2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Learning Cellular Phenotypes through Supervision

Helen Theissen[1], Tapabrata Chakraborti[1], Stefano Malacrino[1],
Korsuk Sirinukunwattana[1,2], Daniel Royston[2,3] and Jens Rittscher[1,2]

***Abstract*— Image-based cell phenotyping is an important and open problem in computational pathology. The two principal challenges are: 1) making the cell cluster properties insensitive to experimental settings (like seed point and feature selection) and 2) ensuring that the phenotypes emerging are biologically relevant and support clinical reporting. To gauge robustness, we first compare the consistency of the phenotypes using self-supervised and supervised features. Through case classification, we analyse the relevance of the self-supervised and supervised feature sets with respect to the clinical diagnosis. In addition, we demonstrate how we can add model explainability through Shapley values to identify more disease relevant cellular phenotypes and measure their importance in context of the disease. Here, myeloproliferative neoplasms, a haematopoietic stem cell disorder, where one particular cell type is of diagnostic relevance is used as an exemplar. The experiments conducted on a set of bone marrow trephines demonstrate an improvement of 7.4 % in accuracy for case classification using cellular phenotypes derived from the supervised scenario.**

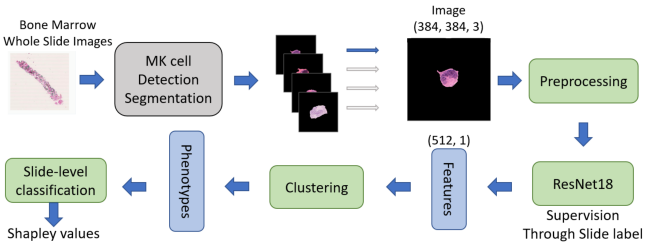


Fig. 1: Overview of the supervised feature extraction: Whole slide images pass through a detection and segmentation pipeline described in [5]. This yields image patches containing a megakaryocyte (MK) at the centre excluding the tissue background. The images are fed through a preprocessing stage including histogram equalisation and resizing before reaching the ResNet18 block. Supervision is provided by slide-level labels. The cellular phenotypes are extracted using clustering analysis. The case classification model is based on the proportion of phenotypes in each whole slide image.

I. Introduction

Methods for quantifying cellular morphology and cell phenotyping typically fall into two categories: Cell classification [1], [2] and clustering. Both methods require feature extraction and their performance depends on the quality of the extracted features. However, classification tasks require extensive annotations at the cell level. Unsupervised learning based on generative models is a popular methodology to extract morphological features. Ruan and Murphy [3] evaluate the performance of different autoencoder models based on the cell outline for shape reconstruction. They report that a higher dimensionality of the latent space in general benefits the models. Recently, a study on quantifying the morphological heterogeneity of cells and nuclei was published [4]. This approach provides classification, subtyping and visualisation of cells based on the cell contours. Sirinukunwattana *et al.* present a methodology for assessing megakaryocytes (MKs) relying on clustering to identify a set of representative phenotypes [5] for a single cell type.

Myeloproliferative neoplasms (MPN) are haematopoietic stem cell disorders and rare cancers [6]. Disease complications may include thromboembolic events such as stroke and heart attack. In the most severe cases, MPN patients can progress to frequently fatal bone marrow failure due to myelofibrosis and acute leukaemia [7]. Thus, early detection of MPNs is crucial especially in patients with high risk of bone marrow failure. However, the histological feature selection and quantification is subjective and not sufficiently defined to allow an accurate quantification [8]. In the context of MPNs, we are interested in quantifying the appearance of a single MK cell type whose morphology can vary considerably within the same whole slide image. Morphological criteria used for characterising megakaryocytes include cell-size, nuclear complexity, irregularity and spatial clustering. This makes cellular phenotyping of MK cells challenging.

