Computer Methods and Programs in Biomedicine 243 (2024) 107924

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Evaluation of deep learning training strategies for the classification of bone marrow cell images

Stefan Glüge [a,*], Stefan Balabanov [b], Viktor Hendrik Koelzer [c], Thomas Ott [a]

[a] Institute of Computational Life Sciences, Zurich University of Applied Sciences, Schloss 1, 8820 Wädenswil, Switzerland
[b] Department of Medical Oncology and Haematology, University Hospital Zurich and University of Zurich, Rämistrasse 100, 8091 Zurich, Switzerland
[c] Department of Pathology and Molecular Pathology, University Hospital Zurich and University of Zurich, Schmelzbergstrasse 12, 8091 Zurich, Switzerland

ARTICLE INFO

MSC:
68T45
68T05
68U10

Keywords:
Hematopoiesis
In-domain pre-training
Deep learning
Hematopathology

ABSTRACT

Background and Objective: The classification of bone marrow (BM) cells by light microscopy is an important cornerstone of hematological diagnosis, performed thousands of times a day by highly trained specialists in laboratories worldwide. As the manual evaluation of blood or BM smears is very time-consuming and prone to inter-observer variation, new reliable automated systems are needed.
Methods: We aim to improve the automatic classification performance of hematological cell types. Therefore, we evaluate four state-of-the-art Convolutional Neural Network (CNN) architectures on a dataset of 171,374 microscopic cytological single-cell images obtained from BM smears from 945 patients diagnosed with a variety of hematological diseases. We further evaluate the effect of an in-domain vs. out-of-domain pre-training, and assess whether class activation maps provide human-interpretable explanations for the models' predictions.
Results: The best performing pre-trained model (Regnet_y_32gf) yields a mean precision, recall, and F1 scores of 0.787 ± 0.060, 0.755 ± 0.061, and 0.762 ± 0.050, respectively. This is a 53.5% improvement in precision and 7.3% improvement in recall over previous results with CNNs (ResNeXt-50) that were trained from scratch. The out-of-domain pre-training apparently yields general feature extractors/filters that apply very well to the BM cell classification use case. The class activation maps on cell types with characteristic morphological features were found to be consistent with the explanations of a human domain expert. For example, the Auer rods in the cytoplasm were the predictive cellular feature for correctly classified images of faggot cells.
Conclusions: Our study provides data that can help hematology laboratories to choose the optimal training strategy for blood cell classification deep learning models to improve computer-assisted blood and bone marrow cell identification. It also highlights the need for more specific training data, i.e. images of difficult-to-classify classes, including cells labeled with disease information.

1. Introduction

The examination of cell morphology in BM and peripheral blood (PB) is the basis for the diagnosis of malignant and non-malignant hematologic diseases [30,25]. Due to its technical feasibility and established clinical value for disease classification, BM and PB cytology is an essential part of the diagnosis of hematological diseases [46]. Traditionally, classification of cell morphology is performed manually by human experts using light microscopy. In addition to being tedious and time-consuming, manual inspection and classification of cells suffers from subjectivity and low sensitivity [13].

The use of digital microscopy and machine learning to classify cells in PB and BM has great potential to achieve more accurate and stable results, while minimizing the need for human intervention (time savings) and has great potential to reduce classification errors by providing an unbiased second opinion.

A large dataset with ground truth labels is the fundamental requirement for a successful application of deep and complex CNN architectures. Therefore, machine learning is usually applied in domains where such data are available, such as magnetic resonance imaging [29]. One way to use these models in domains with limited amounts of data is transfer learning [45,42], which has also been successfully applied to related tasks, such as digital holography [11,6].

* Corresponding author.
E-mail address: stefan.gluege@zhaw.ch (S. Glüge).

https://doi.org/10.1016/j.cmpb.2023.107924
Received 11 May 2023; Received in revised form 28 September 2023; Accepted 6 November 2023
Available online 13 November 2023

While the fields of histopathology and cytopathology are related, the single-cell nature of BM datasets introduces a relevant domain shift that does not allow for easy methodological transfer.

The main contributions of the current manuscript are as follows:

- The evaluation of four common CNN architectures on the BM cell classification problem, which achieved the best top 1/top 5 accuracy on ImageNet [40].
- Establish a benchmark for the BM cell classification problem, since both the models and the data [31] are openly available.
- Investigate the effect of different pre-training strategies, i.e., in-domain vs. out-of-domain, providing a systematic approach to achieve state-of-the-art performance across a wide range of cell types.
- Evaluate whether class activation maps of model predictions provide interpretable explanations to domain experts.

1.1. Related work

The first attempts to classify BM cells were based on the extraction of handcrafted single-cell features and the application of standard classifiers, such as support vector machines, random forests [22], and hierarchical decision trees [23]. Later, deep learning approaches, namely deep CNNs, were investigated, but only on small sample sizes or disease classes [1,3].

Matek et al. [32] presented two CNN-based classifiers for single-cell images of BM leukocytes. The best results were obtained with a ResNeXt-50 model [51] trained from scratch. Along with their approach, and perhaps more importantly, they published a large dataset of expert-annotated single-cell images [31] (cf. Sec. 2.1.1). This great resource can now be used by the community to advance the field.

Mori et al. [33] introduced the use of a pre-trained ResNet-152 in the classification of bone marrow dysplasia. Their system was evaluated on a rather small dataset (1,797 images labeled by 4 degrees of dysplasia). The reported sensitivity, specificity, and accuracy were 85.2%, 98.9%, and 98.2%, respectively.

Dehaene et al. [10] showed the positive effect of an in-domain pre-training in the weakly supervised learning scenario of WSIs classification in histopathology: An in-domain feature extractor pre-trained on histology images outperformed a frozen feature extractor pre-trained on ImageNet [40]. Furthermore, the learned embedding space was shown to exhibit biologically meaningful separation of tissue structures.

Boldú et al. [7] created a dataset from blood smears containing 16,450 single-cell images from 100 healthy patients, 191 patients with viral infections, and 148 patients with acute leukemia. VGG16, ResNet101, DenseNet121 and SENet154 were evaluated on the problem of acute leukemia classification. All CNNs were pre-trained on ImageNet and fine-tuned to cell images. They report an accuracy of $86.9\% \pm 0.68$ for VGG16 on the 6-class cell classification task.

Some research has specifically addressed the problem of large class imbalance in cell datasets. Guo et al. [17] present a class balance classification method for classifying 15 types of BM cells on a dataset of 7484 images with an imbalance ratio of 31 : 1 (3097 lymphocytes, 98 platelets). They achieved precision, sensitivity, and specificity values of 84.53%, 84.44% and 99.29%, respectively. Hazra et al. [18] addressed the problem of underrepresented classes by using a Generative Adversarial Network (GAN) to generate synthetic data and balance their dataset. After this data augmentation, their classification CNN achieved accuracy, specificity, and sensitivity greater than 95%.

Recently, Wang et al. [49] constructed a remarkable large dataset of 131,300 expert-annotated single cell images. They report an overall accuracy on the cell classification task of 89.53%. Furthermore, they applied their Multi-Level Feature Learning Network (MLFL-Net) model to the prediction of leukemia types of hematological diseases. It produced the same diagnostic prediction as the experts for 74 out of the cohort of 80 patients (92.5%).

