Multi-class semantic cell segmentation and classification of aplasia in bone marrow histology images

Leander van Eekelen*[a, b], Hans Pinckaers[b, c], Konnie M. Hebeda[b], and Geert Litjens[b, c]

[a]Faculty of Biomedical Engineering, Technical University Eindhoven
[b]Computational Pathology Group, Radboud University Medical Center, The Netherlands
[c]Department of Pathology, Radboud University Medical Center, The Netherlands

ABSTRACT

Bone marrow biopsies play a central role in hematopathology for diagnosing a variety of diseases, staging lymphomas or performing follow-up progression. Tasks performed while examining biopsies include counting cells and estimating the ratio of various hematopoietic lineages. Inter- and intra-observer variability between hematopathologists in the outcome of these tasks has been shown to be significant, which could result in multiple pathologists diagnosing some patients differently. To that end, this paper presents a fully-convolutional neural network (FCNN) architecture to segment six classes in bone marrow trephine biopsies, which could assist hematopathologists in identifying and delineating cells, thus reducing inter- and intra-observer variability. Additionally, to show an application of the neural network to a clinically relevant task, the output of the network is used to train a classifier capable of distinguishing between normocellular and aplastic bone marrow. Results indicate the network is successfully capable of segmenting cells with an average detection rate of 83%. The classifier for distinguishing normocellular/aplastic bone marrow reaches an AUC of 0.990, showing that is is capable of automatically identifying aplasia.

Keywords: Digital pathology, bone marrow trephine biopsy, cell segmentation, aplastic anemia

1. INTRODUCTION

Bone marrow biopsies are the gold-standard to diagnose various suspected hematopathologies, such as leukemia or anemia. Besides diagnosis, samples can also be of use to stage lymphomas or follow-up progression of various diseases.[1] Bone marrow biopsies are often evaluated by specialized pathologists. These hematopathologist perform multiple tasks during the inspection of the biopsies, for example, inspecting marrow architecture and estimating the quantity and ratio of various hematopoietic lineages. The outcome of these tasks show inter- and intra-observer variability among pathologists, meaning that there is inherent uncertainty and subjectivity in the estimates.[2] Because these tasks play a key role in diagnosing hematopathologies, different pathologists may diagnose some patients differently.

The poor agreement on histological staging, counting cells and estimating cellularity of hematopathologists[3–6] could be explained by the repetitive actions and subjective interpretation that are needed to evaluate the biopsies. With the introduction and gradual adoption of digital pathology workflow,[7] where biopsy slides are digitized into whole-slide images (WSI), it has become possible to apply automated image analysis techniques to biopsy slides. A segmentation algorithm to identify and delineate cell types could support hematopathologists and help reduce inter- and intra-observer variability, increasing consistency of, for example, estimating ratios of hematopoietic lineages. This will support hematopathologists in reaching an accurate diagnosis more often.

In this paper, a fully-convolutional neural network (FCNN) architecture[8] is presented, trained using only sparse annotations, capable of segmenting six different classes in bone marrow trephine biopsy whole-slide images: myelopoietic and erythropoietic cells, matured erythrocytes, megakaryocytes, bone and lipocytes. High resolution segmentation of individual cells is accomplished by utilizing filter dilation[8] in the network at test time. To show the clinical relevance of this work, the classification results of the deep neural network are used to train a classification model that can distinguish between normocellular bone marrow and aplastic bone marrow,

leander.vaneekelen@radboudumc.nl

characteristic of aplastic anemia. Previously performed work on the automatic image analysis of bone marrow tissue has focused on detection and classification of only erythropoietic and myelopoietic cells.[9] This work, to the best of the knowledge of the authors, is the first to perform cell segmentation of all major cell types in bone marrow tissue and directly use this segmentation for a clinical application.