Recently, Shapley values have gained attention in the context of model explainability. Lundberg *et al.* propose SHAP as a method to interpret models intuitively [9]. This method was adapted to explain the effects of patient features on the prediction of intraoperative hypoxaemia [10]. Recent approaches used handcrafted features to quantify cell morphology for classification [11], [12]. We posit that traditional handcrafted features lack the degrees of freedom to represent the complex morphology of MKs. Therefore, we consider deep features. Identifying features by optimising the image reconstruction loss does not make use of side information on the disease into account. Ferlaino *et al.* [13] demonstrate that supervised deep features extracted from cell classification capture intra-class variability. Thus, we use existing slide-level labels as a surrogate to train a cell classifier. Thus the extracted features focus on the differences between MK morphologies associated to different disease states. A brief

[1]Institute of Biomedical Engineering (IBME) and the Big Data Institute (BDI), Dept. of Engineering Science, University of Oxford, Oxford, UK
[2]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, Oxfordshire, UK
[3]Department of Cellular Pathology, John Radcliffe Hospital, Oxford University NHS Foundation Trust, Oxford, United Kingdom
*HT is funded by the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC), grant number EP/L016052/1. TC is supported by the UKRI Innovate UK DART Programme and the Oxford CRUK Cancer Centre.

overview of the proposed method is presented in Fig. 1, which is further elaborated in the next section.

Thus, the contributions of this paper are threefold: 1) Supervised feature extraction preserves intra-class heterogeneity, so that applying a clustering analysis improves the diagnostic relevance of the resulting cellular phenotypes. 2) By comparing supervised and unsupervised features we demonstrate that our supervised feature extraction approach increases their relevance without sacrificing robustness. 3) We use Shapley values to assess the importance of the extracted cellular phenotypes.

II. Methodology

We describe the methods for extracting supervised features and assessing feature relevance by applying Shapley values. The MK cells were extracted from the bone marrow trephine slides using the procedure described by Sirinukunwattana *et al.* [5]. A deforming autoencoder was applied to the MK patches to extract a latent representation based on minimising the reconstruction loss.

A. Increasing Relevance Through Supervision

Our objective is to extract a representation of MK cells with high relevance to the disease. As our data does not provide cell level labels this is achieved by using the slide level label for supervision in a weakly supervised setup. That is, we do not know to which degree a MK cell is abnormal, but we do know whether it stems from a diseased sample or not. Therefore, we classify MKs into two groups: the group of cells belonging to "normal" reactive samples and diseased samples. The feature extractor comprises a preprocessing stage and a ResNet18 binary classifier.

As the whole slide images suffer from non-uniform staining across slides and within the slides, we applied histogram equalisation on each channel (RGB) separately to account for colour variation. The background around the cells is excluded by using the segmentation mask obtained from the previous cell segmentation. The ResNet18 block is pretrained on ImageNet and then fine-tuned by classifying the MK cells into diseased and non-diseased classes. The resulting prediction for each cell can be viewed as the likelihood of a cell belonging to a diseased sample. MK cell features can be extracted from the last convolutional layer which results in a 512-dimensional feature vector.

B. Phenotyping through Cluster Analysis

The supervised and unsupervised features are fed to clustering analysis to identify cell phenotypes. In order to achieve comparable results, this stage should treat both sets similarly, that is the number of clusters should be the same and derived from the same clustering algorithm. For this reason, K-Means clustering is applied to identify clusters in both feature sets where each cluster will represent a cell phenotype.

As the auto-encoder yields rotationally variant features by definition, we decided to use the same mean feature vector chosen as Sirinukunwattana *et al.* [5]. It is composed of the feature sets for the image, the 180 ° rotation, the horizontal and vertical flip. In contrast to this, we use the ResNet features without applying this procedure. The reason for this is that even though CNN features are not inherently rotationally invariant, they can be trained to become rotationally invariant using appropriate data augmentation during training.

TABLE I: Bone marrow trephine dataset

Number	Total cohort	Reactive	ET	PV	MF
Slides	131	43	48	19	26
Cells	53866	7332	17724	12486	16324

The auto-encoder-based and ResNet-based feature extractors render feature vectors of 128 and 512 dimensions, respectively. To reduce the run time of the clustering, dimensionality was reduced to 30 dimensions for both, a trade-off between reducing the dimensionality enough to decrease the run time sufficiently, without losing too much information from the extracted features. Subsequently, the feature vectors were fed to K-Means clustering to obtain the cell phenotypes.

