Analysis of automatically generated embedding guides for cell classification

Philipp Gräbel
Institute of Imaging and Computer Vision
RWTH Aachen University
Aachen, Germany
graebel@lfb.rwth-aachen.de

Julian Thull
Institute of Imaging and Computer Vision
RWTH Aachen University
Aachen, Germany

Martina Crysandt
Department of Hematology, Oncology, Hemostaseology and Stem Cell Transplantation
University Hospital RWTH Aachen University
Aachen, Germany

Barbara M. Klinkhammer
Institute of Pathology
University Hospital RWTH Aachen University
Aachen, Germany

Peter Boor
Institute of Pathology
University Hospital RWTH Aachen University
Aachen, Germany

Tim H. Brümmendorf
Department of Hematology, Oncology, Hemostaseology and Stem Cell Transplantation
University Hospital RWTH Aachen University
Aachen, Germany

Dorit Merhof
Institute of Imaging and Computer Vision
RWTH Aachen University
Aachen, Germany

***Abstract*—Automated cell classification in human bone marrow microscopy images could lead to faster acquisition and, therefore, to a considerably larger number of cells for the statistical cell count analysis. As basis for the diagnosis of hematopoietic diseases such as leukemia, this would be a significant improvement of clinical workflows. The classification of such cells, however, is challenging, partially due to dependencies between different cell types. In 2021, guided representation learning was introduced as an approach to include this domain knowledge into training by providing "embedding guides" as an optimization target for individual cell types.**

In this work, we propose improvements to guided representation learning by automatically generating guides based on graph optimization algorithms. We incorporate information about the visual similarity and the impact on diagnosis of misclassifications. We show that this reduces critical false predictions and improves the overall classification F-score by up to 2.5 percentage points.

***Index Terms*—representation learning, cell classification, embedding guides**

I. Introduction

Many diseases of the blood-forming system, including various types of leukemia, do not exclusively result in morphological changes of blood cells but in an alteration of the cell type distribution. Consequently, the key part of diagnosing such diseases lies in correctly identifying cell types of a large number of samples to obtain a statistically reliable estimation of this distribution. This analysis is typically performed visually based on stained microscopy images from bone marrow samples. Due to practical constraints, such as a limited amount of trained medical experts to perform this time-consuming task, the number of evaluated cells is typically fairly low.

Using deep neural networks, a much higher throughput could be achieved, resulting in a larger sample size for the estimation of the cell distribution. This could lead to more statistically accurate results and more objective estimations, particularly considering the inter-rater disagreement. With

This work was supported by the German Research Foundation (DFG) through the grants SFB/TRR57, SFB/TRR219, BO3755/6-1, and ME3737/3-1. It was further funded by the Federal Ministry of Education and Research (BMBF) and the Ministry of Culture and Science of North Rhine-Westphalia as part of the National High-Performance Computing four Computational Engineering Sciences (NHR4CES) consortium.

978-1-6654-6964-7/22/$31.00 ©2022 IEEE

sufficiently accurate models, this could be the basis for a faster and more reliable diagnosis.

However, the domain of microscopy images of bone marrow samples is highly demanding and not straight forward to conquer for neural networks. Not only is the image data itself challenging – with varying degrees of cell density and potential artifacts – the cell classification itself needs special considerations. In human bone marrow, cells develop from the hematopoietic stem cells and mature into specific types. This results in relationships between cell types that should not be ignored when using them as classes for a neural network. For example, immature cells across different lineages share more characteristics than fully formed mature cells. Furthermore, within a lineage, the maturation process is continuous such that two adjacent maturity stages are more similar compared to non-adjacent stages. This not only results in visual cues but also changes the severity of mis-classifications: an "off-by-one error" is less harmful than predicting a wrong lineage – this fact is also supported by inter-rater disagreements.

Related Work

A common approach for incorporating domain knowledge (such as the aforementioned relationships between cells) into neural network training is offered by representation learning techniques. In order to learn suitable embeddings (representations), loss functions such as the Triplet Margin Loss (TML) [1] enforce clusters representing the classes.