2. MATERIAL AND METHODS

2.1 Materials

2.1.1 Dataset for bone marrow cell segmentation

For training and evaluating the network, 24 biopsies were used, each extracted from a different patient at the Radboud University Medical Center. A hematopathologist confirmed that these biopsies were normocellular. The biopsies were stained with a periodic acid-Schiff (PAS) stain: this stain aids the hematopathologist in differentiating between myelopoietic cells (PAS positive) and erythropoietic cells (PAS negative)[1] and was hypothesized to help the performance of the network as well due to this visual contrast. All biopsies were scanned, resulting in whole-slide images (WSI), using a *3DHistech Panoramic Flash II 250 scanner* with a pixel resolution of approximately 0.25 m. The 24 images were randomly divided into three sets for training, validation and testing of 14, 5, 5 images respectively. The validation set was used to tune the hyperparameters of the various techniques used for segmentation. The test set was used to evaluate the segmentation performance and was not used until the finalization of the network.

For the training and validation set, sparse annotations were made by an experienced hematopathologist and two trained non-experts, in the form of point annotations for erythropoiesis, megakaryocytes, fat cells and myelopoiesis and polygon masks for areas of erythrocytes and bone. Erythropoiesis and myelopoiesis was only annotated by the hematopathologist, as the visual distinction was deemed to be too complex for the non-experts. Point annotations were transformed to circles with dataset-wide average diameters of the respective cells, to increase the available datapoints for the dataset (Figure 1a, b).

For the test set, 9 bounding boxes of on average 500 by 500 μm were drawn across 2 images, sampling a representative distribution of cells. All cells within these bounding boxes were annotated by the hematopathologist to avoid potential bias that can occur when randomly selecting cells across an entire image. Cells that did not belong to the six classes (endothelium, macrophages, etc.) or could not be recognized were excluded from evaluation. The point annotations were transformed in a similar fashion to the training and validation sets.

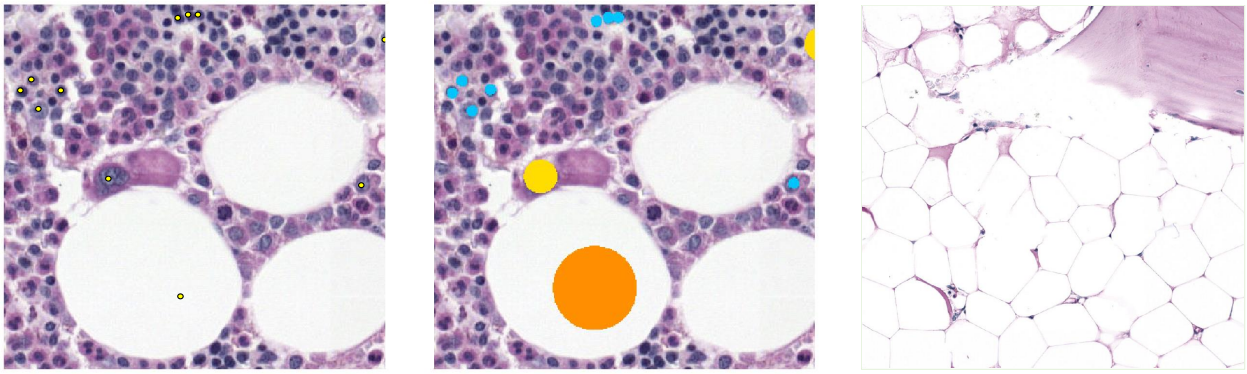

Figure 1: Point annotations were transformed to circles to increase the number of available datapoints and quantify the segmentation performance of the FCNN. On the right, a patch of aplastic bone marrow is shown to indicate the contrast between normocellular and aplastic bone marrow.

2.1.2 Dataset for classification of normocellular versus aplastic bone marrow

For the classification of aplastic bone marrow, positive cases were drawn from 31 patients diagnosed with aplastic anemia in the Radboud University Medical Center between 2016 and 2019. Archival biopsies of these patients were scanned and processed in the same way as the bone marrow cell segmentation dataset, which served as negative cases for this dataset. An example patch is shown in Figure 1c: aplastic bone marrow is characterized by a drastic reduction in cellularity and increase in fat cells.[1]