C. Feature Relevance Metric

In order to measure the relevance of the features to this disease we use two methods: 1) Case classification and 2) Shapley values. In terms of case classification, we can determine the general relevance of the features by using the traditional classification metrics such as accuracy, precision, recall and F1 score. Those inform on how well the classifier can make a prediction based on the input feature set. The input features are defined as the proportion of MK cells with each phenotype (cluster) in a slide.

Shapley values on the other hand have been applied successfully in other work [10] to gauge the importance of features in a machine learning model. They stem from coalitional game theory and allow to calculate the individual gain of a player in a cooperative game. Here, we use them to compare the importance of the cell phenotypes in context of the disease using the aforementioned case classification model trained on the self-supervised and supervised features.

III. Experiments

A. Bone Marrow Trephines from the Oxford Archive

In total 131 bone marrow trephine samples were sourced from the OUH NHS Foundation Trust archive which included 45 ET, 18 PV, 25 MF and 43 reactive or non-neoplastic samples. The latter group comprises patients without any evidence of either an underlying myeloid disorder nor a bone marrow malignancy. The other patient groups were selected based on an established or newly-diagnosed MPN according to the latest WHO classification scheme from 2016. The dataset we used for our experiments contains 53,866 MK cells of which more than 50 % were manually validated by a human expert. (Ref. Table I)

B. Training the Feature Extractor

The feature extraction stage was trained using 5-fold cross-validation (CV). Each CV set contains 27 slides with an approximately equal number of cells. It was ensured that all MPN disease subtypes were uniformly distributed across the

five sets. As evaluation metrics we used accuracy, precision, recall and AUC-ROC. Data imbalance was dealt with by a weighted random sampling in each batch. Parameter updating was done using the ADAM optimizer. Binary cross entropy loss was minimised during training. The batch size was set to 64 and the learning rate to 0.0001. We trained the model for 100 epochs. During training data augmentation was applied to reduce overfitting. Since cell size and shape potentially carry biological meaning we avoid transformations which change these. Instead we use colour perturbation, rotation, noise and horizontal and vertical flips.

C. Cellular Phenotyping and Cluster Quality

For this part we use the previously used CV sets. This means we extract features of the training set from the auto-encoder and ResNet18. These are then fed into K-Means clustering after dimensionality reduction through PCA. The hyper-parameters of K-Means are determined using the features extracted from the validation set. The cluster quality is assessed based on the silhouette score which is a standard metric for K-means. Subsequently, the optimal setting is chosen for cell phenotyping and applied to the test set. The procedure is repeated across all CV sets.

D. Case classification and Feature Relevance

For the case classification the proportion of cells of each phenotype in a slide is defined as the input feature vector. We chose a support vector machine as our sample size is close to the number of dimensions. Again, we partition the data into the previous CV sets and training, validation and test sets, respectively. Performance of the case classifier depending on the set of features, provides a measure for the relevance of the self-supervised and supervised features with respect to the disease. We calculate the Shapley values for each cell phenotype. Since the model-agnostic functionality is applied here, the resulting Shapley values are not an exact calculation, but rather an approximation. Nevertheless, they yield a description for the individual relevance of the cell phenotypes in context of the prediction.

IV. Results and Analysis

This section deals with the performance and results of the supervised feature extraction using a ResNet18 structure, the clustering analysis for cell phenotyping and the resulting relevance of the supervised and unsupervised features.

A. Feature Extraction

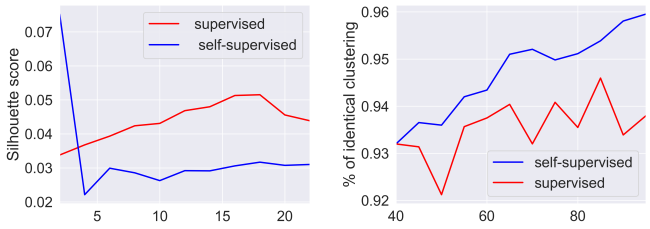
Table II depicts accuracy, recall and precision for applying the trained feature extractor on the test sets of the five CV sets A, B, C, D and E. The mean across all test sets is 88.7 %, 91.8 % and 96.2 %, for accuracy, recall and precision, respectively.