Table 1
Overview of the dataset used in our study. Given that the BM cell images are the target domain, a pre-training on cervical cells or WSI patches is considered to be in-domain, whereas a pre-training on ImageNet is considered to be out-of-domain.

Dataset	#Images	#Classes	Resolution	Domain
Bone marrow cells [32]	171,374	21	250×250	single cell
Comparison Detector [28]	48,587	11	variable	single cell
PatchCamelyon [47]	262,144	2	96×96	WSI patch
ImageNet [40]	1,281,167	1,000	variable	natural scene/object

Table 2
Color channel: Mean and standard deviation for each dataset.

Dataset	Red	Green	Blue
Bone marrow cells	0.5630 ± 0.2421	0.4959 ± 0.2835	0.7353 ± 0.1767
Comparison Detector	0.7255 ± 0.2705	0.7826 ± 0.2380	0.8270 ± 0.1834
PatchCamelyon	0.7008 ± 0.2350	0.5384 ± 0.2774	0.6916 ± 0.2129
ImageNet	0.485 ± 0.229	0.456 ± 0.224	0.406 ± 0.225

2. Methods

2.1. Datasets

In this section, we present the datasets used in our study. The BM cell dataset is our target domain, while different datasets were used to initialize the models. Table 1 gives an overview of the number of images and classes for each dataset. We also show the original image resolution and domain of the images. All datasets provide the images in standard RGB format.

Additionally, we show the mean and standard deviation of the color channels for each dataset in Table 2. These values were used to normalize the images during model training (cf. Sec. 2.3).

2.1.1. Bone marrow cell dataset

Matek et al. [32] published a dataset of 171,374 expert-annotated single BM cell images from 945 patients diagnosed with a variety of hematologic diseases [31].

Diagnostically relevant cell images (250×250-pixel) were annotated into 21 classes. Fig. 1 shows four randomly selected samples from the dataset. The number of images per class varies widely from 8 up to $\approx 30,000$ and is listed in Table 5, column #Images.

2.1.2. Comparison detector (CD) dataset

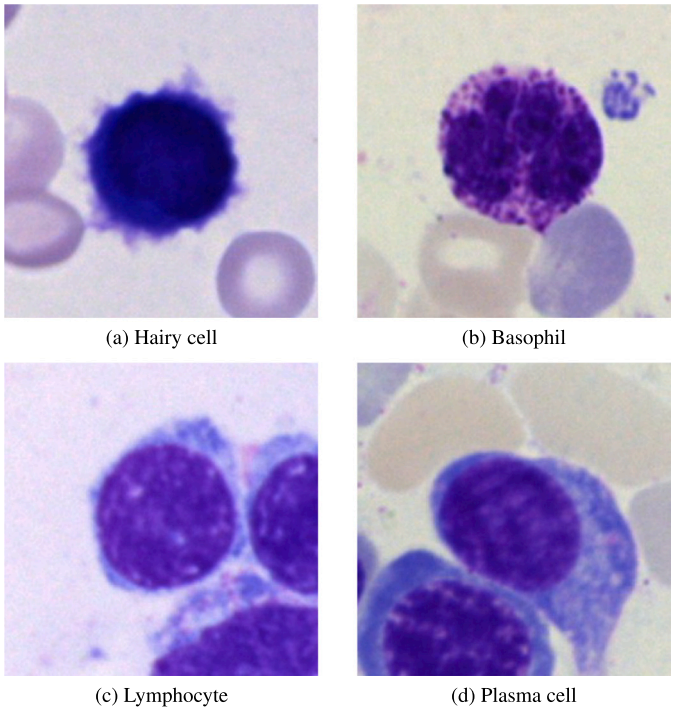
Liang et al. [28] established a dataset consisting of 7,410 cervical microscopical images cropped from WSIs.[1] A total of 48,587 object instance bounding boxes were labeled by experienced pathologists. Each instance belongs to one of 11 categories. As for the BM cell dataset, the number of images per class varies widely between 123 up to $\approx 26,000$. Fig. 2 shows four randomly selected samples from the dataset.

2.1.3. PatchCamelyon (PCam) dataset

Veeling et al. [47] presented the PatchCamelyon dataset. It consists of 327,680 color images (96×96 pixels) extracted from histopathological scans of lymph node sections. Each image is annotated with a binary label indicating the presence or absence of metastatic carcinoma. In total, the dataset consists of 262,144 images for training, and 32.768 for validation and testing. Fig. 3[2] shows some randomly selected examples from the dataset.

[1] The dataset is available at https://github.com/kuku-sichuan/ComparisonDetector.

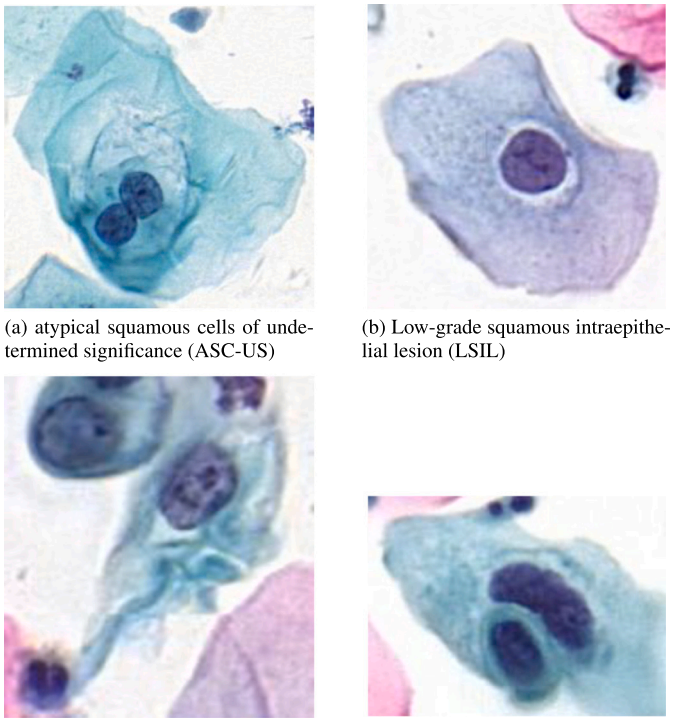
[2] Bas Veeling, example images from PCam, MIT License, available from https://github.com/basveeling/pcam (accessed August 31, 2022).



(a) Hairy cell (b) Basophil

(c) Lymphocyte (d) Plasma cell

Fig. 1. Example of four images from the bone marrow cell dataset with their corresponding class label.



(a) atypical squamous cells of undetermined significance (ASC-US) (b) Low-grade squamous intraepithelial lesion (LSIL)

(c) High-grade squamous intraepithelial lesion (HSIL) (d) Squamous-cell carcinoma (SCC)

Fig. 2. Example of four images of cervical cells from the comparison detector dataset with their corresponding class label.

2.1.4. ImageNet

Since 2010, the ImageNet dataset has been used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [40]. The classification part of the dataset contains 1,000 categories of 1.2 million images (aka. ImageNet-1K). For image classification, ImageNet has provided a solid foundation for benchmarking advances in computer vision research. It serves as the primary dataset for pre-training for computer-vision transfer learning models. In addition, improving performance on ImageNet is often considered as a litmus test for general applicability to downstream tasks [38].

Table 3
CNN models used in our study. We show the top 1 and top 5 accuracy on ImageNet (Acc@1/Acc@5) and the number of trainable parameters (#Params).

