Artificial Intelligence In Medicine 143 (2023) 102628

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Malignant Mesothelioma subtyping via sampling driven multiple instance prediction on tissue image and cell morphology data

Mark Eastwood [a,*], Silviu Tudor Marc [b], Xiaohong Gao [b], Heba Sailem [c,d], Judith Offman [d,e], Emmanouil Karteris [f], Angeles Montero Fernandez [g], Danny Jonigk [h,i], William Cookson [j], Miriam Moffatt [j], Sanjay Popat [j], Fayyaz Minhas [a,1], Jan Lukas Robertus [j,1]

[a] Tissue Image Analytics Center, University of Warwick, United Kingdom
[b] Department of Computer Science, University of Middlesex, United Kingdom
[c] Institute of Biomedical Engineering, University of Oxford, United Kingdom
[d] Kings College London, United Kingdom
[e] Wolfson Institute of Population Health, Queen Mary University of London, United Kingdom
[f] Brunel University, United Kingdom
[g] Manchester University, United Kingdom
[h] German Center for Lung Research (DZL), BREATH, Hanover, Germany
[i] Institute of Pathology, Medical Faculty of RWTH Aachen University, Aachen, Germany
[j] National Heart and Lung Institute, Imperial College London, United Kingdom

ARTICLE INFO

Keywords:
Malignant Mesothelioma
Multiple instance learning
Computational pathology
Deep learning
Cancer subtyping

ABSTRACT

Malignant Mesothelioma is a difficult to diagnose and highly lethal cancer usually associated with asbestos exposure. It can be broadly classified into three subtypes: Epithelioid, Sarcomatoid, and a hybrid Biphasic subtype in which significant components of both of the previous subtypes are present. Early diagnosis and identification of the subtype informs treatment and can help improve patient outcome. However, the subtyping of malignant mesothelioma, and specifically the recognition of transitional features from routine histology slides has a high level of inter-observer variability.

In this work, we propose an end-to-end multiple instance learning (MIL) approach for malignant mesothelioma subtyping. This uses an adaptive instance-based sampling scheme for training deep convolutional neural networks on bags of image patches that allows learning on a wider range of relevant instances compared to max or top-N based MIL approaches. We also investigate augmenting the instance representation to include aggregate cellular morphology features from cell segmentation. The proposed MIL approach enables identification of malignant mesothelial subtypes of specific tissue regions. From this a continuous characterisation of a sample according to predominance of sarcomatoid vs epithelioid regions is possible, thus avoiding the arbitrary and highly subjective categorisation by currently used subtypes. Instance scoring also enables studying tumor heterogeneity and identifying patterns associated with different subtypes. We have evaluated the proposed method on a dataset of 234 tissue micro-array cores with an AUROC of 0.89 ± 0.05 for this task. The dataset and developed methodology is available for the community at: https://github.com/measty/PINS.

1. Introduction

Malignant Mesothelioma (MM) is an aggressive cancer of the pleural lining, primarily associated with asbestos exposure [1]. It has a long latency period from initial exposure, to eventual carcinogenesis, and is difficult to diagnose due to its nonspecific clinical manifestations. As a result, diagnosis is usually confirmed in an advanced stage [2], leading to the 5 year survival rate being less than 5% [3]. Hence there is an urgent clinical need to detect MM at its early onset when treatment is more effective. MM is classified into 3 subtypes [4], Epithelioid (EM), Biphasic (BM) and Sarcomatoid (SM) Mesothelioma, with Biphasic characterised by a mix of epithelioid and sarcomatoid components, including Transitional Mesothelioma (TM). Epithelioid mesothelioma are characterised by malignant cells that are cytologically round with varying grading of atypia. Sarcomatoid mesothelioma cells are generally recognised as malignant spindle cells [5]. The Epithelioid subtype

* Corresponding author.
E-mail address: Mark.Eastwood@warwick.ac.uk (M. Eastwood).
[1] Joint last authors.

https://doi.org/10.1016/j.artmed.2023.102628
Received 26 January 2023; Received in revised form 30 June 2023; Accepted 14 July 2023
Available online 17 July 2023

is more common, and is associated with relatively more favourable outcomes, whereas Biphasic and Sarcomatoid are associated with a progressively worse prognosis. Recent studies have also shown that the presence of transitional features of TM, which share intermediate cytology between epithelioid and spindle cell also indicate a poorer prognosis [6]. In MM, TM may represent an aspect of Epithelial Mesenchymal Transition (EMT), with cells differentiating between EM and SM, suggesting that MM cases may fall more naturally on a continuum of characterisation according to the relative prevalence of EM, SM and TM components. Part of the motivation for this work is to go beyond the current 2021 WHO 3 basic subtypes and move towards a system whereby we use sub-visual signals on individual cell level, to specify quantitatively where a MM sample lies on the EM-SM continuum.

While distinction of these three histological subtypes of MM is crucial to patient treatment, management and prognosis, it is challenging to differentiate EM, SM and BM through visual analysis as they tend to present similar features to transitional patterns at some stages.

A number of deep learning methods for analysing mesothelioma images have been developed recently. For example, SpindleMesoNET [7] can separate malignant SM from benign spindle cell mesothelial proliferations. This method uses region annotations on whole slide images to train a resnet patch classifier. This differs from our learning task, as we do not have region annotations and must rely only on core-level labels.