B. Clustering and Cluster Quality

In order to find a reasonable number of clusters we use k-means clustering on the training set and subsequently applied the trained method on the validation set. The silhouette score was chosen as the performance metric. It ranges between

TABLE II: Results for the supervised feature extractor

CV set	accuracy	recall	precision
A	0.8753	0.9099	0.9421
B	0.8556	0.8704	0.9461
C	0.8452	0.8508	0.9644
D	0.9340	0.9485	0.9762
E	0.9226	0.9315	0.9816
Overall	0.887	0.9184	0.9621



(a) Number of clusters **(b)** Percentage of training set

Fig. 2: Left: Performance of k-means clustering on the validation set: The silhouette scores with respect to the number of clusters is depicted for supervised (red) and self-supervised (blue) features. Right: Average trace of the pair confusion matrix obtained from fitting the clustering to 10 different subsets of training data for varying fractions and applying it to the validation data.

-1 and +1, where a high negative result means that data points have been assigned to the wrong clusters. Values close to zero indicate overlapping clusters. The silhouette scores using self-supervised and supervised features against cluster number is presented in Figure 2a. The reason for the silhouette score staying close to zero, might be the fact that only one cell type is examined and thus cluster boundaries are somewhat arbitrary. This is especially the case for an auto-encoder, which is trained to find a latent representation to reconstruct the image. In contrast, the supervised setting enforces features which separate cells which are more prevalent in diseased from ones in normal slides.

In order to gauge the stability of the resulting clusters we compared 10 different random states of the K-Means initialisation and clustering fitted to 10 different subsets of the training data. The pair confusion matrix takes into account all pairings of data points of the validation data and counting those pairs which were assigned to the same or different clusters under both clusterings. For varying initialisations, on average 96 % and 93.7 % of all pairings were assigned to correctly for unsupervised and supervised, respectively. Figure 2b shows the average trace of the pair confusion matrix for a varying size of the training data subset. As expected the clustering performance increases as more training data is being used. Again, the self-supervised features perform slightly better than the supervised features.

C. Feature Relevance

A slide representation composed of the proportion of cells belonging to the 18 cell phenotypes is used as the feature vector for case classification, where the label 0 and 1 corresponds to reactive and MPN diagnosis, respectively. C, γ denote the regularisation parameter, the kernel coefficient

TABLE III: Hyperparameter tuning for all CV sets (A, B, C, D and E): self-supervised (left) and supervised (right)

CV	C	γ	accuracy	CV	C	γ	accuracy
A	100	0.0001	0.89	A	10	0.001	0.95
B	100	0.0001	0.82	B	100	0.001	0.95
C	10	0.001	0.86	C	10	0.001	0.95
D	100	0.001	0.84	D	10	0.001	0.93
E	10	0.001	0.93	E	10	0.001	0.93

TABLE IV: CV case classification performance on the test sets

CV (self-supervised)	recall	precision	f1-score	accuracy
A	0.85	0.79	0.81	0.85
B	0.79	0.84	0.77	0.78
C	0.85	0.81	0.82	0.85
D	0.78	0.76	0.77	0.81
E	0.95	0.89	0.91	0.93
All	0.844	0.818	0.816	0.844

and the radial basis function kernel was used. Table III shows the optimal setting for SVM hyperparameters for all CV sets based on self-supervised (left) and supervised features (right), respectively. Subsequently, the optimal setting was used on the test set as depicted in tables IV and V for self-supervised and supervised, respectively. The macro-average is used for precision, recall and F1-score. The performance results for the supervised features show a significant improvement when compared to the self-supervised features.