2.2 Methods

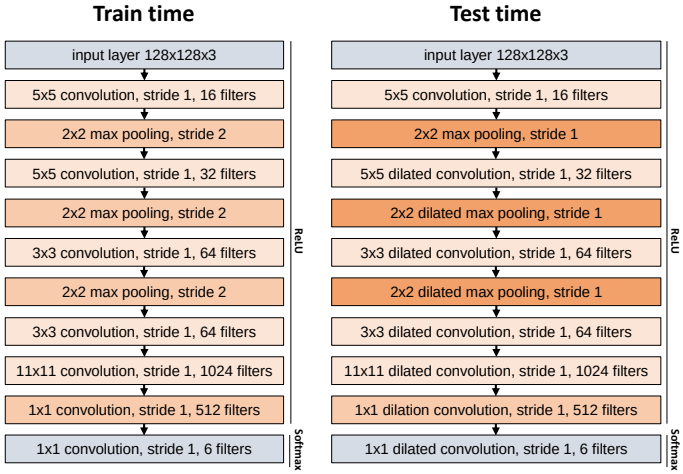

Figure 2: Overview of the network architecture during train and test time. During test time, dilated convolutional and max pooling layers are used to produce a prediction map of equal resolution as the input image, improving the quality of segmentation significantly.

2.2.1 Network architecture

For the task of bone marrow cell semantic segmentation, a 10-layer fully convolutional neural network (FCNN) was trained for the pixel-wise classification of six classes: myelopoietic and erythropoietic cells, erythrocytes, megakaryocytes, bone and lipocytes. Together, the pixel-wise classification forms a segmentation of the classes. The network was trained on patches of 128x128 of resolution 0.5 μm/px, empirically determined during initial experiments on the validation set to the best balance between spatial context surrounding the cells and cell-specific details.

The network (Figure 2) first begins with two repetitions of a 5x5 convolution followed by max-pooling. Then, one convolution of 3x3 followed by max-pooling occurs. Lastly, the network widens in channels for two convolutions of 3x3 and 11x11 respectively, before narrowing down for the last two convolutional layers, both 1x1. All max-pooling operations had a kernel of 2x2 with a stride of 2. Rectified linear units were used as non-linear activation function for all convolutions except the last one, where a softmax function was used.

The network was trained on a NVIDIA GTX 1080TI for a hundred epochs. The network weights with the lowest loss on the validation set were used as a final model. Experiments with data augmentation (rotations, flips and additive noise/blurring) indicated no significant gain in performance or even decrease in performance, so it was omitted. The network was optimized using Adam[10] with a constant learning rate of 0.0001 and a batch size of 32. Cross entropy was used as the loss function.

2.2.2 Test time

All WSIs were inferenced using a patch size of 512x512. During inference, the loss of resolution due to pooling layers was prevented by using filter dilation,[8] resulting in a likelihood map of the same resolution as the input image. This was needed to achieve segmentation of individual cells. Filter dilation works by setting all layers D_i with stride $s_i > 1$ to stride $\hat{s}_i = 1$ and dilating filters of all subsequent layers L_j (convolutional and pooling, $j > l$) by a factor of $\hat{d}_j = \prod_{i<j} s_i$.

Additionally, two noise reducing techniques were used to suppress spurious predictions, noticed throughout the development of the network. Firstly, test time augmentation is applied by inferencing an image six times (four 90 degree flips and x/y-axis flips), then averaging over the results. Secondly, connected component analysis was used to remove spuriously detected objects. Objects of class i were removed if the largest axis of its bounding box was smaller than 50 percent of the class i dataset-wide average diameter. Removed objects were filled in with the most occurring class immediately surrounding their bounding boxes.

2.2.3 Evaluation

Due to a lack of full annotations, the Dice score could not be used a metric for segmentation performance. Instead, segmentation performance is approximated by determining per cell from class C_i with area A_i if $\geq$50 percent of A_i is predicted as class C_j. If $i = j$, the prediction counts as a true positive. If $i \neq j$, it counts as a false positive. The segmentation performance on the validation and test set is visualized in two normalized confusion matrices.