To address the challenges of assessing stromal invasion in small biopsies, the most accurate indicator of malignancy, the separation of benign and malignant mesothelial proliferations has been investigated [8], in both epithelial and spindle cell mesothelial processes. A recent approach for survival prediction of MM patients called MesoNet [9] uses an MIL solver originally developed for computer vision applications [10]. This has also been applied to classification of lymph node metastases in [11]. The model uses a resnet50 base followed by a 1-d convolution to give an instance level score. A small MLP prediction head on the top and bottom two instances then provides the bag-level label. Models based on learning on extremal instances can suffer from learning on only a small subset of the relevant instances during training.

These examples demonstrate the successful application of machine learning to some prediction tasks on MM tissue, however automated subtyping of mesothelioma from Hematoxylin and Eosin (H&E) stained tissue sections remains an open problem that has not been addressed in the literature.

One of the issues associated with development of automated computational pathology approaches for predicting malignant mesothelioma subtypes is that pathologist-assigned ground-truth labels for these images are typically available only at the case level. However, we are often interested in properties of smaller regions of a sample. To address this, tissue images can be tiled into patches for training of deep learning models and the case-level labels used as bag labels. Thus, mesothelioma subtyping can be categorised as a multiple instance learning (MIL) or weak-supervision problem. This class of problem was first introduced in [12], and various approaches have since been proposed for use in such problems.

An attention-inspired pooling method for MIL instance aggregation is proposed in [13]. Another attention-based MIL approach is introduced in [14]. Here, a dual stream approach is used where the final bag score is the mean of max instance pooling and an attention based weighted average of instances attended to by the max instance. This model is applied to Camelyon-16 and TCGA lung cancer datasets. Large datasets on prostate cancer, basal cell carcinoma and breast cancer metastases are assembled in [15] and used to train an MIL model backpropagating only the top instance per bag. In the IDaRS algorithm proposed in [16], for each slide the training instances used in epoch t are the top k ranked instances by prediction score from the previous epoch $t-1$, augmented by a number of randomly selected patches from the slide. This approach was used to predict the status of molecular pathways and detect key mutations in colorectal cancer. Of the approaches detailed in the literature, this is the closest conceptually to the approach taken in this paper. Our approach differs in that instead of the union of top N and a purely random sampling of instances, we instead sample at each iteration according to the current model score. By avoiding to use an arbitrary top N cutoff, and instead stochastically sampling the more positive scoring patches through a probability distribution based on how highly each patch is scored, our approach can adapt more closely to the actual distribution of positive instances in each bag, thus making better use of all instances for training.

Building on our previous work [17], here we present a simple yet effective approach to multiple instance learning for MM subtype prediction with the following major contributions:

1. The introduction of a novel MIL-based method for computational pathology tasks which addresses shortcomings identified in similar methods regarding robustness to initialisation and learning on only a small number of the relevant instances in the training data. Instead of learning on some variation of top N instances, we learn on instances sampled according to model score. Thus, learning is focused on the more positive instances but in a more natural, adaptive and smooth way that does not rely on an arbitrary, discontinuous cut off to select instances to be used.
2. The collection of a dataset of MM tissue cores labelled by subtype, which we make publicly available for further study by the community. We also address a prediction task, automated subtyping of MM tissue, which has not been covered in the literature to date.
3. The incorporation into the model of patch level cell morphology statistics derived from analysis of cell segmentation on the tissue images, as a way to introduce domain knowledge (specifically, the knowledge that cells are important histological entities within the tissue) into the model.

2. Data and preprocessing

The dataset used in this work is a collection of H&E stained Tissue Micro-arrays (TMAs) of tumor tissue biopsies collected from St. George's Hospital. It consists of 4 TMA slides each with an average size of $40,000 \times 40,000$ pixels scanned using a Hamamatsu Nanozoomer S360 scanner at 20× (0.4415 μm per pixel) with a total of 279 cores covering 102 separate cases (patients). After removal of dropped and severely damaged/incomplete cores, we are left with 234 cores, with 148 EM, 61 BM, and 25 SM cores. We perform Vahadane stain normalisation [18] to minimise systematic stain variability between slides and cores. We tile each core into patches of 224×224 pixels at 20× magnification. Patches consisting of less than 50% tissue, as determined by a tissue mask created via luminosity thresholding, were discarded. Only core-level labels are provided, detailed annotations describing how different regions of the core contribute to the core-level label are not available.

