition, Resources. **Ali Gooya:** Conceptualization, Supervision. **Derek Magee:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom [EP/S024336/1]; Cancer Research UK [C588/A19167, C8216/A6129, and C588/A10721 and NIH CA83115]; and the Medical Research Council, United Kingdom [MR/S001530/1].

This work also made use of the LMC WSIs which were digitised by the National Pathology Imaging Co-operative, NPIC (Project no. 104687) which is supported by a £50 m 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).

Furthermore 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]. The Centre is co-ordinated by the Universities of Durham, Manchester and York. We would like to thank 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.

We would like to also thank the Leeds Melanoma Cohort patients for their involvement and generosity in providing data for this study. We would also like to thank the patients from the TCGA, as the results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103097.

References

Abbasi-Sureshjani, S., Yüce, A.l., Schönenberger, S., Skujevskis, M., Schalles, U., Gaire, F., Korski, K., 2021. Molecular subtype prediction for breast cancer using H&E specialized backbone. In: Proceedings of the MICCAI Workshop on Computational Pathology. PMLR, (ISSN: 2640-3498) pp. 1–9, URL: https://proceedings.mlr.press/v156/abbasi-sureshjani21a.html.

Acharya, N., Sabatos-Peyton, C., Anderson, A.C., 2020. Tim-3 finds its place in the cancer immunotherapy landscape. J. Immunother. Cancer 8 (1), e000911. http://dx.doi.org/10.1136/jitc-2020-000911, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326247/.

Alam, M.R., Abdul-Ghafar, J., Yim, K., Thakur, N., Lee, S.H., Jang, H.-J., Jung, C.K., Chong, Y., 2022. Recent applications of artificial intelligence from histopathologic image-based prediction of microsatellite instability in solid cancers: A systematic review. Cancers 14 (11), 2590. http://dx.doi.org/10.3390/cancers14112590.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M.L., Snajder, R., Krogsdam, A.M., Waldner, M.J., Bindea, G., Mlecnik, B., Galon, J., Trajanoski, Z., 2015. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol. 16 (1), 64. http://dx.doi.org/10.1186/s13059-015-0620-6.

Awan, R., Koohbanani, N.A., Shaban, M., Lisowska, A., Rajpoot, N., 2018. Context-aware learning using transferable features for classification of breast cancer histology images. URL: http://arxiv.org/abs/1803.00386, arXiv:1803.00386 [cs].

Balch, C.M., Gershenwald, J.E., Soong, S.-j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., Eggermont, A.M., Flaherty, K.T., Gimotty, P.A., Kirkwood, J.M., McMasters, K.M., Mihm, M.C., Morton, D.L., Ross, M.I., Sober, A.J., Sondak, V.K., 2009. Final version of 2009 AJCC melanoma staging and classification. J. Clin. Oncol. 27 (36), 6199–6206. http://dx.doi.org/10.1200/JCO.2009.23.4799, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2793035/.

Breslow, N.E., 1975. Analysis of survival data under the proportional hazards model. Int. Statist. Rev. / Rev. Int. Stat. 43 (1), 45–57. http://dx.doi.org/10.2307/1402659, URL: https://www.jstor.org/stable/1402659, Publisher: [Wiley, International Statistical Institute (ISI)].

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., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309. http://dx.doi.org/10.1038/s41591-019-0508-1, URL: https://www.nature.com/articles/s41591-019-0508-1. Number: 8 Publisher: Nature Publishing Group.

Campanella, G., Ho, D., Häggström, I., Becker, A.S., Chang, J., Vanderbilt, C., Fuchs, T.J., 2022. H&E-based computational biomarker enables universal EGFR screening for lung adenocarcinoma. http://dx.doi.org/10.48550/arXiv.2206.10573, URL: http://arxiv.org/abs/2206.10573, arXiv:2206.10573 [cs, q-bio].

Campanella, G., Silva, V.W.K., Fuchs, T.J., 2018. Terabyte-scale deep multiple instance learning for classification and localization in pathology. arXiv:1805.06983 [cs], URL: http://arxiv.org/abs/1805.06983, arXiv:1805.06983.