Model	Acc@1	Acc@5	#Params
VGG-19 BN [43]	74.218	91.842	143,678,248
ResNet-152 [19]	82.284	96.002	60,192,808
Regnet_y_32gf [36]	83.368	96.498	145,046,770
ViT_l_32 [12]	76.972	93.07	306,535,400

2.2. Models and model training

We tested the following model architectures for image classification: VGG [43] with batch normalization (BN), ResNet [19], RegNet [36] and VisionTransformer (ViT) [12]. The models are provided in PyTorch [34].[3] We chose the model configurations that gave the best top 1/top 5 accuracy on ImageNet [40]. Table 3 provides an overview of the model configurations used in our study. A more detailed introduction to the different architectures is further provided in the Supplementary Material, Sec. 1.2 Model architectures.

To adapt the models for the BM cell classification task, we removed the last fully connected layer and used 21 linear units. Model training was performed for 75 epochs with a batch size of 32. PyTorch's implementation of stochastic gradient descent optimization [8] was used with a fixed momentum of 0.9. We also applied a learning rate decay by a factor of 0.1 if the validation loss did not improve within the last 3 epochs of training. The models were evaluated on the validation set after each epoch, and the models with the highest validation accuracy were evaluated on the held-out test data.

To find the most promising initial learning rate for each model, we used the PyTorch implementation of the learning rate range test[4] detailed in [44]. We did not optimize other hyperparameters, such as batch size and optimizer, because we are mainly interested in comparing model architectures and different pre-training strategies.

2.3. Data preparation and augmentation

For the network training, we used a stratified 5-fold train-validation-test split. In each split, we trained a network using 80% and 20% of the available images for each class for training and testing, respectively. Repeating the stratified split five times ensures that each image was in the test set once in each experiment. Within the training set, 20% of the samples were used as a validation set during training.

The images were resized to 224×224 pixels and normalized to mean $\pm$ standard deviation of the channels of the full dataset (cf. Table 2). In addition, the following augmentation functions were used during fine-tuning on the BM cell classification task and during in-domain pre-training:

- Random cropping of the image with a random size between 0.08 and 1 of the original image size, and a random aspect ratio of the crop between 0.75 and 1.33.[5]
- Random rotation of the image between 0° and 180°.
- Random horizontal flipping of the image with probability 0.5
- Random vertical flipping of the image with probability 0.5

[3] https://pytorch.org/vision/stable/models.html version: 0.13.
[4] https://pypi.org/project/torch-lr-finder/.
[5] not used during in-domain pre-training cf. Sec. 2.4.2.

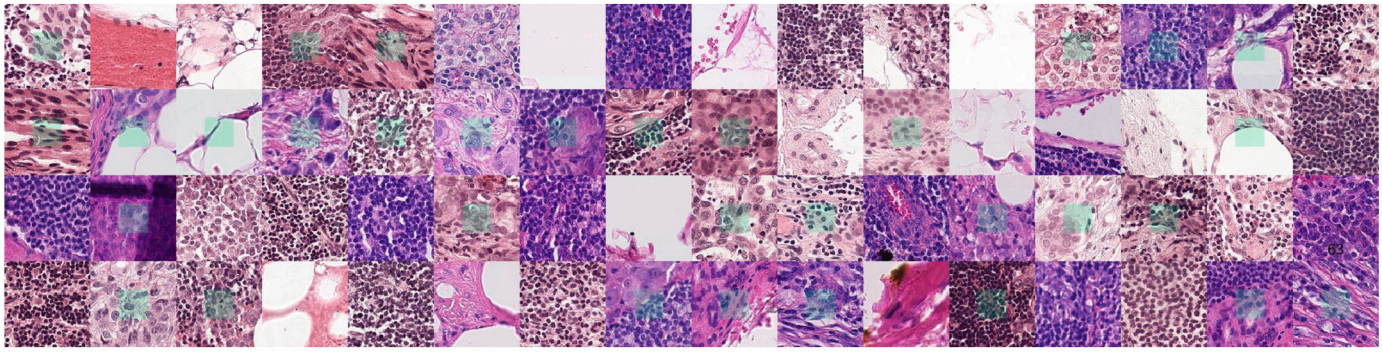


Fig. 3. Example images from PCam. Green boxes indicate tumor tissue in the center region, corresponding to a positive label.

To compensate for the strong class imbalance, we used a weighted random sampler during training, which ensures that the network sees the same number of (augmented) samples for all classes.

For the out-of-domain pre-training, we used pre-trained models available in the torchvision library. Additional steps for data preparation and augmentation were not performed (cf. Sec. 2.4.1).

2.4. Pre-training

Pre-training has long been used to improve performance in visual tasks [16]. The features learned by a CNN trained on a large dataset such as ImageNet tend to transfer well to other domains.

We hypothesize that a domain-specific pre-training might help in the development of features that facilitate separability between cell classes, rather than using out-of-domain examples such as provided by ImageNet. Therefore, we set up several experiments with different pre-training strategies, namely

1. no pre-training / random initialization,
2. out-of-domain pre-training,
3. in-domain pre-training,
4. out-of-domain + in-domain pre-training.

2.4.1. Out-of-domain pre-training

Pre-training on ImageNet is considered out-of-domain for the BM cell classification task. Ridnik et al. [38] Therefore, we used the pre-trained models that are available through torchvision. We refer to the torchvision page https://pytorch.org/vision/0.12/models.html for more details.

2.4.2. In-domain pre-training

We consider the CD (cf. Sec. 2.1.2) and PCam (cf. Sec. 2.1.3) datasets to be in the same domain as the BM cell images. The pre-training was performed on the randomly initialized models with the same training parameters as the later fine-tuning (cf. Sec. 2.2). Data preparation was performed as in the fine-tuning phase (cf. Sec. 2.3) with the following adjustments: normalization to the dataset-specific color channel values (cf. Table 2) and without random image cropping.

2.4.3. Out-of-domain + in-domain pre-training

In this scenario, we combined out-of-domain pre-training on ImageNet followed by an in-domain pre-training on CD and PCam, respectively.

2.5. Gradient-weighted class activation mapping

To build confidence in the classification results of deep CNNs, it is essential to provide some human interpretable explanations for the models' predictions.

We used the PyTorch implementation [15] of Gradient-weighted Class Activation Mapping (Grad-CAM) [41] to address this problem. This technique uses the gradients of any target concept, such as 'faggot cell', that flow into the final convolutional layer to produce a coarse localization map (heatmap) that highlights the important regions in the image for predicting the concept. For a more detailed explanation see Sec. 1.5 Gradient-weighted Class Activation Mapping in the Supplementary Material.

While these visualizations are often referred to as 'visual explanation', expert interpretation remains critical to actually *explain* the decision, or at least to judge whether the decision is reasonable. This is what makes post hoc explanations problematic, as argued, for example by Rudin [39].

3. Results

Table 4 shows the mean precision, recall and F1 scores that were obtained in the 5-fold cross-validation of the different models under different pre-training conditions. We chose these specific scores to ensure the comparability of our results with the numbers reported in Matek et al. [32]. A more detailed definition of the scores and the individual steps on how we computed the numbers in Table 4 are given in the Supplementary Material Sec. 1.1 Evaluation metrics.

Table 5 shows the class-wise scores for the Regnet_y_32gf pre-trained on ImageNet + CD side by side with the scores reported in Matek et al. [32]. Note that we calculated the F1 score for the ResNeXt-50 from the published precision and recall means without the corresponding standard deviations.