3. Problem formulation

As the biphasic subtype is a mix of epithelioid and sarcomatoid components, the subtyping task can be modelled as a two class problem, where the three subtypes act as a crude measure of how much of the positive class are present. If we can train a model that will provide instance scores that express how likely each patch is to be the positive class, we can move towards a more expressive characterisation of mesothelioma cores according to the proportion of sarcomatoid component they contain. As subtype labels are only available at the core level and not for individual image patches within each core, we model the subtype prediction task as a binary Multiple Instance Learning (MIL) problem, with sarcomatoid as the positive class. Under the MIL paradigm [12], an example is represented by a bag of instances, and a bag is considered positive if it contains at least one positive sample. The goal of an MIL predictor is to use training data consisting of bags with bag level labels only to predict both bag and instance level labels

in testing. Formally, let $B = \left\{x_1, \ldots, x_{n_B}\right\}$ be a bag corresponding to a single TMA core in our dataset, where x_i are instances (patches) within the bag. The number of instances n_B can vary across bags. Each core, represented by bag B, is associated with a label $Y_B \in \{0, 1\}$ in the training dataset. In our formulation, both sarcomatoid and biphasic cores are taken as positive bags ($Y_B = 1$), as in both cases a noticeable sarcomatoid component is present whereas epithelioid-labelled cores become negative examples ($Y_B = 0$). Our goal is then to build a machine learning model $F(B; \Phi)$ with trainable parameters Φ that can use a labelled training dataset $D = \{(B_1, Y_1), (B_2, Y_2), \ldots, (B_M, Y_M)\}$ to generate a predicted label for a test core B. This is done by denoted by aggregating instance level predictions $z_i = f(x_i; \phi)$ to give $Z_B = F(B; \Phi) = Agg(\{z_i = f(x_i; \phi) | x_i \in B\})$ through an appropriate aggregation function $Agg(\cdot)$ such as max or average across top most positive instances.

Modelling the mesothelioma subtyping problem through MIL allows us to use the weakly supervisory signal from core-level labels to learn an instance-level scoring, with which we can identify predominantly EM or SM regions in a core. This enables us to quantify where each tissue component falls in the EM-to-SM continuum according to the proportion of positive (sarcomatoid) instances. This fine-grained and natural characterisation of a tumor can lead to more informed decisions regarding treatment etc. to be made.

4. Sampling-based MIL training for CNNs

We propose a simple but powerful approach for solving the MIL problem underlying mesothelioma subtyping based on the fundamental definition of MIL. In the binary case, MIL can be paraphrased as 'only the most positive instance in a bag counts'. Recall from Section 3 we label a bag as positive if it contains at least one positive instance. Intuitively, then, during training we wish to make the most positive scoring instances of negative bags less positive, and the positive instances of positive bags more positive. We would also like to avoid forcing negative instances in a positive bag to become positive labelled. Many approaches [9,11], rank instances according to an instance score, and learn only on the max (or top N) of these. However, this has some potential problems:

1. We learn only on very few instances. A significant proportion of the bag may be positive, but only the top few will contribute to learning per bag. This may be fine if we have many example bags to learn from, but can become a problem if we have relatively few bags as the model may rapidly over fit the resulting small number of top instances.
2. The method can be susceptible to unfortunate initialisation. If the initial weights of the model happen to score some unimportant instances highly, a situation may arise where the model is learning on a small subset of instances which have little to no relation to the bag labels, and may get stuck in an extremely sub-optimal local minimum.

In our approach, we minimise these issues by randomly sampling instances from each bag with a probability that is a continuous function of their instance score, sampling higher scoring instances more often. Formally, for each bag B we define a probability distribution P_B (initially uniform) over instances in B. Given the prediction scores $z_i = f(x_i, \phi) \in [0, 1]$ for an instance $x_i \in B$, from a CNN f with learnable weights ϕ, we set

$$P_B(i) = \frac{z_i^\alpha + c}{\sum_j (z_j^\alpha + c)}. \quad (1)$$

In Eq. (1), c is a small constant which limits how small $P_B(i)$ can get so that all instances are occasionally sampled, and α controls how heavily we weight for positive instances. For each training epoch, we sample 20% of the patches in each bag according to the distribution in Eq. (1) for training. In the extreme of $\alpha = 0$, all instances are weighted equally and we simply learn on all patches with label inherited from the bag label, disregarding the MIL setting. In the case of $\alpha \to \infty$ (and assuming c is reduced accordingly), we recover something similar to the max-based MIL approach of [10] or [9], where we learn only on the maximal instance of each bag. The pseudo-code for our method can be found in Algorithm. 1, and it is illustrated diagrammatically in Fig. 1.

Algorithm 1 Pseudo-code for MIL CNN Training

```
Initialise P_B: uniform distribution for all training
bags B
for e in epochs:
    S: Sample 20% instances ~ P_B from each
       training core
    For batch of instances X and bag labels Y in S:
        Z = f(X, φ)
        L = CE(Z, Y)      #cross-entropy loss
        Update φ to minimise L
    Save φ_best if validation AUC improves
    For instances x_i ∈ B in each training bag B:
        z_i = f(x_i, φ)   #inference pass
    Update P_B's according to Eq. (1)
Return best model f(·, φ_best)
```

This approach mitigates the problems mentioned earlier, as

1. We learn from all positive instances in a bag, not just the top N. As the probability distribution is calculated per bag (core), the method adjusts to the varying proportion of positive instances in different bags.
2. It is robust to initialisation, as initial probability distributions are likely to be fairly flat, and (assuming α not large) the sampling does not focus heavily on positive instances until the model has started to become more sure of its predictions (i.e when its outputs z_i become more polarised).