Instead of only minimizing intra-class distances and maximizing inter-class distances, more sophisticated methods can be employed to enforce constraints based on domain knowledge, for example with "embedding guides" [2]. These guides encode class relationships in the form of a two-dimensional embedding that was created manually by experts. In training, they can be enforced in various ways – most commonly by minimizing the distance between learnt embeddings and the guide. In particular, the authors propose the "Inverse UMAP Loss", for which an n-dimensional embedding position $e^c_{\text{nD, epoch}}$ is created for each class c for each epoch from the two-dimensional points of the guide e^c_{2D} using UMAP [3]. The computation is based on UMAP trained on the computed embedding after each epoch:

$$e^c_{nD,\text{epoch}} = \kappa\cdot e^c_{nD,\text{epoch}-1} + (1-\kappa)\cdot \text{UMAP}^{-1}(e^c_{2D}) \quad \forall c. \tag{1}$$

This loss is combined with classical TML:

$$\mathcal{L}^{\text{UMAP}} = (1-\alpha)\mathcal{L}_{\text{TML}}(\mathcal{T}_B) + \frac{\alpha}{|B|}\sum_{i=0}^{|B|-1}\mathcal{L}_1(e^{y_i}_{nD}, e_i), \tag{2}$$

where $\mathcal{T}_B$ represents all mined triplets in the batch B that violate the margin (triplet mining).

They furthermore propose the "Distance Loss", which enforces neighborhood relations while ignoring absolute positions by minimizing the difference of distances of pairs within the predicted and guide embeddings:

$$d(i,j) = L_1(L_1(e_i, e_j), L_1(e^{y_i}_{2d}, e^{y_j}_{2d})) \tag{3}$$

$$\mathcal{L}^{\text{DIST}} = \frac{2}{|B|(|B|-1)}\sum_{i=0}^{|B|-1}\sum_{j=0}^{i-1} d(i,j). \tag{4}$$

This loss also incorporates the TML.

Even though these methods show improved classification scores, only a single guide was investigated.

A large part of this work relies on graph visualization approaches that are used to generate embedding guides. In particular, we utilize the Kamada-Kawai [4], the Fruchtermann-Reingold [5], and the Distributed Recursive Layout [6] algorithms. The Kamada-Kawai algorithm minimizes the energies in a virtual system that represents graph nodes as particles connected by physical springs. The forces are computed using Hooke's law and the strength of the springs is determined by the value in an adjacency matrix. The Fruchtermann-Reingold algorithm follows a similar approach but with a fixed edge length for all edges. For the optimization, attracting forces are used for neighboring nodes and repelling forces for all nodes. The layout is computed iteratively, by first applying the attractive forces and the repelling forces, followed by a temperature computation to limit the total displacement of the layout in each iteration. The Distributed Recursive Layout algorithm is a multi-level force-directed method that uses simulated annealing. It iteratively refines the embedding in multiple resolutions through subsequent downsampling operations on the number of nodes.

Contribution

In this work, we propose alternatives to the manually crafted embedding guide [2]. We propose various methods of automatically generating guides based on visual similarity of cell types and the impact of mis-classifications for the diagnosis. We show that these approaches are not only viable options for reducing human supervision in guide creation but also improve results for hematopoietic cell classification.

II. Image Data

In this work, we focus on the classification of hematopoietic cells from human bone marrow samples. Each sample is stained using Pappenheim-staining [7], which is commonly used for visual analysis to make cell types easily distinguishable. The samples are digitized using automatic immersion oiling and a magnification of 63×. Medical experts subsequently select representative regions in each whole slide image, similarly to the manual workflow in clinical practice.

Individual cells are identified in a semi-automatic workflow [8]. Firstly, automatic detection based on U-Net [9] and Watershed [10], [11] is performed. Secondly, the results are manually validated and, if necessary, corrected.