Chatziioannou, E., Roßner, J., Aung, T.N., Rimm, D.L., Niessner, H., Keim, U., Serna-Higuita, L.M., Bonzheim, I., Cuellar, L.K., Westphal, D., Steininger, J., Meier, F., Pop, O.T., Forchhammer, S., Flatz, L., Eigentler, T., Garbe, C., Röcken, M., Amaral, T., Sinnberg, T., 2023. 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, http://dx.doi.org/10.1016/j.ebiom.2023.104644, URL: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00209-8/fulltext#%20, Publisher: Elsevier.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. arXiv:2206.02647.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. URL: http://arxiv.org/abs/2002.05709, arXiv:2002.05709 [cs, stat].

Ciga, O., Xu, T., Martel, A.L., 2021. Self supervised contrastive learning for digital histopathology. arXiv:2011.13971 [cs, eess], URL: http://arxiv.org/abs/2011.13971, arXiv:2011.13971.

Clark, W.H., Elder, D.E., Guerry, D., Braitman, L.E., Trock, B.J., Schultz, D., Synnestvedt, M., Halpern, A.C., 1989. Model predicting survival in stage I melanoma based on tumor progression. J. Natl. Cancer Inst. 81 (24), 1893–1904. http://dx.doi.org/10.1093/jnci/81.24.1893.

Clarke, E.L., Treanor, D., 2017. Colour in digital pathology: a review. Histopathology 70 (2), 153–163.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. 24 (10), 1559–1567. http://dx.doi.org/10.1038/s41591-018-0177-5, URL: https://www.nature.com/articles/s41591-018-0177-5, Number: 10 Publisher: Nature Publishing Group.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., Girard, N., Elemento, O., Nicholson, A.G., Blay, J.-Y., Galateau-Sallé, F., Wainrib, G., Clozel, T., 2019. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. 25 (10), 1519–1525. http://dx.doi.org/10.1038/s41591-019-0583-3, URL: https://www.nature.com/articles/s41591-019-0583-3, Number: 10 Publisher: Nature Publishing Group.

Couture, H.D., 2022. Deep learning-based prediction of molecular tumor biomarkers from H&E: A practical review. J. Pers. Med. 12 (12), 2022. http://dx.doi.org/10.3390/jpm12122022, URL: https://www.mdpi.com/2075-4426/12/12/2022, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J.S., Perou, C.M., Troester, M.A., Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 4, 30. http://dx.doi.org/10.1038/s41523-018-0079-1.

Cox, D.R., 1972. Regression models and life-tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2), 187–202. http://dx.doi.org/10.1111/j.2517-6161.1972.tb00899.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks. arXiv:1612.08083 [cs], URL: http://arxiv.org/abs/1612.08083, arXiv:1612.08083.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (ISSN: 1063-6919) pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Díaz-Vico, D., Dorronsoro, J.R., 2020. Deep least squares Fisher discriminant analysis. IEEE Trans. Neural Netw. Learn. Syst. 31 (8), 2752–2763. http://dx.doi.org/10.1109/TNNLS.2019.2906302, Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89 (1–2), 31–71. http://dx.doi.org/10.1016/S0004-3702(96)00034-3, URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370296000343.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer 1 (8), 800–810. http://dx.doi.org/10.1038/s43018-020-0085-8, URL: https://www.nature.com/articles/s43018-020-0085-8, Number: 8 Publisher: Nature Publishing Group.

Gershenwald, J.E., Scolyer, R.A., 2018. Melanoma staging: American joint committee on cancer (AJCC) 8th edition and beyond. Ann. Surg. Oncol. 25 (8), 2105–2110. http://dx.doi.org/10.1245/s10434-018-6513-7.

Ghaffari Laleh, N., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., Buelow, R.D., Grabsch, H.I., Brenner, H., Chang-Claude, J., Alwers, E., Brinker, T.J., Khader, F., Truhn, D., Gaisa, N.T., Boor, P., Hoffmeister, M., Schulz, V., Kather, J.N., 2022. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. 79, 102474. http://dx.doi.org/10.1016/j.media.2022.102474, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841522001219.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised Learning. http://dx.doi.org/10.48550/arXiv.2006.07733, URL: http://arxiv.org/abs/2006.07733, arXiv:2006.07733 [cs, stat].