Our approach improves on similar models in the literature by removing the need for some arbitrary, discontinuous cut-off in the way we select instances to learn from during training, while still focusing the training on the most positive instances, which are the most important examples in a MIL setting. As the probability distribution to be sampled from is a continuous function calculated on the fly for each bag from the current model predictions at each iteration, it is adaptive to the different distributions of positive instances present in the bags in the training set.

5. Incorporating morphological features

We would ideally want that a deep learning model trained on histopathological images would learn to identify the morphological features of cells present in a patch that are relevant to the problem, together with any other important features of the tissue images. However this may not be the case when learning on a relatively small amount of weakly labelled data. In these cases, it can help to provide domain knowledge directly. To this end, we have used stardist [19] to segment the cells in each TMA core, and have associated with each image patch the cells contained within it. For each cell, we have calculated, using QuPath [20], a number of morphological features as follows:

- Shape features: Area, length, circularity, Max and Min diameter for both nucleus and whole cell
- Intensity features: Mean, Median and Standard Deviation for Hematoxylin and Eosin channels over cell nucleus, cell cytoplasm and whole cell
- Shape/intensity smoothed: Above features smoothed over nearby cells using a gaussian kernel of diameter 50 μm

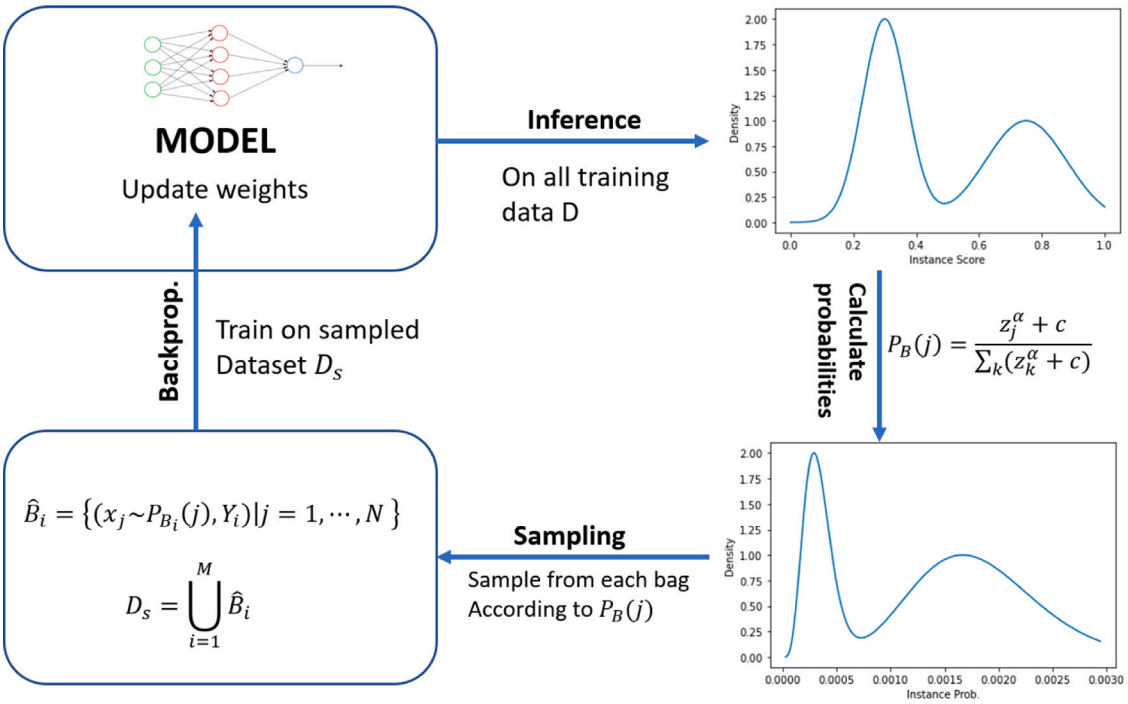

Fig. 1. Overview of proposed method, showing the instance scoring → weight calculation → sampling → training loop.