The cell type of each individual cell is annotated by medical experts, typically in collaboration of two experts. As this annotation is used for classification, the terms "cell type" and

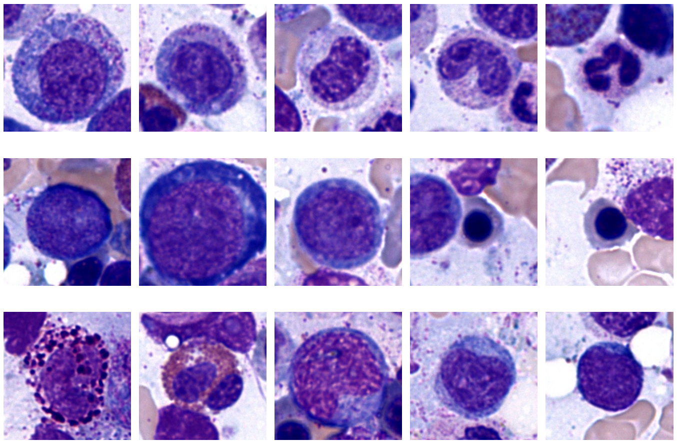


Fig. 1. Example cell images for each cell type. Top row: promyelocyte, myelocyte, metamyelocyte, band granulocyte and segmented granulocyte. Middle row: blast cell, proerythroblast, as well as basophilic, polychromatic and orthochromatic erythroblast. Bottom row: basophilic and eosinophilic granulocyte as well as a promonocyte, monocyte and lymphocyte.

"class" are used interchangeably in the following. For each cell, patches of size 224×224 px are extracted from the Whole Slide Image based on the annotations. Examples are shown in Figure 1.

In total, the dataset comprises 4 560 annotated cell patches. The number of samples per cell type varies greatly, in accordance with the typical distribution of cell types in the human bone marrow. This class-imbalance is one challenge that needs to be considered by classification approaches.

III. Methods

In the following sections, we will present the proposed methods starting with similarity and severity matrices that form the foundation of embedding guide generation. This is followed by an analysis of the resulting guide, an introduction to the training methods and specifications of the experimental setup.

A. Similarity and Severity Assessment

As the aim of this work is to incorporate biological domain knowledge, particularly the relationships between various cell types, this information needs to be encoded in a suitable format. We chose to focus on two essential aspects as described in the introduction: visual similarity and impact on the diagnosis.

1) Visual Similarity: To define visual similarity, we asked medical experts to encode how easily a sample of a specific cell type can be mis-interpreted as a specific different cell type. This is denoted by a integer number between 1 and 5. 1 indicates that it is virtually impossible to confuse cell type A as cell type B – indicating severe visual differences. 5 indicates that the two cell types are very similar in appearance and can thus be easily confused. The resulting matrix is presented in Figure 2.

Most visually similar cells can be found within a cell lineage across the maturity stages. Particularly adjacent maturity stages of the neutrophilic granulocytes show visual similarities. It should further be be noted that this matrix is not symmetric.

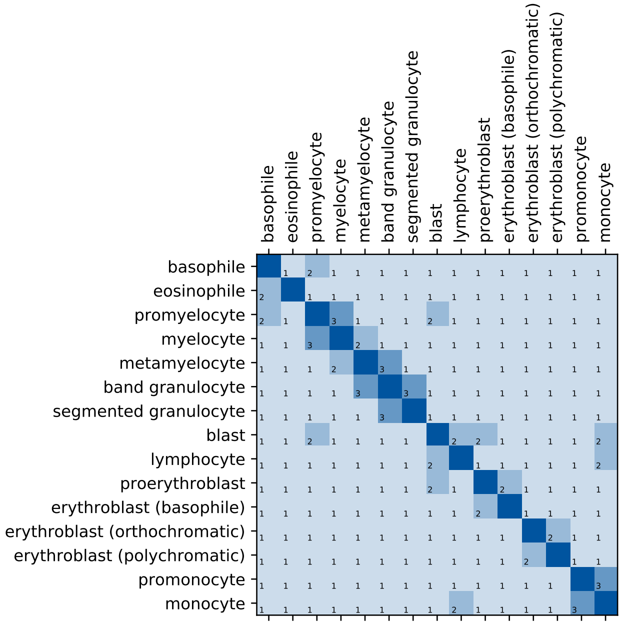


Fig. 2. Visual similarity matrix. Cells should be interpreted as "how easy is it to falsely classify the cell denoted by the row (true class) as the cell denoted by the column (predicted class)?" with values between 1 (severe visual differences, impossible to confuse) and 5 (visually similar, easily to confuse).