Guo, B., Li, X., Yang, M., Jonnagaddala, J., Zhang, H., Xu, X.S., 2023. 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. J. Pathol.: Clin. Res. 9 (3), 223–235. http://dx.doi.org/10.1002/cjp2.312, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10073932/.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, (ISSN: 1063-6919) pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Heath, A.P., Ferretti, V., Agrawal, S., An, M., Angelakos, J.C., Arya, R., Bajari, R., Baqar, B., Barnowski, J.H.B., Burt, J., Catton, A., Chan, B.F., Chu, F., Cullion, K., Davidsen, T., Do, P.-M., Dompierre, C., Ferguson, M.L., Fitzsimons, M.S., Ford, M., Fukuma, M., Gaheen, S., Ganji, G.L., Garcia, T.I., George, S.S., Gerhard, D.S., Gerthoffert, F., Gomez, F., Han, K., Hernandez, K.M., Issac, B., Jackson, R., Jensen, M.A., Joshi, S., Kadam, A., Khurana, A., Kim, K.M.J., Kraft, V.E., Li, S., Lichtenberg, T.M., Lodato, J., Lolla, L., Martinov, P., Mazzone, J.A., Miller, D.P., Miller, I., Miller, J.S., Miyauchi, K., Murphy, M.W., Nullet, T., Ogwara, R.O., Ortuño, F.M., Pedrosa, J., Pham, P.L., Popov, M.Y., Porter, J.J., Powell, R., Rademacher, K., Reid, C.P., Rich, S., Rogel, B., Sahni, H., Savage, J.H., Schmitt, K.A., Simmons, T.J., Sislow, J., Spring, J., Stein, L., Sullivan, S., Tang, Y., Thiagarajan, M., Troyer, H.D., Wang, C., Wang, Z., West, B.L., Wilmer, A., Wilson, S., Wu, K., Wysocki, W.P., Xiang, L., Yamada, J.T., Yang, L., Yu, C., Yung, C.K., Zenklusen, J.C., Zhang, J., Zhang, Z., Zhao, Y., Zubair, A., Staudt, L.M., Grossman, R.L., 2021. The NCI genomic data commons. Nature Genet. 53 (3), 257–262. http://dx.doi.org/10.1038/s41588-021-00791-5, URL: https://www.nature.com/articles/s41588-021-00791-5, Number: 3 Publisher: Nature Publishing Group.

Hildebrand, L.A., Pierce, C.J., Dennis, M., Paracha, M., Maoz, A., 2021. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. Cancers 13 (3), 391. http://dx.doi.org/10.3390/cancers13030391.

Huang, A.C., Zappasodi, R., 2022. A decade of checkpoint blockade immunotherapy in melanoma: understanding the molecular basis for immune sensitivity and resistance. Nat. Immunol. 23 (5), 660–670. http://dx.doi.org/10.1038/s41590-022-01141-1.

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. arXiv:1802.04712 [cs, stat], URL: http://arxiv.org/abs/1802.04712, arXiv:1802.04712.

Kacew, A.J., Strohbehn, G.W., Saulsberry, L., Laiteerapong, N., Cipriani, N.A., Kather, J.N., Pearson, A.T., 2021. Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping. Front. Oncol. 11, URL: https://www.frontiersin.org/articles/10.3389/fonc.2021.630953.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53 (282), 457–481. http://dx.doi.org/10.1080/01621459.1958.10501452, URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452.

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., Grabsch, H.I., Yoshikawa, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Trautwein, C., Luedde, T., 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. 25 (7), 1054–1056. http://dx.doi.org/10.1038/s41591-019-0462-y, URL: https://www.nature.com/articles/s41591-019-0462-y, Number: 7 Publisher: Nature Publishing Group.