2.2.4 Classification of normocellular versus aplastic bone marrow

For the task of classifying positive and negative cases of aplastic bone marrow, the network was applied to all slides, generating prediction maps. Because there were not enough WSIs available to predict cases of aplastic bone marrow on a biopsy basis, two sets of 1000 patches of positive and negative cases were sampled from their respective prediction maps. On average 40 patches were sampled per WSI. The patches were 512x512 of resolution 2.0 μm/px, empirically determined to be a good compromise between representativity of WSI-wide cell class distribution and the number of patches retrievable from each WSI without overlap. Each patch was turned into a feature vector of six components by counting the number of pixel occurrences of each class and standardizing each feature to zero mean and a standard deviation of one. The 2000 feature vectors were randomly distributed over a training, validation and test set of 1200, 400 and 400 vectors respectively, keeping the ratio between positive and negative samples equal.

The feature vectors were used to train three classifiers: a support vector machine (SVM), a k-nearest neighbor classifier (kNN) and a multi-variate logistic regression model. The SVM classifier was trained using a radial basis function (RBF) kernel, a gamma of 1/6, and a standard penalty parameter C of 1.0. The kNN classifier was optimized using a grid search across $k = 1, 2, \ldots, 100$ on the validation set, giving k = 5 as the best performing hyperparameter. The logistic regression classifier was trained using l2 regularization and a standard penalty parameter C of 1.0. A ROC curve was generated for all classifiers and the classifier with the highest AUC was selected to be applied on the test set.

3. RESULTS

The confusion matrices depicting the segmentation performance of the neural network on the validation and test are shown side by side in Figure 3. The lower bound in the test set for the correctly segmented percentage of cells is 70 percent for the erythrocytes class, while bone was segmented perfectly across both the validation and test set. This resulted in an average detection rate of 83%. Figure 5 shows visual results of the segmentation output of the network. The dataset for classification of normocellular/aplastic bone marrow was transformed using principal component analysis along its two principal components (with myelopoietic cells and fat cells respectively predicting 45 and 25 percent of the variance present in the data) and is visualized in Figure 4a. The SVM classifier performed best on the validation set with an AUC of 0.999 and was applied on the test set to achieve a final AUC of 0.990. The ROC curves with accompanying AUC metric are reported in Figure 4b for all classifiers used on the validation set and the SVM on the test set.

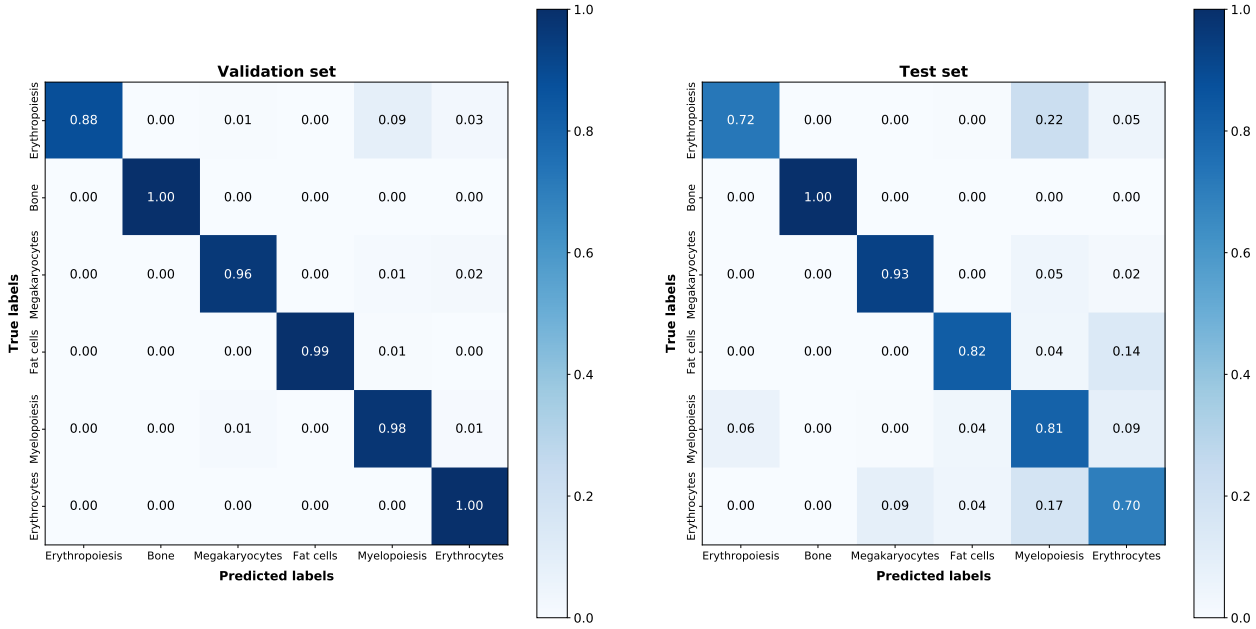

Figure 3: Confusion matrices for segmentation performance on the validation and test set. Most errors by the neural network are made in distinguishing between erythropoiesis, myelopoiesis and erythrocytes.