2) Severity: In a similar way, the impact on the diagnosis can be encoded. To this end, we focus on the model disease Chronic Myeloid Leukemia (CML), which is particularly characterized by a shift in the cell type distribution. Again, the value range is defined between 1 (low severity) and 5 (severe impact). The value indicates the severity of a mis-classification between two cell types.

The resulting matrix is presented in Figure 3. Again, the matrix is not fully symmetric. The numbers indicate, that almost every mis-classification has a severe impact. The only major exceptions are adjacent maturity stages within a single cell lineage, for example neutrophilic granulocytes or cells of the erythropoiesis.

B. Guide Generation

Based on the matrices presented in the previous section, two-dimensional embedding guides are created. Such guides are essentially mappings of each cell type c to a two-dimensional coordinate e^c_{2D}.
The similarity and severity matrices can be processed and subsequently interpreted as adjacency matrices, which are used in *force-directed* algorithms often found in visualization approaches for graphs. Typical examples for these, which we utilize in this work, are the Kamada-Kawai (KK), the Fruchtermann-Reingold (FR) and the Distributed Recursive Layout (DRL) algorithms as described in the introduction. We only utilize unweighted adjacency values, as not all of the

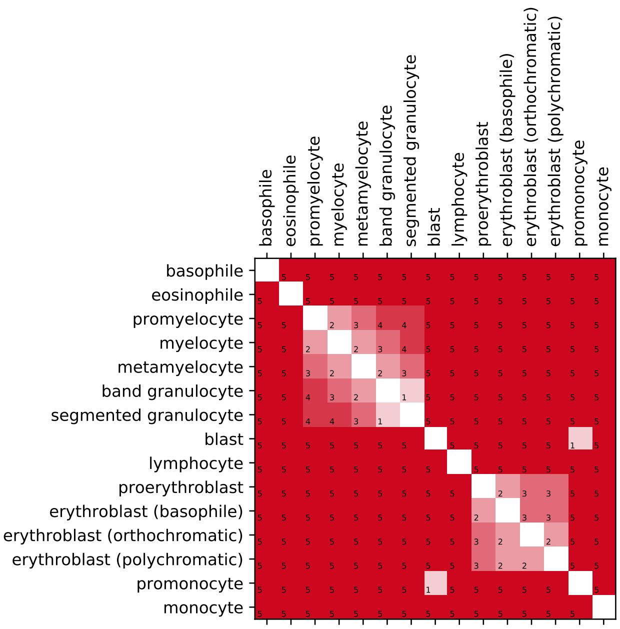


Fig. 3. Severity matrix. Cells should be interpreted as “how severe is the impact on the diagnosis of Chronic Myeloid Leukemia (CML) if the cell denoted by the row (true class) is falsely classified as the cell denoted by the column (predicted class)?” with values between 1 (low severity) to 5 (very severe).