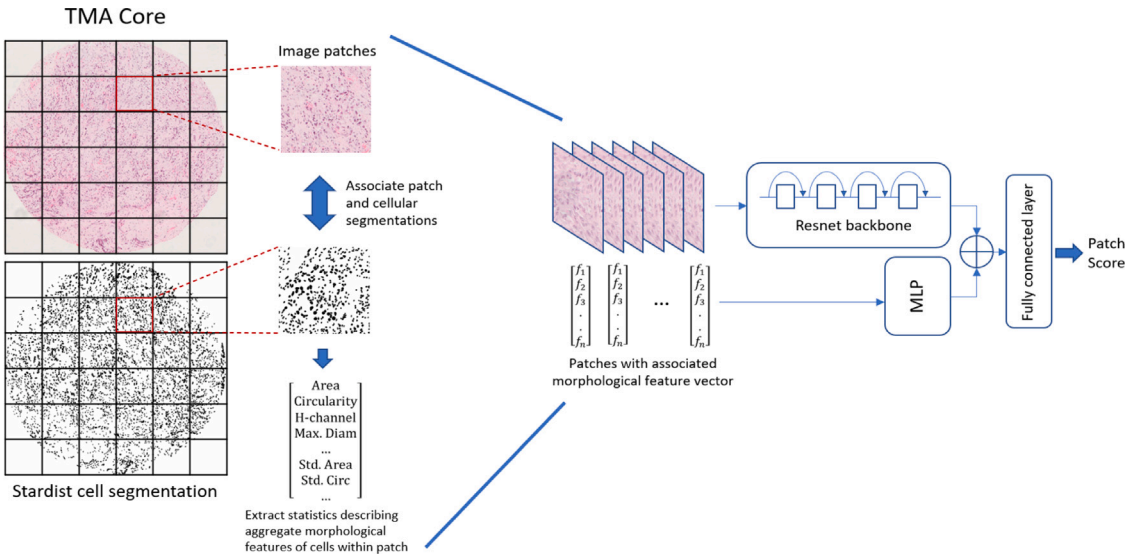

Fig. 2. Overview of model architecture and data pipeline. Cores are patched into 224 × 224 patches. For each patch, aggregate statistics on morphological features of cells that intersect with it are calculated. The patch image and morphological feature vector are passed to the model which outputs the patch score.

- Delaunay cluster features: number of neighbours, edge length statistics, cluster means of above features.
- Haralick texture features [21] on a small circular region around detection: calculated on the Eosin channel, the Hematoxylin channel and on the OD sum.

We have then calculated simple per-patch statistics (mean and standard deviation over the cells contained in the patch) for each of these morphological features, together with a cell count, and provided these as an additional input to the model for each patch. This results in a feature vector of length 321. The model incorporates this information via a small MLP, whose output is concatenated to the 512 features output by the resnet34 backbone before the final prediction head, as shown in Fig. 2. The remainder of our model training paradigm is kept the same.

In this way we can introduce domain knowledge, providing to the model the knowledge that cells are important histological entities within tumor tissue, and features describing their appearance are likely to be useful for subtype prediction. This means the model does not need to learn the concept of cells, and the relevant features of the cells, from scratch, which may be expecting too much of a model given relatively limited training data.

A better way for such domain information to enter the model would be through the use of a computational-pathology specific pre-trained backbone having learned a highly expressive representation for Cpath images. However currently no such general purpose Cpath backbone has emerged, and most CPath applications when doing transfer learning still rely on Imagenet pre-trained models.

6. Results and discussion

We use a ResNet34 pre-trained on ImageNet as the backbone in our CNN model [22], due to its consistently strong performance over a wide range of application areas including computational pathology [15,23], combined with its relatively small footprint. Larger models were not expected to provide much improvement due to the relatively small size of the dataset, meaning the additional capacity provided by larger networks would not be well utilised. We train our model using the Adam optimiser [24] with batch size of 64 over a maximum of 200 epochs with early stopping. Random rotations with equal probability of 0.25 for 0, 90, 180 or 270° rotation, in combination with flips with probability $p = 0.5$, and a small amount of colour jitter using the pytorch ColorJitter function (with strength arguments brightness=0.1, contrast=0.05, saturation=0.2, and hue=0.2) were applied to images during training. The learning rate used was 5×10^{-5}, weight decay 10^{-4}, with $\alpha = 2$ and $c = 0.01$ (See Eq. (1)). We choose a relatively low learning rate over a larger number of epochs because we update the probabilities used for sampling after each epoch, so we do not want the 'true' distribution to change too quickly over a single epoch. To address class imbalance, losses per class were weighted inversely to their class counts. We use a one-cycle learning rate schedule as introduced in [25]. During inference on cores, we aggregate the instance scores by averaging the top 5 instances. This is more robust than max aggregation, where a single poorly scored instance can completely change the aggregated score. Our model is implemented in PyTorch; code and data is available at https://github.com/measty/PINS. Models were trained on a workstation with an nvidia RTX 3080 12Gb graphics card, 64Gb RAM and an AMD Ryzen 9 5900X CPU. Training times varied due to early stopping, but were on the order of a day for a full cross-validation run.

For performance evaluation we employ a hold-one-out cross-validation strategy over slides, so that for each fold all cores of a single slide are held out as the test set. This is done to avoid any potential bias from systematic differences between slides, and to ensure no mixing of cores from the same patient occurs between the training and testing sets. The cores to be used for training are split 75%–25% into train and validation sets, respectively.

The results of our prediction model (which we name PINS for the Positive INstance Sampling that lies at its core) are reported in Table 1, together with baseline results from max-based MIL, which is the approach used to train the patch model in [15], and the model resulting from training on all patches with no regard for the MIL setting during training (naive-MIL in Table 1). We also provide results of CLAM [26], an attention based MIL approach, on our dataset. Our model achieves an AUROC of 0.83 and average precision (AP) of 0.73. The ROC curve for our method can be found in Fig. 3 . As can be seen from Table 1, the max-based MIL strategy performs poorly. This is likely due to the relatively small size of our training dataset which is orders of magnitude smaller than the very large dataset used in [15], which reported excellent performance using this strategy. Limiting learning to only one instance per core in each epoch exacerbates problems inherent in small datasets, as the model may rapidly overfit the top patches of positive bags. In contrast, our method allows learning from a wider selection of positive instances according to the model estimate of the proportion of positive instances in each bag resulting in much improved performance. Purely patch-based learning, that is simply learning on all patches labelled according to their bag label, performs surprisingly well, scoring quite close to our MIL method. This is likely due to the relatively high proportion of positive instances that are expected to be present in many of the positive bags (for example a sarcomatoid core is expected to comprise of mostly positive instances). This makes the implicit assumption a patch-based model makes, namely that all instances share the label of the bag, less wrong for this dataset compared to other MIL problems.