To provide some insight into which cell types are more difficult to differentiate by the trained models, we show the confusion matrix of the Regnet_y_32gf pre-trained on ImageNet + CD on the test data (mean of 5-fold cross-validation) in Fig. 4. The lowest class-wise accuracy (0.3 - 0.6) is observed for Abnormal Eosinophils (ABE, 8 samples), Immature Lymphocytes (LYI, 65 samples), Faggot Cells (FGC, 47 samples), Basophils (BAS, 441) and Metamyelocytes (MMZ, 3,055 samples).

To understand the network decision-making process for cell classification, we performed a Grad-CAM analysis. Examples of correct and incorrect classifications and the corresponding heatmaps of selected cell classes are shown in Fig. 5 a-h and Fig. 6 a-l. We chose to use images from of cell types with characteristic morphological features. FGCs with multiple bundles of Auer rods in the cytoplasm are a characteristic example for correct classification (e.g., Fig. 5a). According to the activation maps, these Auer rods were the most predictive cellular feature in correctly classified images for this particular cell type (Fig. 5a-d). This was particularly true, when the cells had a wide and bright cytoplasm, allowing for clear recognition of the Auer rods (Fig. 5a-c). In misclassified images, the activation maps highlighted areas in the cytoplasm (Fig. 5e-f) or in the nucleus (Fig. 5g-h), which showed the artifactual formation of rod-like textures.

The most prominent morphologic feature of BAS is the purple granules in the cytoplasm (e.g., Fig. 6a-d), which allows differentiation from other granulocyte populations. For the BAS test set ($n = 88$), 48 images were classified correctly and 40 images were misclassified by the

Table 4
Mean ± standard deviation of Precision/Recall/F1 scores obtained in the 5-fold cross-validation of the different models on the BM classification task for different pre-training strategies. The best scores are highlighted. For comparison, we also show the results published in [32]. The pre-training strategies are denoted as follows: *random* – no pre-training, the models are initialized with random weights, *ImageNet* – out-of-domain pre-training, the models are initialized with a pre-training on ImageNet, *PCam / CD* – in-domain pre-training, the models are initialized with a pre-training on PCam/CD dataset, *ImageNet + PCam/CD* – out-of-domain + in-domain pre-training, the models are initialized with a pre-training on ImageNet followed by a pre-training on PCam/CD dataset.

Model & pre-training	Precision	Recall	F1 Score
VGG-19 BN			
Random	0.667 ± 0.039	0.744 ± 0.058	0.695 ± 0.038
ImageNet	0.705 ± 0.037	0.748 ± 0.038	0.720 ± 0.028
PCam	0.682 ± 0.052	0.763 ± 0.056	0.712 ± 0.047
CD	0.648 ± 0.045	0.782 ± 0.065	0.691 ± 0.045
ImageNet + PCam	0.722 ± 0.057	0.772 ± 0.054	0.742 ± 0.049
ImageNet + CD	0.701 ± 0.062	0.751 ± 0.061	0.720 ± 0.054
ResNet-152			
Random	0.670 ± 0.050	0.731 ± 0.060	0.689 ± 0.044
ImageNet	0.732 ± 0.039	0.745 ± 0.049	0.733 ± 0.034
PCam	0.656 ± 0.069	0.738 ± 0.075	0.683 ± 0.065
CD	0.672 ± 0.048	0.735 ± 0.062	0.695 ± 0.046
ImageNet + PCam	0.739 ± 0.061	0.757 ± 0.046	0.744 ± 0.044
ImageNet + CD	0.734 ± 0.036	0.740 ± 0.023	0.730 ± 0.022
Regnet_y_32gf			
Random	0.709 ± 0.040	0.698 ± 0.038	0.695 ± 0.025
ImageNet	0.770 ± 0.030	0.731 ± 0.040	0.740 ± 0.030
PCam	0.712 ± 0.053	0.709 ± 0.058	0.705 ± 0.049
CD	0.707 ± 0.038	0.698 ± 0.032	0.697 ± 0.028
ImageNet + PCam	0.784 ± 0.063	0.735 ± 0.062	0.748 ± 0.053
ImageNet + CD	0.787 ± 0.060	0.755 ± 0.061	0.762 ± 0.050
ViT_l_32			
Random	0.538 ± 0.028	0.576 ± 0.037	0.547 ± 0.024
ImageNet	0.769 ± 0.056	0.687 ± 0.058	0.712 ± 0.049
PCam	0.539 ± 0.039	0.584 ± 0.055	0.550 ± 0.038
CD	0.552 ± 0.049	0.588 ± 0.059	0.561 ± 0.045
ImageNet + PCam	0.743 ± 0.069	0.734 ± 0.058	0.732 ± 0.050
ImageNet + CD	0.762 ± 0.061	0.701 ± 0.053	0.722 ± 0.048
ResNeXt-50 [32]			
Random	0.510 ± 0.048	0.689 ± 0.087	0.545

Regnet_y_32gf pre-trained on ImageNet + CD. In the case of correct classification, manual evaluation of the Grad-CAM action maps showed that the model indeed focused on the characteristic cytoplasmic granules of BAS (Fig. 6a-d). The model misclassified images when the focus was on the nucleus or when the granules were less prominent, as shown in Fig. 6e-h. This suggests that the granular chromatin structures may have been mistaken for cellular granules. Furthermore, manual inspection of the misclassified BAS images revealed an incorrect ground truth as the reason for misclassification in some cases (Fig. 6i-l). Interestingly, our model classified the images shown in Fig. 6i-l to the correct cell type, compensating for this error in the ground truth of the dataset.

4. Discussion

In this study, we evaluated different deep learning training strategies for the classification of BM cell images. First, we compared different model architectures that achieve state-of-the-art performance on ImageNet. Overall, depending on the evaluation score, different models can be considered as the best performs.

Without any pre-training, the VGG-19 BN, ResNet-152 and Regnet_y_32gf outperform the previously published results of the ResNeXt-50, while the ViT_l_32 architecture does not (cf. Table 4 rows "random"). ViT structurally lacks locality inductive bias and requires a large amount of training data to obtain an acceptable visual representation. Therefore, learning on a small dataset requires pre-training on a large dataset, which may limit its applicability to our current use case [26]. However, larger models in particular could benefit from an out-of-domain pre-training on the larger ImageNet-21K dataset, as shown in [38].

For precision, a good measure when the cost of false positives is high, the Regnet_y_32gf pre-trained on ImageNet + CD performed best. In terms of recall, typically used when the cost of false negatives is high, the VGG-19 BN pre-trained on CD performed best. All pre-trained models tested outperformed their randomly initialized counterparts.

In general, we observed an advantage for out-of-domain pre-training, i.e., ImageNet vs. PCam and CD, respectively. These results suggest that pre-training on a large out-of-domain dataset yields better features to separate BM cell images compared to pre-training on a smaller, but domain-specific, dataset (cf. Table 4). This could be due to the specific properties of the ImageNet training examples, including many center-clipped objects, which show some similarity to the task of classifying single cells, rather than patch-based histopathology domain-specific features, which represent a more limited representation of textures. Notably, the combination of ImageNet + CD/PCam tended to yield a slight improvement over ImageNet pre-training alone. However, the performance differences are within the range of the standard deviation of the 5-fold cross-validation. This suggests that the features

Table 5
Classwise precision, recall and F1 score obtained in the 5-fold cross-validation for the Regnet_y_32gf pretrained on ImageNet + CD compared to the published results of a ResNeXt-50 architecture [32]. Additionally, we show the number of samples (#Images) for each class in the data set.