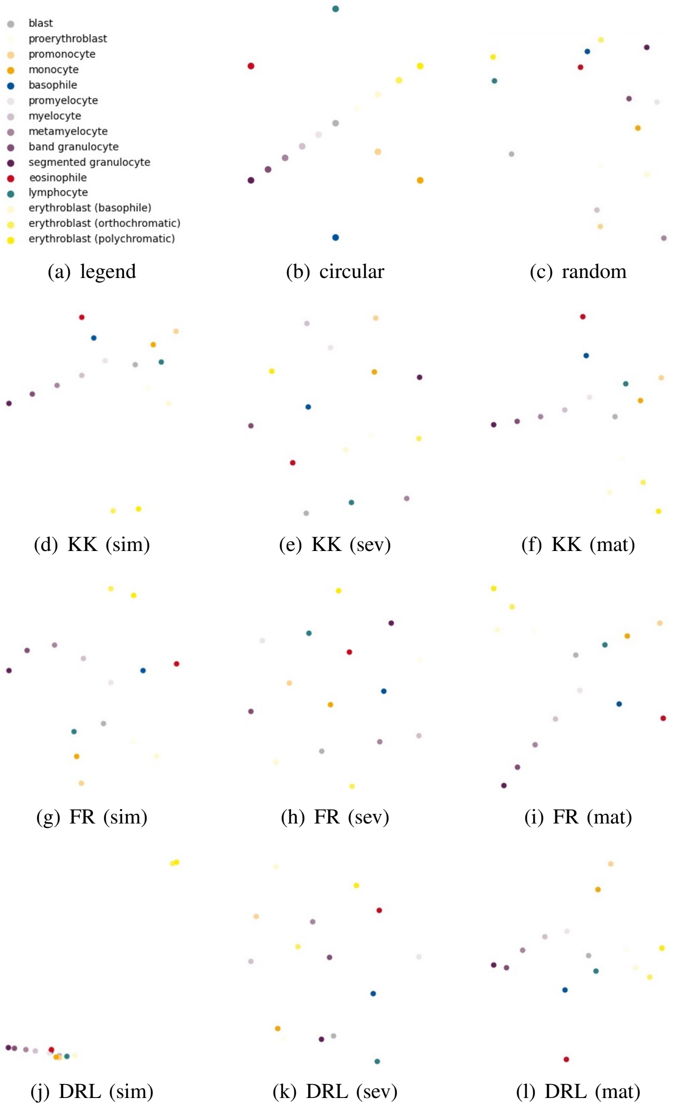
methods (Fruchtermann-Reingold) can handle weighted adjacency matrices. We transform the matrices into unweighted adjacency matrices by removing the lowest entries, normalizing the entries into the range $[0, 1]$ and subsequently rounding each entry to either 0 or 1. Additionally, we propose to incorporate the maturity information by setting the adjacency values of cells of the same cell lineage to 1.

In the Kamada-Kawai algorithm, we utilize the adjacency matrix directly to compute the parameters of the virtual springs. These are applied such that more adjacent cells – for examples visually similar cells or cell pairs that have a low impact on the diagnosis if mis-classified – are “pulled” closer together.

The Fruchtermann-Reingold algorithm is set up in a similar way. The required initial embedding for the iterative optimization is produced randomly.

In the Distributed Recursive Layout algorithm, the similarity and severity matrices are applied similarly to the KK-algorithm.

1) Baselines: We furthermore use the circular embedding proposed by Gräbel et al. [2] and a randomly generated embeddings as baselines.



(a) legend (b) circular (c) random

(d) KK (sim) (e) KK (sev) (f) KK (mat)

(g) FR (sim) (h) FR (sev) (i) FR (mat)

(j) DRL (sim) (k) DRL (sev) (l) DRL (mat)

Fig. 4. Embedding guides created using the Kamada-Kawai (KK), the Fruchtermann-Reingold (FR) or the Distributed Recursive Layout (DRL) algorithm. The parameters are based on either the visual similarity matrix (sim), the severity matrix (sev) or the similarity matrix with inclusion of maturity progress information (mat). Furthermore, the manually created circular embedding guide as well as a random guide are shown as baselines. Each point corresponds to the two-dimensional embedding of a cell type. Figure a) contains the color legend for the mapping to individual cell types.

Figure 4 shows the resulting guides. Many automatically generated embeddings follow similar characteristics as the circular reference embedding guide when including the maturity progressions: particularly the neutrophilic granulocytes and cells of the erythropoiesis are often shown along a line or curve. Without this information and based purely on the matrices, the embedding guides appear less structured and they utilize more of the available embedding space.

C. Loss Computation

For employing the embedding guides in the training process, we utilize the same methods as proposed in [2]: the Inverse UMAP Loss and the Distance Loss. For both, we determine the hyper-parameters in a preliminary hyper-parameter analysis on a subset of the data. This yields $\alpha = 0.4$ and $\kappa = 0.2$ as the most suitable values for the Inverse UMAP Loss and $\alpha = 0.4$ for the Distance Loss.

D. Experimental Setup

As encoder network, we utilize a DenseNet-121 [12], which has shown superior performance in similar tasks. The network is pre-trained on ImageNet [13]. For data augmentation and regularization, we utilize dropout ($p = 0.05$, random crop and random rotation.

The embedding has a dimensionality of 256. In addition to the losses mentioned previously, we minimize the norm of the embedding vector with a factor of 0.001 as additional regularization. The margin for TML is set to 0.1.