Kim, R.H., Nomikou, S., Coudray, N., Jour, G., Dawood, Z., Hong, R., Esteva, E., Sakellaropoulos, T., Donnelly, D., Moran, U., Hatzimemos, A., Weber, J.S., Razavian, N., Aifantis, I., Fenyo, D., Snuderl, M., Shapiro, R., Berman, R.S., Osman, I., Tsirigos, A., 2020. A deep learning approach for rapid mutational screening in melanoma. bioRxiv, 610311, http://dx.doi.org/10.1101/610311, URL: https://www.biorxiv.org/content/10.1101/610311v2, Publisher: Cold Spring Harbor Laboratory Section: New Results.

Kim, R.H., Nomikou, S., Coudray, N., Jour, G., Dawood, Z., Hong, R., Esteva, E., Sakellaropoulos, T., Donnelly, D., Moran, U., Hatzimemos, A., Weber, J.S., Razavian, N., Aifantis, I., Fenyo, D., Snuderl, M., Shapiro, R., Berman, R.S., Osman, I., Tsirigos, A., 2022. Deep learning and pathomics analyses reveal cell nuclei as important features for mutation prediction of BRAF-mutated melanomas. J. Invest. Dermatol. 142 (6), 1650–1658.e6. http://dx.doi.org/10.1016/j.jid.2021.09.034, URL: https://www.sciencedirect.com/science/article/pii/S0022202X21024052.

Kjeldsen, J.W., Lorentzen, C.L., Martinenaite, E., Ellebaek, E., Donia, M., Holmstroem, R.B., Klausen, T.W., Madsen, C.O., Ahmed, S.M., Weis-Banke, S.E., Holmström, M.O., Hendel, H.W., Ehrnrooth, E., Zocca, M.-B., Pedersen, A.W., Andersen, M.H., Svane, I.M., 2021. A phase 1/2 trial of an immune-modulatory vaccine against IDO/PD-L1 in combination with nivolumab in metastatic melanoma. Nat. Med. 27 (12), 2212–2223. http://dx.doi.org/10.1038/s41591-021-01544-x, URL: https://www.nature.com/articles/s41591-021-01544-x, Number: 12 Publisher: Nature Publishing Group.

Lapin, M., Hein, M., Schiele, B., 2016. Loss functions for top-k error: Analysis and insights. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1468–1477.

Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. arXiv:2011.08939.

Li, J., Li, W., Sisk, A., Ye, H., Wallace, W.D., Speier, W., Arnold, C.W., 2020. A multi-resolution model for histopathology image classification and localization with multiple instance learning. URL: http://arxiv.org/abs/2011.02679, arXiv:2011.02679 [cs, eess].

Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J., 2021b. DT-MIL: Deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, Cham, pp. 206–216.

Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W., 2023. Interventional bag multi-instance learning on whole-slide pathological images. arXiv:2303.06873.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2021. On the variance of the adaptive learning rate and beyond. arXiv:1908.03265.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. arXiv:1803.01534.

Lu, W., Toss, M., Rakha, E., Rajpoot, N., Minhas, F., 2021b. SlideGraph+: Whole slide image level graphs to predict HER2status in breast cancer. URL: http://arxiv.org/abs/2110.06042, arXiv:2110.06042 [cs].

Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2020. Data efficient and weakly supervised computational pathology on whole slide images. arXiv:2004.09666 [cs, eess, q-bio], URL: http://arxiv.org/abs/2004.09666, arXiv:2004.09666.

Lu, M.Y., Zhao, M., Shady, M., Lipkova, J., Chen, T.Y., Williamson, D.F.K., Mahmood, F., 2021a. Deep learning-based computational pathology predicts origins for cancers of unknown primary. Nature 594 (7861), 106–110. http://dx.doi.org/10.1038/s41586-021-03512-4, URL: http://arxiv.org/abs/2006.13932, arXiv:2006.13932 [cs, q-bio].

Nahhas, O.S.M.E., Loeffler, C.M.L., Carrero, Z.I., van Treeck, M., Kolbinger, F.R., Hewitt, K.J., Muti, H.S., Graziani, M., Zeng, Q., Calderaro, J., Ortiz-Brüchle, N., Yuan, T., Hoffmeister, M., Brenner, H., Brobeil, A., Reis-Filho, J.S., Kather, J.N., 2023. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. http://dx.doi.org/10.48550/arXiv.2304.05153, URL: http://arxiv.org/abs/2304.05153, arXiv:2304.05153 [cs].