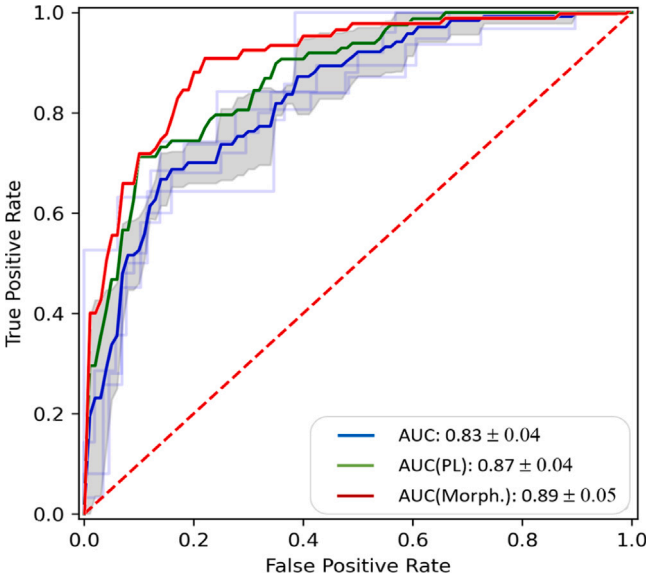

Fig. 3. ROC curves over 4-slide folds. Green and red plots shows curves after adjustment for labels from expert pathologist. Morph. denotes model on instances augmented with patch-level cell morphology features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

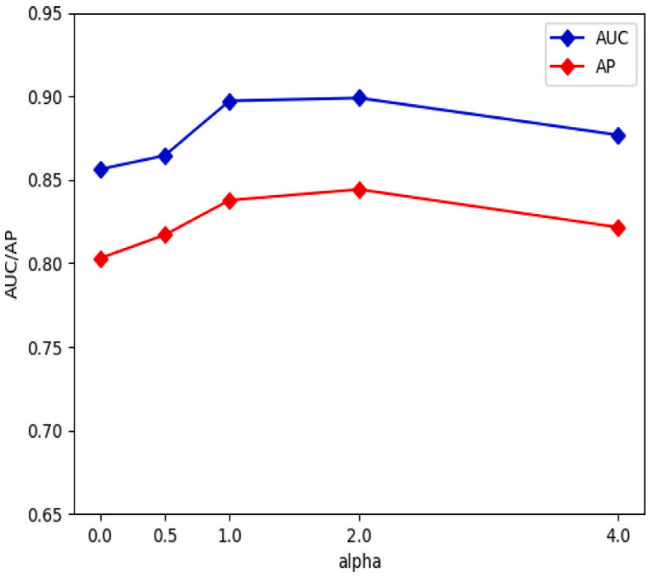

Fig. 4. AUC and AP of model as α is varied. α around 1–2 allows sampling on positive instances to occur without focusing to quickly or too sharply on a very small subset of the most positive instances.

Labels on histopathology images are often noisy, as the classification into clinical categories is subjective and opinion can vary significantly between pathologists. This is especially true in the context of MM, which is particularly difficult to diagnose. Thus, we sought an independent opinion from an expert pathologist on a small set of examples that were most consistently misclassified, to see to what extent the model could be justified on examples where its predictions differed from the original labelling.

Table 1
Summary of results (mean±stdev). PINS (P) indicates metric for a method after adjustment for labels from expert pathologist. PINS-M denotes model including morphological feature vector.

Metric	AUC-ROC	Avg. precision	Sensitivity	Specificity	Accuracy	f1 score
max-MIL	0.70 ± 0.04	0.58 ± 0.12	0.54 ± 0.07	0.73 ± 0.09	0.68 ± 0.03	0.54 ± 0.05
naive-MIL	0.81 ± 0.04	0.68 ± 0.11	0.72 ± 0.08	0.71 ± 0.1	0.74 ± 0.04	0.67 ± 0.03
PINS	0.83 ± 0.04	0.73 ± 0.09	0.77 ± 0.12	0.68 ± 0.11	0.75 ± 0.06	0.68 ± 0.11
PINS (P)	0.87 ± 0.04	0.81 ± 0.07	0.82 ± 0.1	0.71 ± 0.13	0.77 ± 0.03	0.72 ± 0.07
PINS-M (P)	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.84 \pm 0.08}$	$\mathbf{0.87 \pm 0.06}$	$\mathbf{0.77 \pm 0.05}$	$\mathbf{0.81 \pm 0.04}$	$\mathbf{0.77 \pm 0.05}$
CLAM (P)	0.84 ± 0.07	0.74 ± 0.11	0.75 ± 0.11	$\mathbf{0.77 \pm 0.02}$	0.77 ± 0.03	0.71 ± 0.06