Training and evaluation is performed in a six-fold cross-validation, with one fold each for validation and testing. Training is performed until the macro F-score on the validation set does not improve for 50 epochs. This threshold has been determined based on observation of the validation loss and local minima to put a larger emphasis on results compared to computation time. The classification scores are computed from the embeddings using a Support Vector Machine (SVM) [14] with an Radial Basis Function (RBF) kernel [15]. The SVM is tuned with respect to hyper-parameters γ and c using grid searches every 5 training epochs.

In addition to the macro F-score, which is less impacted by the problem of class imbalance and thus the most commonly used evaluation criterion for this approach, we evaluate the mis-classification severity (MCS). The MCS score is obtained by element-wise multiplication of the confusion matrix $C \in \mathbb{R}^{mxm}$ and severity matrix $S \in \mathbb{R}^{mxm}$ divided by the number of samples n.

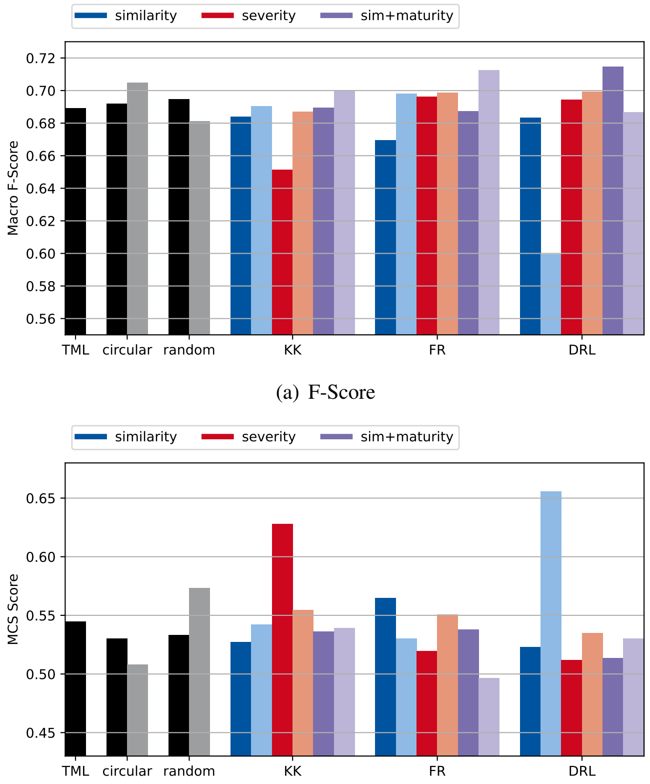
$$\text{MCS} = \frac{1}{n} \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} S_{i,j} C_{i,j} \tag{5}$$

A high MCS value indicates very severe mis-classifications, while a lower value indicates mis-classifications that have a lower impact on the diagnosis of CML. While this score correlates with the F-score, it has a higher emphasis on the domain knowledge: for example, mis-classifications between adjacent maturity stages will have a lower impact on the MCS score than on the F-score.

IV. Results

Figure 5 shows the results for each of the presented guides as well as for the baselines. Many of the proposed guides improve both classification score as well as mis-classification severity on the given dataset. Often, they perform better not only compared to the Triplet Margin Loss, but also compared to the circular embedding guide. The randomized embedding guide is able of improving results as well when used with inverse UMAP optimization but is detrimental with the distance loss.

The highest classification scores are achieved using the inverse UMAP optimization with a guide generated using the DRL algorithm based on the combination of similarity matrix and maturity information. Also for the mis-classification severity, this method achieves excellent results. Close results for classification and superior results for MCS are found using the Furchterman-Reingold method again based on similarity and maturity using distance loss optimization.



(a) F-Score

(b) MCS Score

Fig. 5. Classification results as macro F-Score and Mis-classification Severity (MCS) score for each guide using either the inverse UMAP loss (darker shade) or the distance loss (lighter shade). For comparison, the TML loss using no guide is shown as well. For the F-Score, higher values are better, for the MCS, lower values are better.

V. Discussion

First, it should be noted that the embedding guide proposed in [2] slightly outperforms training without guides on the given dataset. This, however, is also the case for a randomized embedding when using the inverse UMAP loss. This can be explained by a clearer optimization target with guided loss functions, which might be helpful given the class imbalance and overall dataset size.