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., 2009. Serum 25-Hydroxyvitamin D3 levels are associated with breslow thickness at presentation and survival from melanoma. J. Clin. Oncol. 27 (32), 5439–5444. http://dx.doi.org/10.1200/JCO.2009.22.1135, URL: https://ascopubs.org/doi/10.1200/JCO.2009.22.1135, Publisher: Wolters Kluwer.

Newton-Bishop, J.A., Davies, J.R., Latheef, F., Randerson-Moor, J., Chan, M., Gascoyne, J., Waseem, S., Haynes, S., O'Donovan, C., Bishop, D.T., 2015. 25-Hydroxyvitamin D2/D3 levels and factors associated with systemic inflammation and melanoma survival in the Leeds Melanoma Cohort. Int. J. Cancer 136 (12), 2890–2899. http://dx.doi.org/10.1002/ijc.29334, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29334, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.29334.

Noorbakhsh, J., Farahmand, S., Foroughi pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-ha, M., Zarringhalam, K., Chuang, J.H., 2020. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nature Commun. 11 (1), 6367. http://dx.doi.org/10.1038/s41467-020-20030-5, URL: http://www.nature.com/articles/s41467-020-20030-5.

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., 2018. β-Catenin–mediated immune evasion pathway frequently operates in primary cutaneous melanomas. http://dx.doi.org/10.1172/JCI95351, URL: https://www.jci.org/articles/view/95351/pdf, Publisher: American Society for Clinical Investigation.

Park, J.H., Kim, E.Y., Luchini, C., Eccher, A., Tizaoui, K., Shin, J.I., Lim, B.J., 2022. Artificial intelligence for predicting microsatellite instability based on tumor histomorphology: A systematic review. Int. J. Mol. Sci. 23 (5), 2462. http://dx.doi.org/10.3390/ijms23052462.

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., 2019. Genetic and environmental determinants of immune response to cutaneous melanoma. Cancer Res. 79 (10), 2684–2696. http://dx.doi.org/10.1158/0008-5472.CAN-18-2864, URL: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-18-2864.

Qu, H., Zhou, M., Yan, Z., Wang, H., Rustgi, V.K., Zhang, S., Gevaert, O., Metaxas, D.N., 2021. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. Npj Precis. Oncol. 5 (1), 1–11. http://dx.doi.org/10.1038/s41698-021-00225-9, URL: https://www.nature.com/articles/s41698-021-00225-9, Number: 1 Publisher: Nature Publishing Group.

Rawat, R.R., Ortega, I., Roy, P., Sha, F., Shibata, D., Ruderman, D., Agus, D.B., 2020. Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images. Sci. Rep. 10 (1), 7275. http://dx.doi.org/10.1038/s41598-020-64156-4, URL: https://www.nature.com/articles/s41598-020-64156-4, Number: 1 Publisher: Nature Publishing Group.

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., 2015. Improved overall survival in melanoma with combined dabrafenib and trametinib. N. Engl. J. Med. 372 (1), 30–39. http://dx.doi.org/10.1056/NEJMoa1412690.

Rotte, A., 2019. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. J. Exp. Clin. Cancer Res. 38 (1), 255. http://dx.doi.org/10.1186/s13046-019-1259-z.

Saillard, C., Dehaene, O., Marchand, T., Moindrot, O., Kamoun, A., Schmauch, B., Jegou, S., 2021. Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. arXiv:2109.05819 [cs, eess], URL: http://arxiv.org/abs/2109.05819, arXiv:2109.05819.

Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2021. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images. URL: http://arxiv.org/abs/2107.09405, arXiv:2107.09405 [cs, eess].

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y., 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. arXiv:2106.00908.