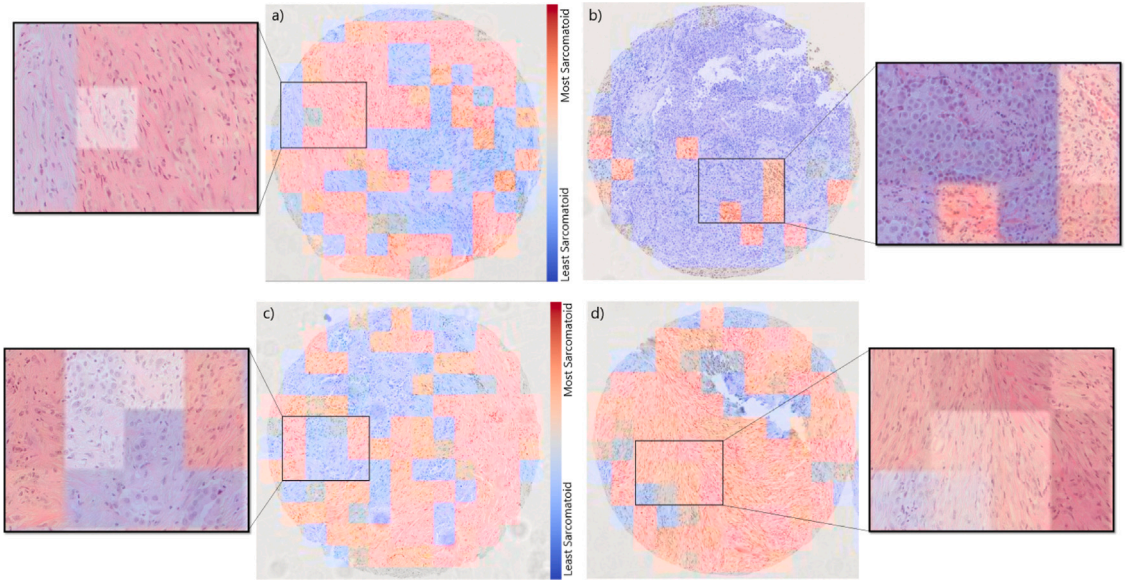

Fig. 5. Representative Heatmaps of model predictions. (a) a core labelled Epitheloid, which was consistently misclassified as positive (i.e significant SM component present). This agreed with the second opinion obtained from an expert pathologist, making this an example of a justified misclassification. From the closeup, spindle-like morphology of cells can be seen. (b) A correctly-predicted epithelioid-predominant core. As can be seen in (b) and the closeup of (c), patches demonstrating the typical rounded cell morphology of the EM subtype appear in bluer shades. (c) A correctly-predicted biphasic core with an even mix of EM and SM components. (d) A Sarcomatoid core, correctly predicted. In comparison to (a) and (c), has a much higher proportion of the core identified as SM.

In BM, a TMA core may represent a focal area that is specifically, either epithelioid or sarcomatoid. Of 14 consistently miss-classified cores, the opinion of the expert pathologist was that in 9 cases the model could be justified in its prediction given the representative core that was available for assessment. Further, 3 of the remaining cases contained very few tumor cells or were otherwise very challenging cases. Adjusting the ground truth for the 9 justified misclassifications to align with the pathologists assessment of the cores improves AUC (see Fig. 3) from 0.83 to 0.87, and AP from 0.73 to 0.81.

When providing the model both an rgb patch image and a vector of aggregate morphological features of the cells contained in the patch, performance further improves to an AUC of 0.89 and AP of 0.84. This confirms the expectation that in cases where training data is limited, if we can find a way to provide additional domain knowledge to the model we can achieve better performance. The confusion matrix for this model aggregated over all folds is

$$CM = \begin{bmatrix} 113 & 35 \\ 10 & 76 \end{bmatrix}$$

revealing that our models errors are skewed slightly towards false positive classifications.

An important parameter in our sampling-based MIL approach is α, which as discussed in Section 4 controls how heavily we weight on the instance score when determining the probability distribution to be used when sampling training instances. We have investigated the effect of varying this parameter between $\alpha = 0$ (no weighting by score) and $\alpha = 4$, a very heavy weighting on instance score. Results are shown in Fig. 4. We expect α between 1 and 2 will be appropriate in most cases, allowing training to focus on positive examples without the distribution becoming too heavily focused on high-scoring patches before the model has undergone sufficient training for scores on positive and negative instances to diverge significantly as the model starts to learn what positive instances look like. An extremely high α also forfeits one of the main advantages of our approach, which is to allow all positive instances, not just the very few highest scoring, to participate in training.

Heatmaps illustrating the output of our network are discussed in Fig. 5. Quantifying the proportion of a core which is predicted as SM subtype in this way could enable a much less subjective characterisation of a tissue sample. It also allows a more fine-grained characterisation of a core if desired.