D. Cell Phenotype Importance

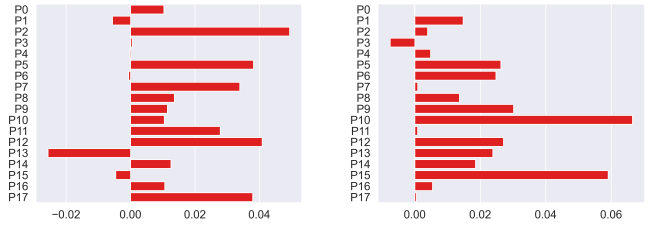
For assessing the importance of the individual phenotypes we use Shapley values. For each slide we obtain 18 Shapley values, one per phenotype, which sum up to the classification model output. Figure 3 depicts the mean Shapley value of each input feature of the case classification for self-supervised 3a and supervised 3b features with respect to the true class label. The increase in average impact across all CV sets and phenotypes on the model output is 0.198 for the supervised compared to self-supervised setting.

V. Conclusion

We demonstrate how identifying cellular phenotypes can benefit from additional clinical variables such as diagnosis by using supervised feature extraction. First, we compared self-supervised and supervised features with respect to clustering robustness. Through case classification we assessed the relevance of the supervised and self-supervised features in context of disease. By exploiting model explainability based on Shapley values the individual importance of the identified cellular phenotypes to the clinical problem was measured. The experiments conducted on the MK cells of a bone marrow trephine dataset showed an improvement of 7.4 % in accuracy for the case classification using cellular phenotypes from the supervised scenario. These promising results demonstrate that supervision using clinical side information can be used to inform cellular phenotyping through supervised feature extraction.

TABLE V: CV case classification performance on the test sets

CV (supervised)	recall	precision	f1-score	accuracy
A	0.93	0.81	0.85	0.89
B	0.94	0.97	0.96	0.96
C	0.85	0.89	0.85	0.85
D	0.97	0.94	0.95	0.96
E	0.92	0.92	0.92	0.93
All	0.922	0.906	0.906	0.918



(a) Self-supervised features **(b)** Supervised features

Fig. 3: Impact of phenotypes on the slide-level classifier output (mean Shapley value) of each feature on the case classification: The value zero corresponds to the mean model output across all predictions. Negative values counteract a correct prediction of the true class label. On average the positive impact of each phenotype is 0.03 and 0.041 for self-supervised supervised features making an increase 0.198 for all phenotypes for the supervised setting.

Declaration

JR and KS are co-founders of Ground Truth Labs, there are no other author conflicts. All relevant ethical approval and privacy consent were ensured for human patient data.

References

[1] L. Strbkova *et al.*, "Automated classification of cell morphology by coherence-controlled holographic microscopy," *Journal of biomedical optics*, 22(8): 086008, 2017.

[2] K. Yao, N. D. Rochman, and S. X. Sun, "Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning," *Scientific reports*, 9(1): 1-13, 2019.

[3] X. Ruan and R. F. Murphy, "Evaluation of methods for generative modeling of cell and nuclear shape," *Bioinformatics*, 35(14): 2475-2485, 2019.

[4] J. M. Phillip *et al.*, "A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei," *Nature protocols*, 16(2): 754-774, 2021.

[5] K. Sirinukunwattana *et al.*, "Artificial intelligence–based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in mpn patients," *Blood advances*, 4(14): 3284-3294, 2020.

[6] G. Gatta *et al.*, "Rare cancers are not so rare: the rare cancer burden in europe," *European journal of cancer*, 47(17): 2493-2511, 2011.

[7] T. Barbui, G. Finazzi, and A. Falanga, "Myeloproliferative neoplasms and thrombosis," *Blood*, 122(13): 2176-2184, 2013.

[8] B. S. Wilkins *et al.*, "Bone marrow pathology in essential thrombocythemia: interobserver reliability and utility for identifying disease subtypes," *Blood*, 111(1): 60-70, 2008.

[9] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[10] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, 2(10): 749, 2018.

[11] N. Petrović *et al.*, "Sickle-cell disease diagnosis support selecting the most appropriate machine learning method: Towards a general and interpretable approach for cell morphology analysis from microscopy images," *Computers in Biology and Medicine*, 126: 104027,2020.

[12] V. K. Lam *et al.*, "Quantitative assessment of cancer cell morphology and motility using telecentric digital holographic microscopy and machine learning," *Cytometry Part A*, 93(3): 334-345, 2018.

[13] M. Ferlaino *et al.*, "Towards deep cellular phenotyping in placental histology," *arXiv preprint arXiv:1804.03270*, 2018.