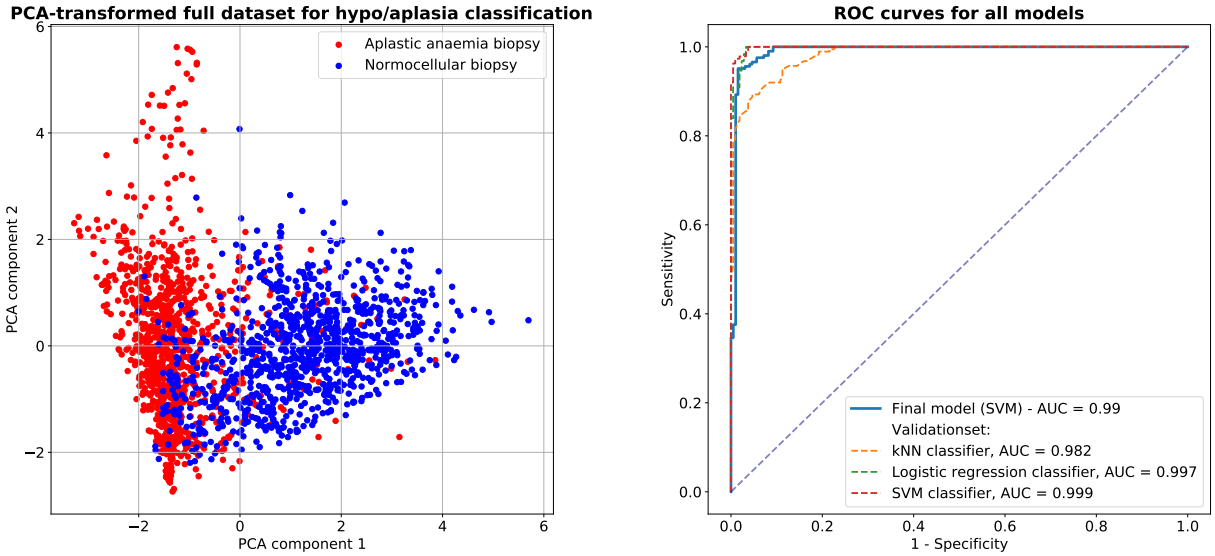

Figure 4: (a, left): The dataset for aplasia classification, when projected along the two principal components using principal component analysis, shows a clear linear separability. Intuitively, this separation makes sense: aplastic tissue has little to no myelopoietic cells and a large amount of fat cells. (b, right): The linear separability of the data explains why the AUC of all methods is very high. The drop between validation and test performance of the SVM is negligible.