		ResNeXt50 random [32]			Regnet_y_32gf ImageNet + CD			
Class		Precision	Recall	F1 Score	Precision	Recall	F1 Score	#Images
Band neutrophils	(NGB)	0.540 ± 0.030	0.650 ± 0.040	0.590	0.717 ± 0.012	0.790 ± 0.008	0.752 ± 0.008	9,968
Segmented neutrophils	(NGS)	0.920 ± 0.020	0.710 ± 0.050	0.801	0.938 ± 0.003	0.897 ± 0.010	0.917 ± 0.004	29,424
Lymphocytes	(LYT)	0.900 ± 0.030	0.700 ± 0.030	0.788	0.922 ± 0.004	0.909 ± 0.008	0.915 ± 0.005	26,242
Monocytes	(MON)	0.570 ± 0.050	0.700 ± 0.030	0.628	0.731 ± 0.015	0.790 ± 0.024	0.759 ± 0.013	
Eosinophils	(EOS)	0.850 ± 0.050	0.910 ± 0.030	0.879	0.958 ± 0.007	0.974 ± 0.007	0.966 ± 0.006	5,883
Basophils	(BAS)	0.140 ± 0.050	0.640 ± 0.070	0.230	0.763 ± 0.067	0.618 ± 0.041	0.682 ± 0.044	441
Metamyelocytes	(MMZ)	0.300 ± 0.050	0.640 ± 0.080	0.409	0.551 ± 0.013	0.579 ± 0.036	0.564 ± 0.015	3,055
Myelocytes	(MYB)	0.520 ± 0.050	0.590 ± 0.060	0.553	0.703 ± 0.013	0.757 ± 0.012	0.729 ± 0.008	6557
Promyelocytes	(PMO)	0.760 ± 0.050	0.720 ± 0.080	0.739	0.873 ± 0.012	0.814 ± 0.010	0.842 ± 0.009	11,994
Blasts	(BLA)	0.750 ± 0.030	0.650 ± 0.030	0.696	0.843 ± 0.010	0.872 ± 0.008	0.857 ± 0.008	11,973
Plasma cells	(PLM)	0.810 ± 0.060	0.840 ± 0.040	0.825	0.918 ± 0.015	0.936 ± 0.008	0.927 ± 0.006	7,629
Smudge cells	(KSC)	0.280 ± 0.090	0.900 ± 0.100	0.427	0.893 ± 0.106	0.875 ± 0.125	0.874 ± 0.044	42
Other cells	(OTH)	0.220 ± 0.060	0.840 ± 0.060	0.349	0.946 ± 0.017	0.827 ± 0.030	0.882 ± 0.023	294
Artefacts	(ART)	0.820 ± 0.050	0.740 ± 0.060	0.778	0.902 ± 0.006	0.897 ± 0.007	0.900 ± 0.003	19,630
Not identifiable	(NIF)	0.270 ± 0.040	0.630 ± 0.040	0.378	0.628 ± 0.019	0.662 ± 0.019	0.644 ± 0.014	3,538
Proerythroblasts	(PEB)	0.570 ± 0.090	0.630 ± 0.130	0.599	0.707 ± 0.025	0.825 ± 0.040	0.761 ± 0.010	2,740
Erythroblasts	(EBO)	0.880 ± 0.010	0.820 ± 0.010	0.849	0.957 ± 0.004	0.936 ± 0.004	0.946 ± 0.001	27,395
Hairy cells	(HAC)	0.350 ± 0.080	0.800 ± 0.060	0.487	0.804 ± 0.034	0.783 ± 0.081	0.790 ± 0.033	409
Abnormal eosinophils	(ABE)	0.020 ± 0.030	0.200 ± 0.400	0.036	0.400 ± 0.548	0.400 ± 0.548	0.400 ± 0.548	8
Immature lymphocytes	(LYI)	0.080 ± 0.030	0.530 ± 0.150	0.139	0.710 ± 0.228	0.292 ± 0.167	0.383 ± 0.185	65
Faggot cells	(FGC)	0.170 ± 0.050	0.630 ± 0.270	0.268	0.655 ± 0.112	0.422 ± 0.093	0.503 ± 0.061	47
mean		0.510 ± 0.048	0.689 ± 0.087	0.545	0.787 ± 0.060	0.755 ± 0.061	0.762 ± 0.050	

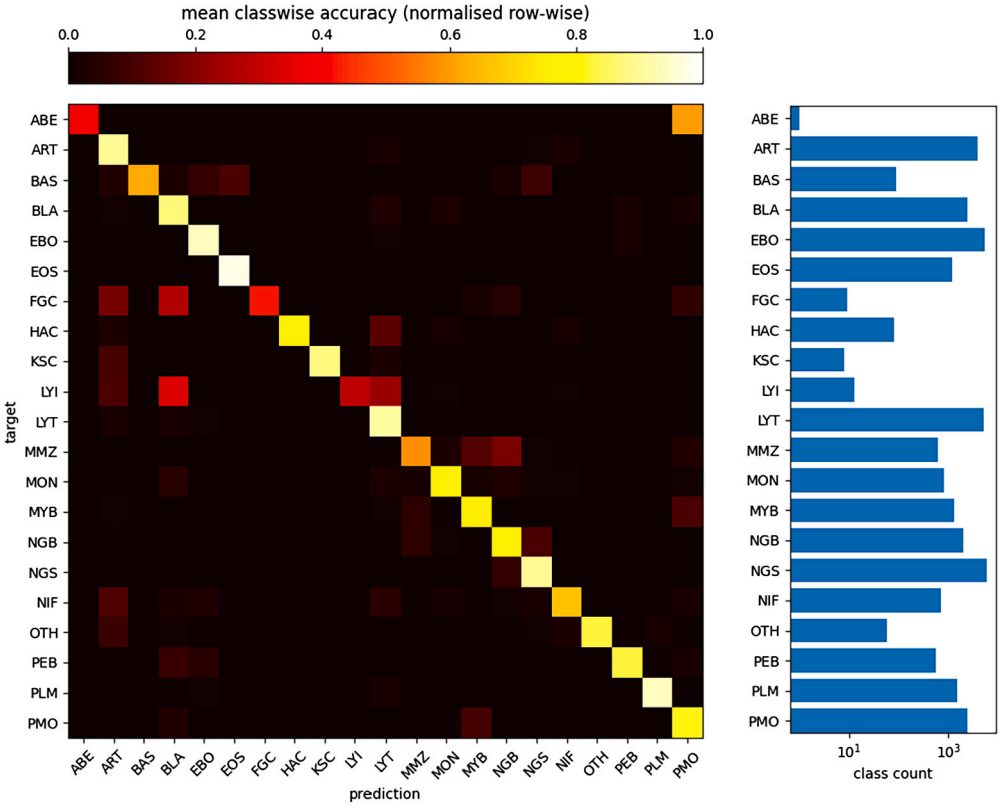


Fig. 4. Confusion matrix on the test set of the Regnet_y_32gf pre-trained on ImageNet + CD. Shown are classwise accuracies as the mean of the 5-fold cross-validation normalized by row to account for class imbalance. The number of single-cell images included in each category is indicated in the logarithmic plot on the right.

learned from ImageNet are sufficiently general for the BM cell classification task. Further fine-tuning on more, domain-specific microscopy images did not lead to better features for the final task.