Interestingly, the distance loss yields better results, which differs from the findings in [2]. A logical explanation lies in the lower stability of the UMAP-based method. Particularly without growing embeddings as proposed in [2], inverse UMAP leads to lower performance.

It should further be noted that the embedding spaces are not pre-defined anymore, as was the case for manually crafted guides. However, the embedding guides in Figure 4 still show a clear, suitably arranged embedding space. Particularly the well-performing guides, which are based on similarity and maturity information, show clear delineations of maturity progressions. With the distance loss, predicted embeddings can be globally rotated arbitrarily with respect to these guides.

While not defined in terms of absolute positions, the relative arrangement of cell types in the embedding space allows visual interpretation of results.

In terms of classification F-score and mis-classification severity, improvements to all baselines can be reported. Particularly the similarity-based guide with incorporation of the maturity information proves to be a very suitable guide. This applies to all guide generation methods but is particularly successful for FR and DRL.

Compared to training with TML, we achieve F-score improvements of up to 2.5 percentage points and a reduction of the mis-classification severity of up to 4.8 percentage points.

Both visual similarity and mis-classification severity cannot be precisely defined and encoded as a number. Therefore, both matrices are at least in part subjective even though multiple experts have collaborated in assigning numbers. Furthermore, a different scale could be chosen. The range of 1 to 5 is a compromise between very granular, which might be more accurate but difficult to assign, and very coarse, which would be technically less useful but easier to assign. As the results show, the encodings have a positive impact in terms of classification scores. Nevertheless, an analysis of other measures will be considered in further research.

VI. Conclusion

In this work, we proposed an alternative method for the generation of embedding guides to incorporate domain knowledge into representation learning. To this end, we encoded the visual similarity between cell types and the severity of mis-classifications with respect to the diagnosis into matrices. These are used as adjacency matrices in graph optimization algorithms to yield suitable guides. We show that these guides lead to improved classification performance and to less severe mis-classification in the case of hematopoietic cell classification.

Acknowledgment

This work is a purely retrospective, pseudonymized analysis of bone marrow samples under the Helsinki Declaration of 1975/2000 with written informed consent of all patients. The authors would like to thank Reinhild Herwartz and Melanie Baumann for their efforts in sample preparation and annotation.

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[2] P. Gräbel, M. Crysandt, B. M. Klinkhammer, P. Boor, T. H. Brümmendorf, and D. Merhof, "Guided representation learning for the classification of hematopoietic cells," 2021.

[3] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[4] T. Kamada, S. Kawai *et al.*, "An algorithm for drawing general undirected graphs," *Information processing letters*, vol. 31, no. 1, pp. 7–15, 1989.

[5] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[6] S. Martin, W. M. Brown, and B. N. Wylie, "Drl: Distributed recursive (graph) layout," Sandia National Laboratories, Tech. Rep., 2007.

[7] T. Binder, H. Diem, R. Fuchs, K. Gutensohn, and T. Nebe, "Pappenheim-färbung: Beschreibung einer hämatologischen standardfärbung–geschichte, chemie, durchführung, artefakte und problemlösungen/pappenheim stain: Description of a hematological standard stain–history, chemistry, procedure, artifacts and problem solutions," *Laboratoriumsmedizin*, vol. 36, no. 5, pp. 293–309, 2012.

[8] P. Gräbel, Özcan Özkan, M. Crysandt, R. Herwartz, M. Bauman, B. M. Klinkhammer, P. Boor, T. H. Brümmendorf, and D. Merhof, "State of the art cell detection in bone marrow slide images," *Journal of Pathology Informatics*, vol. 12, no. 1, p. 36, 2021. [Online]. Available: https://www.jpathinformatics.org/text.asp?2021/12/1/36/326215

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] S. Beucher, "Use of watersheds in contour detection," in *Proceedings of the International Workshop on Image Processing*. CCETT, 1979.

[11] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," *Mathematical morphology in image processing*, vol. 34, pp. 433–481, 1993.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] J.-P. Vert, K. Tsuda, and B. Schölkopf, "A primer on kernel methods," *Kernel methods in computational biology*, vol. 47, pp. 35–70, 2004.