7. Conclusions and future work

In this work, we demonstrate for the first time that an MIL framework can successfully predict presence of a sarcomatoid component in local tissue regions, paving the way for a quantitative categorisation of malignant mesothelioma subtypes. Incorporating the MIL setting into model training by sampling positive instances weighted on instance score instead of considering only the max or top few instances is shown to improve model performance. We believe our approach opens new opportunities for more objective assessment of epithelial-mesenchymal transformation where intra-tumor heterogeneity represents a gradient that can be difficult to assess by routine examination by histopathology. The output of the proposed model can be used to create a smoother continuum of disease classification by determining the extent of the

different cellular sub-populations at the patch level. Future work will be focused on including contextual information and identifying subtype at the cell level in addition to a detailed comparison with other backbone CNNs and larger-scale multi-centric evaluation on whole slide images. Due to a lack of ground truth cell segmentations on mesothelioma tissue, the presented work used a stardist model that has been trained for general cell segmentation on a variety of tissue. While the resulting segmentations were visually validated as being reasonable segmentations by expert pathologists involved in the study, this does represent a limitation of the study as quantitative validation of cell segmentation cannot be reported, and it is likely that segmentation could be improved through the use of a model fine-tuned on mesothelioma-specific cell boundary annotations. A cell segmentation model capable of predicting cellular phenotypes at single cell level could also be considered, to enable a more accurate assessment of tumor heterogeneity and aid pathological assessment. Explainability is an extremely important aspect of AI in the context of medicine, so another avenue of future work could be to add a layer of explainable AI such as [27] to highlight the features in image patches that are particularly relevant to the model prediction.

Declaration of competing interest

The authors have no known conflicts of interest.

Acknowledgements

This work was conducted as part of the PRISM project, kindly funded by Cancer Research UK through the CRUK-STFC Early Detection Innovation Award.

References

[1] Wagner JC, Sleggs CA, Marchand P. Diffuse pleural mesothelioma and asbestos exposure in the North Western Cape Province. Br J Ind Med 1960;17(13782506):260–71. http://dx.doi.org/10.1136/oem.17.4.260.

[2] Lagniau S, Lamote K, van Meerbeeck JP, Vermaelen KY. Biomarkers for early diagnosis of malignant mesothelioma: Do we need another moonshot? Oncotarget 2017;8(28881848):53751–62. http://dx.doi.org/10.18632/oncotarget.17910.

[3] Scherpereel A, Astoul P, Baas P, et al. Guidelines of the European Respiratory Society and the European Society of Thoracic Surgeons for the management of malignant pleural mesothelioma. Eur Respir J 2010;35(3):479–95.

[4] Ai J, Stevenson JP. Current issues in malignant pleural mesothelioma evaluation and management. 2014/07/24. Oncologist 2014;19(25061089):975–84. http://dx.doi.org/10.1634/theoncologist.2014-0122.

[5] WHO Classification of Tumours Editorial Board. Thoracic tumours, vol. 5. 5th ed. WHO Classification of Tumours; 2021.

[6] Dacic S. Pleural mesothelioma classification-update and challenges. Mod Pathol Off J United States Can Acad Pathol Inc 2021.

[7] Naso JR, Levine AB, Farahani H, et al. Deep-learning based classification distinguishes sarcomatoid malignant mesotheliomas from benign spindle cell mesothelial proliferations. Mod Pathol 2021;34(11):2028–35.

[8] Churg A, Colby TV, Cagle P, et al. The separation of benign and malignant mesothelial proliferations. Am J Surg Pathol 2000;24:1183–200.

[9] Courtiol P, Maussion C, Moarii M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 2019;25:1519–25.

[10] Durand T, Thome N, Cord M. WELDON: Weakly supervised learning of deep convolutional neural networks. In: 2016 IEEE conference on computer vision and pattern recognition. 2016, p. 4743–52. http://dx.doi.org/10.1109/CVPR.2016.513.

[11] Courtiol P, Tramel EW, Sanselme M, Wainrib G. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. 2018, arXiv abs/1802.02212.

[12] Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 1997;89(1):31–71. http://dx.doi.org/10.1016/S0004-3702(96)00034-3.

[13] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. 2018, arXiv preprint arXiv:1802.04712.

[14] Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 14318–28.

[15] Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;1–9.

[16] Bilal M, Raza SEA, Azam A, et al. Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. 2021, http://dx.doi.org/10.1101/2021.01.19.21250122, medRxiv.

[17] Eastwood M, Marc S, Gao X, Sailem H, Offman J, Karteris E, et al. Malignant mesothelioma subtyping of tissue images via sampling driven multiple instance prediction. 2022, p. 263–72.

[18] Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans Med Imaging 2016;35:1962–71.

[19] Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. In: Medical image computing and computer assisted intervention - MICCAI 2018 - 21st international conference, Granada, Spain, September 16–20, 2018, proceedings, part II. 2018, p. 265–73.

[20] Bankhead P, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7(1):16878. http://dx.doi.org/10.1038/s41598-017-17204-5.

[21] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;SMC-3(6):610–21. http://dx.doi.org/10.1109/TSMC.1973.4309314.

[22] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015, CoRR abs/1512.03385.

[23] Pinckaers H, Bulten W, van der Laak J, Litjens G. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. IEEE Trans Med Imaging 2021;40(7):1817–26.

[24] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2017, arXiv:1412.6980.

[25] Smith LN, Topin N. Super-convergence: Very fast training of residual networks using large learning rates. 2017, CoRR abs/1708.07120, arXiv:1708.07120.

[26] Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5(6):555–70.

[27] Leonardi G, Montani S, Striani M. Explainable process trace classification: An application to stroke. J Biomed Inform 2022;126:103981.