Compared to the ResNeXt-50 trained from scratch, we obtained a 54.3% improvement in precision (Regnet_y_32gf (ImageNet + CD): 0.787 vs. ResNeXt-50: 0.51) and 9.6% improvement in recall (Regnet_y_32gf (ImageNet + CD): 0.755 vs. ResNeXt-50: 0.689).

Domain expert interpretation provides an explanation for these findings, as these cell types share relevant morphological similarities that can be difficult for human experts to resolve. In the future, cross-

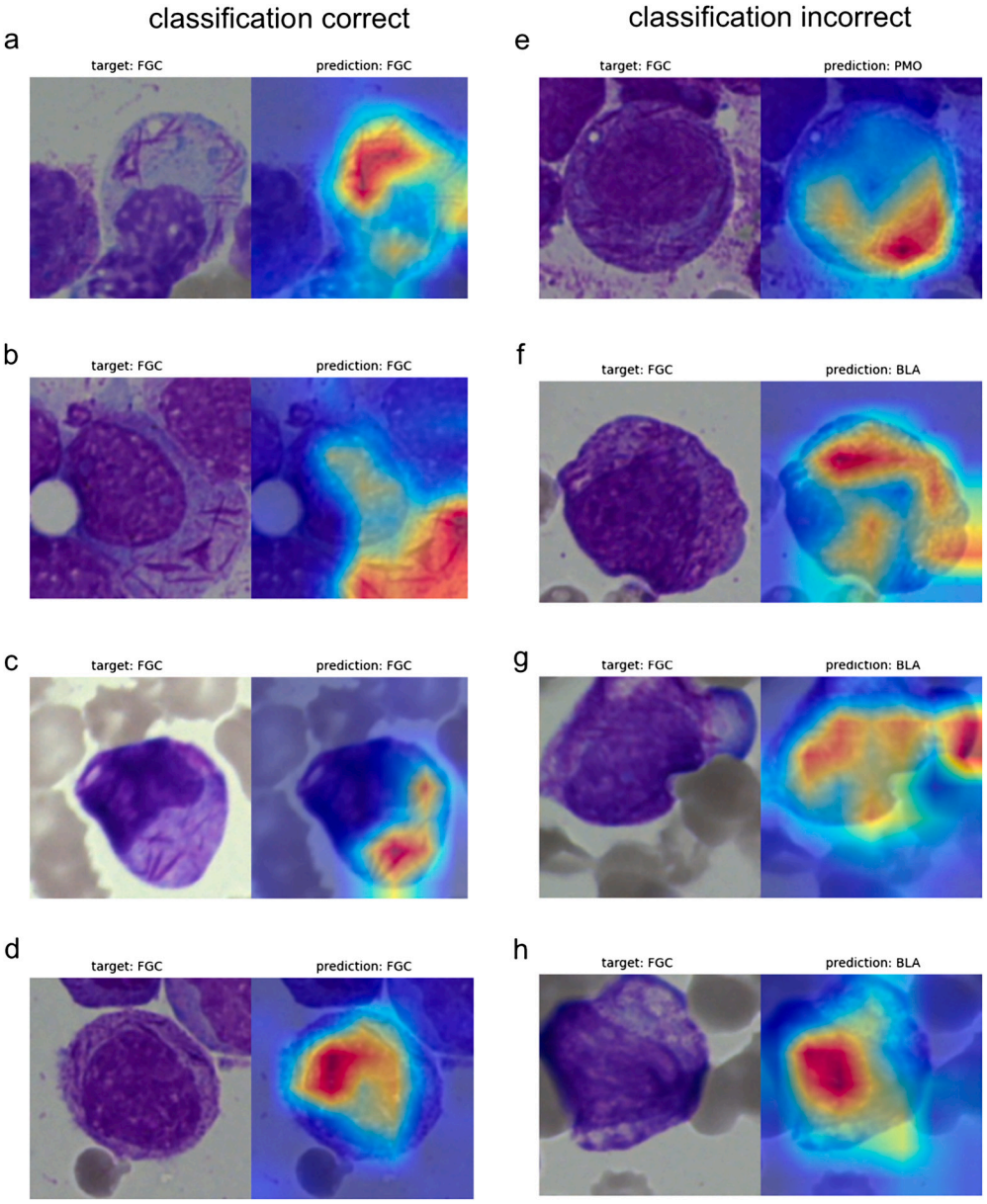


Fig. 5. Grad-CAM activation maps generated from the Faggot cells using the Regnet_y_32gf pre-trained on ImageNet + CD. Example images with corresponding activation maps for correct classified images (a-d) and misclassified images (e-h). Regions showing high activation (in red) provide a strong contribution to the classification result.

modality explanation maps may help to generate even better explanations that are potentially more understandable to human experts [5].

More expert-annotated training data is needed to improve this challenging classification task. Since different cell types are characterized by specific cytoplasmic or nuclear features, feature preselection by cell segmentation in the cytoplasm and nucleus could be another suitable approach to increase correct cell classification [2,35]. In particular, this could be an approach for the correct detection of FGCs, which are characterized by cytoplasmic Auer rods. Correct classification of FGCs is of clinical importance and misclassification, especially false negative classification, has direct negative clinical consequences. FGCs are a morphological hallmark of a very rare subtype of acute leukemia (acute promyelocytic leukemia (APL)), which can be cured in most patients after correct diagnosis [9]. However, APL is associated with severe bleeding complications and early death due to bleeding events if diagnosis and treatment are delayed [37]. In this context, a combination of digital microscopy and automated blood cell detection could lead to earlier diagnosis of APL patients and a reduction in early mortality in these patients.

Current commercially available systems for digital microscopy and computer-assisted cell detection can already provide sufficient accuracy for some blood cell types (e.g., segmented neutrophils, monocytes), especially in healthy individuals [21,24]. However, for other blood cells (e.g., lymphocyte subtypes), the correct detection rate is rather low, and data for disease classification with sufficient accuracy based on blood smear evaluation with these systems are lacking.

In this context, our study indicates the need for more training data, especially samples for difficult-to-classify classes, including cells labeled with disease information. Since collecting images of blood cells labeled by experts is time-consuming, especially for rare cell types data augmentation using generative models has the potential to provide more images for model training [4,18]. In addition, removal of experimen-