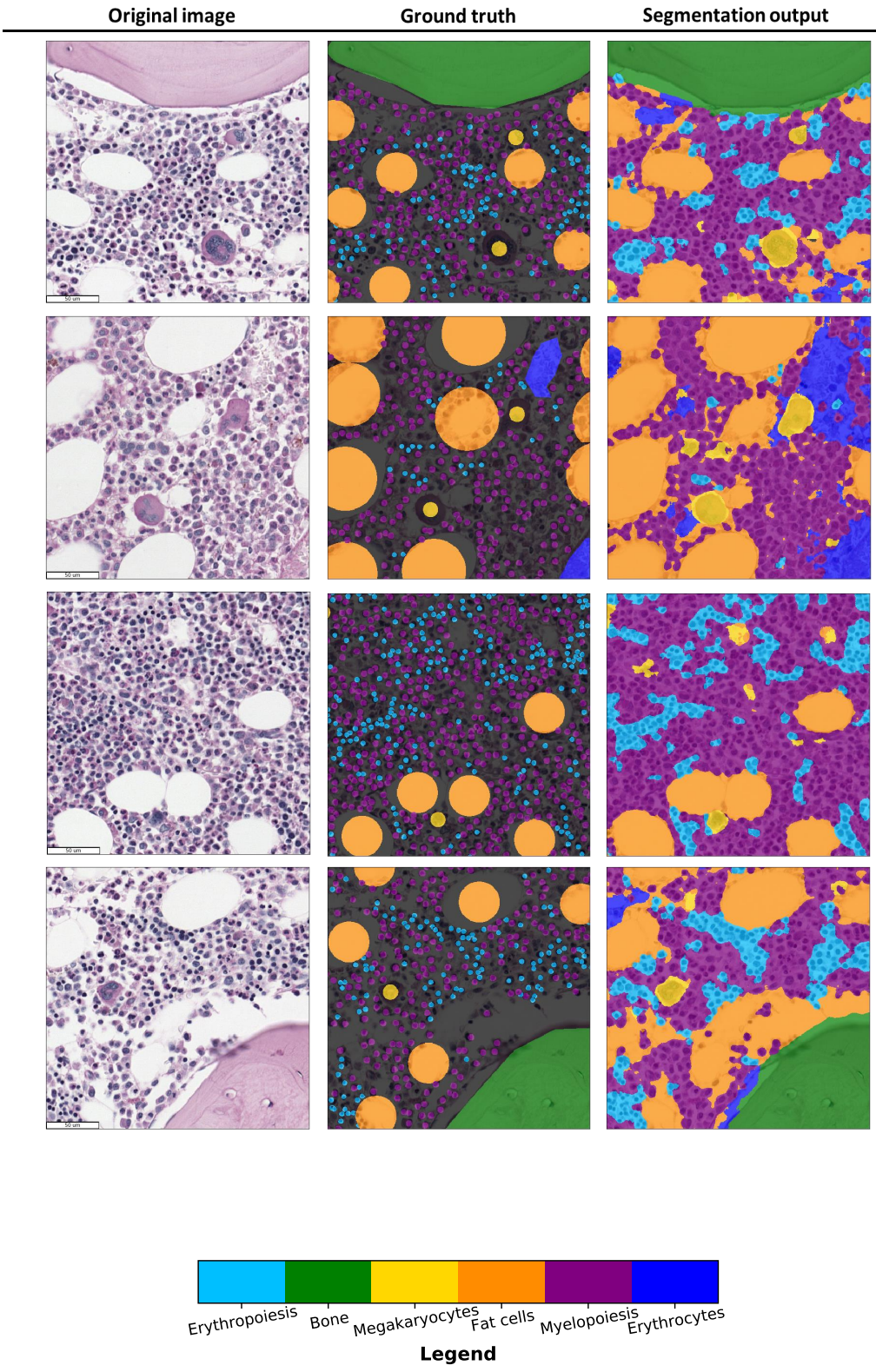

Figure 5: A selection of visual results of the segmentation output. Successes include the good detection of erythropoietic islands, while underperformance include intercellular space, which is often oversegmented as fat, as this was not present in the ground truth. The same goes for erythrocytes, which are typically located in blood vessels or tissue tears (caused due to biopsy procedure).

4. DISCUSSION

This work presented a neural network architecture meant for multi-class semantic segmentation in bone marrow trephine biopsies for the support of hematopathologists in identifying and locating various cell lineages.

The segmentation results demonstrate that the neural network is capable of correctly segmenting clinically relevant cell lineages in bone marrow trephine biopsies at least 70 percent of the time (Figure 3). Bone and megakaryocytes were segmented well (>90%) with no noticeable drop of performance going from the validation to the test set. However, a lower performance (>10%) occurs for all other cells. These discrepancies may be explained (1) by a strong variance in staining intensity across the test slides and no color augmentation to account for this, (2) accidental unrepresentative annotation of the training/validation set by the hematopathologist and (3) the transition from point annotation in the training/validation set to dense annotation in the test set, making the network perform poorly on cellular context.