Sirinukunwattana, K., Domingo, E., Richman, S.D., Redmond, K.L., Blake, A., Verrill, C., Leedham, S.J., Chatzipli, A., Hardy, C., Whalley, C.M., Wu, C.-h., Beggs, A.D., McDermott, U., Dunne, P.D., Meade, A., Walker, S.M., Murray, G.I., Samuel, L., Seymour, M., Tomlinson, I., Quirke, P., Maughan, T., Rittscher, J., Koelzer, V.H., 2021. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. Gut 70 (3), 544–554. http://dx.doi.org/10.1136/gutjnl-2019-319866, URL: https://gut.bmj.com/lookup/doi/10.1136/gutjnl-2019-319866.

Sun, Q., Sun, H., Wu, N., Cong, L., Cong, X., 2020. Prognostic significance of tumor-infiltrating lymphocyte grade in melanoma: A meta-analysis. Dermatol. (Basel, Switzerland) 236 (6), 481–492. http://dx.doi.org/10.1159/000505152.

The Royal College of Pathologists, 2018. Meeting pathology demand: Histopathology workforce census. 2018, URL: https://www.rcpath.org/static/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf.

Tourniaire, P., Ilie, M., Hofman, P., Ayache, N., Delingette, H., 2023. MS-CLAM: Mixed supervision for the classification and localization of tumors in Whole Slide Images. Med. Image Anal. 85, 102763. http://dx.doi.org/10.1016/j.media.2023.102763, URL: https://www.sciencedirect.com/science/article/pii/S1361841523000245.

Ugurel, S., Röhmel, J., Ascierto, P.A., Flaherty, K.T., Grob, J.J., Hauschild, A., Larkin, J., Long, G.V., Lorigan, P., McArthur, G.A., Ribas, A., Robert, C., Schadendorf, D., Garbe, C., 2016. Survival of patients with advanced metastatic melanoma: The impact of novel therapies. Eur. J. Cancer (Oxford, England: 1990) 53, 125–134. http://dx.doi.org/10.1016/j.ejca.2015.09.013.

Valieris, R., Amaro, L., Osório, C.A.B.d.T., Bueno, A.P., Rosales Mitrowsky, R.A., Carraro, D.M., Nunes, D.N., Dias-Neto, E., da Silva, I.T., 2020. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers 12 (12), 3687. http://dx.doi.org/10.3390/cancers12123687, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7763049/.

Wolchok, J.D., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Lao, C.D., Cowey, C.L., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P.F., Smylie, M., Butler, M.O., Hill, A., Márquez-Rodas, I., Haanen, J.B.A.G., Guidoboni, M., Maio, M., Schöffski, P., Carlino, M.S., Lebbé, C., McArthur, G., Ascierto, P.A., Daniels, G.A., Long, G.V., Bas, T., Ritchings, C., Larkin, J., Hodi, F.S., 2021. Long-term outcomes with nivolumab plus ipilimumab or nivolumab alone versus ipilimumab in patients with advanced melanoma. J. Clin. Oncol. JCO.21.02229. http://dx.doi.org/10.1200/JCO.21.02229, URL: https://ascopubs.org/doi/full/10.1200/JCO.21.02229, Publisher: Wolters Kluwer.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V., 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. arXiv:2102.03902.

Xu, H., Park, S., Lee, S.H., Hwang, T.H., 2019. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. http://dx.doi.org/10.1101/554527, URL: https://www.biorxiv.org/content/10.1101/554527v1, Pages: 554527 Section: New Results.

Yu, K.-H., Wang, F., Berry, G.J., Ré, C., Altman, R.B., Snyder, M., Kohane, I.S., 2020. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. J. Am. Med. Inform. Assoc. : JAMIA 27 (5), 757–769. http://dx.doi.org/10.1093/jamia/ocz230, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7309263/.

Zhang, M.R., Lucas, J., Hinton, G., Ba, J., 2019. Lookahead optimizer: k steps forward, 1 step back. arXiv:1907.08610.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. arXiv:2203.12081.

Zormpas-Petridis, K., Failmezger, H., Raza, S.E.A., Roxanis, I., Jamin, Y., Yuan, Y., 2019. Superpixel-based conditional random fields (SuperCRF): Incorporating global and local context for enhanced deep learning in melanoma histopathology. Front. Oncol. 9, URL: https://www.frontiersin.org/articles/10.3389/fonc.2019.01045.