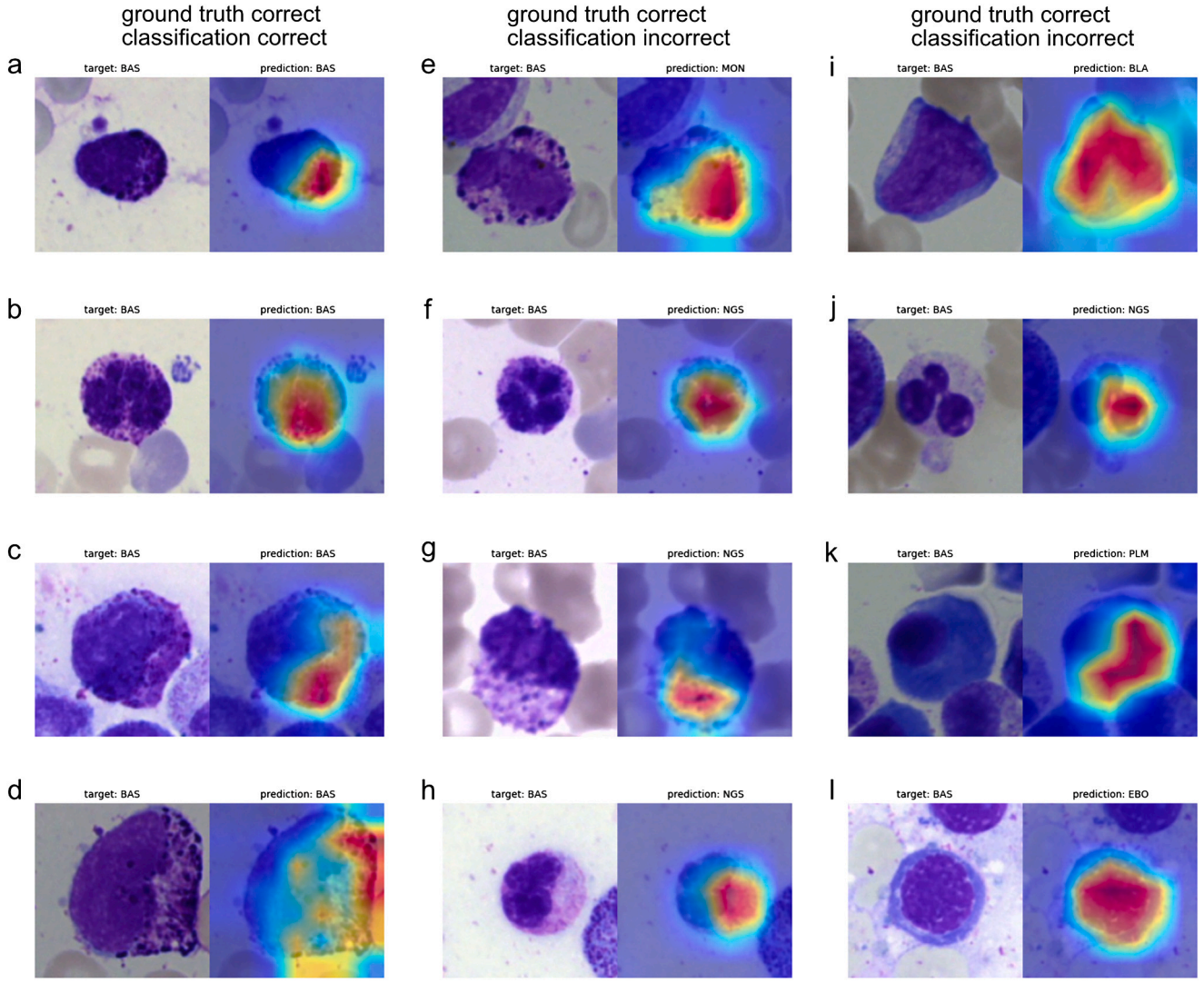


Fig. 6. Grad-CAM activation maps generated from the Basophils using the Regnet_y_32gf pre-trained on ImageNet + CD. Example images with corresponding activation maps for correct classified images (a-d), misclassified images with correct ground truth (e-h) and images not classified as Basophils due to incorrect ground truth (i-l). Regions showing high activation (in red) provide a strong contribution to the classification result.

tal noise in microscopy is often essential, especially for accurate cell classification, as highlighted for example in [14].

We see many promising and exciting results in the field of the automated evaluation of BM cell morphology that have the potential to improve patient outcomes. Besides the work of [32], the work of Wang et al. [49] is probably most comparable to our work, as they used a large dataset of $131,300$ expert-annotated cell images.

However, most of the work in this area has been done on datasets with small sample sizes or datasets that are not publicly available (cf. Sec. 1.1). Like Wagner et al. [48], we argue for the need for open datasets to enable reproducibility and reusability. Establishing benchmarks for model development will rapidly and sustainably advance computational pathology.

Last but not least, we emphasize the importance of AI-assisted decision tools adhering to the recommendations of professional societies and bodies [20,50,27].

Ethical approval

This study did not involve any human or animal participants, personal data, or biological material. Therefore, no ethical approval was required.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Stefan Balabanov reports a relationship with Alexion that includes: board membership. Stefan Balabanov reports a relationship with Amgen Europe GmbH that includes: board membership. Stefan Balabanov reports a relationship with Blueprint Medicines (Switzerland) GmbH that includes: speaking and lecture fees. Stefan Balabanov reports a relationship with Incyte Biosciences Germany GmbH that includes: speaking and lecture fees. Stefan Balabanov reports a relationship with Novartis that includes: speaking and lecture fees. Stefan Balabanov reports a relationship with Takeda Oncology that includes: speaking and lecture fees. Viktor Hendrik Koelzer reports a relationship with Indica Labs that includes: invited speaker. Viktor Hendrik Koelzer reports a relationship with Image Analysis Group that includes: funding grants. Viktor Hendrik Koelzer reports a relationship with SPCC that includes: speaker fees. Viktor Hendrik Koelzer reports a relationship with Roche that includes: funding grants; advisory board. Viktor Hendrik Koelzer reports a relationship with Takeda that includes: advisory board.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2023.107924.

References

[1] A. Acevedo, S. Alférez, A. Merino, L. Puigví, J. Rodellar, Recognition of peripheral blood cell images using convolutional neural networks, Comput. Methods Programs Biomed. 180 (2019) 105020.

[2] A.R. Andrade, L.H. Vogado, R. de MS Veras, R.R. Silva, F.H. Araujo, F.N. Medeiros, Recent computational methods for white blood cell nuclei segmentation: a comparative study, Comput. Methods Programs Biomed. 173 (2019) 1–14.

[3] K. Anilkumar, V. Manoj, T. Sagi, A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of leukemia, Biocybern. Biomed. Eng. 40 (2020) 1406–1420.

[4] O. Bailo, D. Ham, Y.M. Shin, Red blood cell image generation for data augmentation using conditional generative adversarial networks, CoRR, arXiv:1901.06219 [abs], 2019.

[5] F. Bardozzo, M. delli Priscoli, T. Collins, A. Forgione, A. Hostettler, R. Tagliaferri, Cross X-AI: explainable semantic segmentation of laparoscopic images in relation to depth estimation, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8, URL https://api.semanticscholar.org/CorpusID:252626691.

[6] V. Bianco, M. Valentino, J. Behal, D. Pirone, F. Bardozzo, P. Memmolo, L. Miccio, R. Tagliaferri, P. Ferraro, Deep learning assisted Fourier ptychography for cells and tissue analysis, in: P. Ferraro, D. Psaltis, S. Grilli (Eds.), Optical Methods for Inspection, Characterization, and Imaging of Biomaterials VI, in: International Society for Optics and Photonics, SPIE, 2023, p. 126220D.

[7] L. Boldú, A. Merino, A. Acevedo, A. Molina, J. Rodellar, A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images, Comput. Methods Programs Biomed. 202 (2021) 105999, https://doi.org/10.1016/j.cmpb.2021.105999, https://www.sciencedirect.com/science/article/pii/S0169260721000742.

[8] L. Bottou, On-Line Learning and Stochastic Approximations, Cambridge University Press, USA, 1999, pp. 9–42.