The ROC curves of figure 4b and the 0.990 AUC of the final model indicate clearly that the segmentation output of the FCNN can be used to distinguish between normocellular and aplastic bone marrow tissue. The output of this model could be used to support hematopathologists in judging the presence of aplasia, typically present in aplastic anemia, thus reducing workload. The typical failure case of the classifier is on normocellular biopsies from which a sampled patch is uncharacteristically aplastic. These failure cases are ascribable to the patch-based nature of the evaluation; normally, a hematopathologist examining an entire biopsy would be able to see enough context to disregard local aplasia and the same can be expected if the classifier is applied on feature vectors sampled from biopsy-wide data.

Because the ratio between erythropoietic and myelopoietic cells is of clinical value for multiple hematopathologies,[1] future work may focus on classifying such pathologies using the same method as for aplastic bone marrow. The segmentation result could also be improved: the logical next step would be to train the neural network with intensive color augmentation to account for variety in natural cell appearance and stain variety across biopsies. Moreover, the segmentation output could be used to determine other clinically relevant measures or diagnose other pathologies that characterize themselves via changes in cell (surface) ratios. If instance segmentation is achieved, even more clinically relevant measures could be extracted from the network inference, e.g. cell morphology. Lastly, point annotations were extended to circles using a dataset-wide average diameter of the respective cells. However, this resulted in varying degrees of over/undersized annotations, especially for fat cells and megakaryocytes, cells which vary greatly in shape and size. Future work may also utilize more sophisticated ways of extending point annotations to full annotations.

To conclude, a fully convolutional neural network was successfully trained to segment six classes in bone marrow trephine biopsies, with an average detection rate of 83% as a result. Using the output of the neural network, it was found that a classifier can be trained to classify bone marrow biopsies as normocellular/aplastic. Future applications of the segmentation output may give other clinically relevant metrics, such as an estimate of cellularity, a metric normally subjectively estimated by eye. Supporting hematopathologists in this way can improve inter- and intra-observer variability and reduce workload.

AUTHOR CONTRIBUTIONS

L.v.E. performed the experiments, analyzed the results and wrote the manuscript. K.H. performed the data selection, annotated cells and provided advise on the clinical aspects of this study. H.P. and G.L. supervised the work and were involved in setting up the experimental design. All authors reviewed the manuscript and agree with its contents.

REFERENCES

[1] Bain, B. J., Clark, D. M., and Wilkins, B. S., [*Bone marrow pathology*], Wiley-Blackwell (2019).

[2] Morley, A. and Blake, J., "Observer error in histological assessment of marrow hypocellularity.," *Journal of clinical pathology* **28**(2), 104–108 (1975).

[3] Chen, T., McDonald, A., Shadbolt, B., and Talaulikar, D., "Precision of histological bone marrow staging in follicular lymphoma and diffuse large b-cell lymphoma," *Clinical and Investigative Medicine* **35**, E358–E364 (2012).

[4] Hodes, A., Calvo, K. R., Dulau, A., Maric, I., Sun, J., and Braylan, R., "The challenging task of enumerating blasts in the bone marrow," in [*Seminars in hematology*], **56**(1), 58–64, Elsevier (2019).

[5] Kim, Y., Kim, M., Kim, Y., Han, J., and Han, K., "Estimation of bone marrow cellularity using digital image nucleated cell counts in patients receiving chemotherapy," *International journal of laboratory hematology* **36**(5), 548–554 (2014).

[6] Wilkins, B. S., Erber, W. N., Bareford, D., Buck, G., Wheatley, K., East, C. L., Paul, B., Harrison, C. N., Green, A. R., and Campbell, P. J., "Bone marrow pathology in essential thrombocythemia: interobserver reliability and utility for identifying disease subtypes," *Blood* **111**(1), 60–70 (2008).

[7] Al-Janabi, S., Huisman, A., and Van Diest, P. J., "Digital pathology: current status and future perspectives," *Histopathology* **61**(1), 1–9 (2012).

[8] Long, J., Shelhamer, E., and Darrell, T., "Fully convolutional networks for semantic segmentation," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 3431–3440 (2015).

[9] Song, T.-H., Sanchez, V., ElDaly, H., and Rajpoot, N., "Simultaneous cell detection and classification in bone marrow histology images," *IEEE journal of biomedical and health informatics* **23** (2018).

[10] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).