[9] C. Coombs, M. Tavakkoli, M. Tallman, Acute promyelocytic leukemia: where did we start, where are we now, and the future, Blood Cancer J. 5 (2015) e304.

[10] O. Dehaene, A. Camara, O. Moindrot, A. de Lavergne, P. Courtiol, Self-supervision closes the gap between weak and strong supervision in histology, 2020.

[11] M. Delli Priscoli, P. Memmolo, G. Ciaparrone, V. Bianco, F. Merola, L. Miccio, F. Bardozzo, D. Pirone, M. Mugnano, F. Cimmino, M. Capasso, A. Iolascon, P. Ferraro, R. Tagliaferri, Neuroblastoma cells classification through learning approaches by direct analysis of digital holograms, IEEE J. Sel. Top. Quantum Electron. 27 (2021) 1–9, https://doi.org/10.1109/JSTQE.2021.3059532.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[13] X. Fuentes-Arderiu, D. Dot-Bach, Measurement uncertainty in manual differential leukocyte counting, Clin. Chem. Lab. Med. 47 (2009) 112–115.

[14] F.H. Gil Zuluaga, F. Bardozzo, J.I. Rios Patino, R. Tagliaferri, Blind microscopy image denoising with a deep residual and multiscale encoder/decoder network, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 3483–3486.

[15] J. Gildenblat, contributors, Pytorch library for cam methods, https://github.com/jacobgil/pytorch-grad-cam, 2021.

[16] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[17] L. Guo, P. Huang, D. Huang, Z. Li, C. She, Q. Guo, Q. Zhang, J. Li, Q. Ma, J. Li, A classification method to classify bone marrow cells with class imbalance problem, Biomed. Signal Process. Control 72 (2022) 103296.

[18] D. Hazra, Y.C. Byun, W.J. Kim, Enhancing classification of cells procured from bone marrow aspirate smears using generative adversarial networks and sequential convolutional neural network, Comput. Methods Programs Biomed. 224 (2022) 107019.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[20] G. Karimian, E. Petelos, S.M.A.A. Evers, The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review, AI Ethics 2 (2022) 539–551, https://link.springer.com/10.1007/s43681-021-00131-7.

[21] H.N. Kim, M. Hur, H. Kim, S.W. Kim, H.W. Moon, Y.M. Yun, Performance of automated digital cell imaging analyzer Sysmex DI-60, Clin. Chem. Lab. Med. 56 (2018) 94–102.

[22] S. Krappe, M. Benz, T. Wittenberg, T. Haferlach, C. Münzenmayer, Automated classification of bone marrow cells in microscopic images for diagnosis of leukemia: a comparison of two classification schemes with respect to the segmentation quality, in: L.M. Hadjiiski, G.D. Tourassi (Eds.), Medical Imaging 2015: Computer-Aided Diagnosis, International Society for Optics and Photonics, SPIE, 2015, pp. 858–863.

[23] S. Krappe, T. Wittenberg, T. Haferlach, C. Münzenmayer, Automated morphological analysis of bone marrow cells in microscopic images for diagnosis of leukemia: nucleus-plasma separation and cell classification using a hierarchical tree model of hematopoesis, in: G.D. Tourassi, S.G. Armato III (Eds.), Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics, SPIE, 2016, pp. 856–861.

[24] A. Kratz, S.h. Lee, G. Zini, J.A. Riedl, M. Hur, S. Machin, Digital morphology analyzers in hematology: ICSH review and recommendations for Standardization in Haematology, I.C., Int. J. Lab. Hematol. 41 (2019) 437–447.

[25] S.H. Lee, W. Erber, A. Porwit, M. Tomonaga, L. Peterson, ICSIH, ICSH guidelines for the standardization of bone marrow specimens and reports, Int. J. Lab. Hematol. 30 (2008) 349–364.

[26] S.H. Lee, S. Lee, B.C. Song, Vision transformer for small-size datasets, CoRR, arXiv:2112.13492 [abs], 2021.

[27] K. Lekadir, G. Quaglio, A. Tselioudis Garmendia, C. Gallin, Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts, European Parliamentary Research Service, 2022.

[28] Y. Liang, Z. Tang, M. Yan, J. Chen, Q. Liu, Y. Xiang, Comparison detector for cervical cell/clumps detection in the limited data scenario, Neurocomputing 437 (2021) 195–205.

[29] J. Liu, Y. Pan, M. Li, Z. Chen, L. Tang, C. Lu, J. Wang, Applications of deep learning to MRI images: a survey, Big Data Min. Anal. 1 (2018) 1–18.

[30] H. Löffler, J. Rastetter, Atlas of Clinical Hematology, Springer Science & Business Media, 2012.

[31] C. Matek, S. Krappe, C. Münzenmayer, T. Haferlach, C. Marr, An expert-annotated dataset of bone marrow cytology in hematologic malignancies [data set], 2021.

[32] C. Matek, S. Krappe, C. Münzenmayer, T. Haferlach, C. Marr, Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set, Blood 138 (2021) 1917–1927.

[33] J. Mori, S. Kaji, H. Kawai, S. Kida, M. Tsubokura, M. Fukatsu, K. Harada, H. Noji, T. Ikezoe, T. Maeda, A. Matsuda, Assessment of dysplasia in bone marrow smear with convolutional neural network, Sci. Rep. 10 (2020) 14734.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[35] J. Prinyakupt, C. Pluempitiwiriyawej, Segmentation of white blood cells and comparison of cell morphology by linear and naïve Bayes classifiers, Biomed. Eng. Online 14 (2015) 1–19.

[36] I. Radosavovic, R.P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10425–10433.

[37] R. Rahmé, X. Thomas, C. Recher, N. Vey, J. Delaunay, E. Deconinck, P. Hirsch, D. Bordessoule, J. Micol, A. Stamatoullas, et al., Early death in acute promyelocytic leukemia (APL) in French centers: a multicenter study in 399 patients, Leukemia 28 (2014) 2422–2424.

[38] T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik, ImageNet-21k pretraining for the masses, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Curran, 2021, URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98f13708210194c475687be6106a3b84-Paper-round1.pdf.

[39] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[41] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[42] Y. Sharmay, L. Ehsany, S. Syed, D.E. Brown, HistoTransfer: understanding transfer learning for histopathology, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2021, pp. 1–4.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[44] L.N. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 464–472.

[45] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018, Springer International Publishing, Cham, 2018, pp. 270–279.

[46] D.C. Tkachuk, J.V. Hirschmann, M.M. Wintrobe, Wintrobe's Atlas of Clinical Hematology, Lippincott Williams & Wilkins, 2007.

[47] B.S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation equivariant CNNs for digital pathology, arXiv:1806.03962, 2018.

[48] S. Wagner, C. Matek, S. Boushehri, M. Boxberg, L. Lamm, A. Sadafi, D. Waibel, C. Marr, T. Peng, Make deep learning algorithms in computational pathology more reproducible and reusable, Nat. Med. 28 (2022) 1–3, https://doi.org/10.1038/s41591-022-01905-0.

[49] W. Wang, M. Luo, P. Guo, Y. Wei, Y. Tan, H. Shi, Artificial intelligence-assisted diagnosis of hematologic diseases based on bone marrow smears using deep neural networks, Comput. Methods Programs Biomed. 231 (2023) 107343.
[50] WHO, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, World Health Organization, 2021.
[51] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5987–5995.