Annotation-efficient classification combining active learning, pre-training and semi-supervised learning for biomedical images

Sayedali Shetab Boushehri[1,3,5,*] , Ahmad Bin Qasim[1,4,*], Dominik Waibel[1,2], Fabian Schmich[5], Carsten Marr[1]

1 Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
2 Technical University of Munich, School of Life Sciences, Weihenstephan, Germany
3 Technical University of Munich, Department of Mathematics, Munich, Germany
4 Technical University of Munich, Department of Informatics, Munich, Germany
5 Roche Innovation Center Munich, Roche Diagnostics GmbH, Penzberg, Germany
* Equal contribution

Abstract

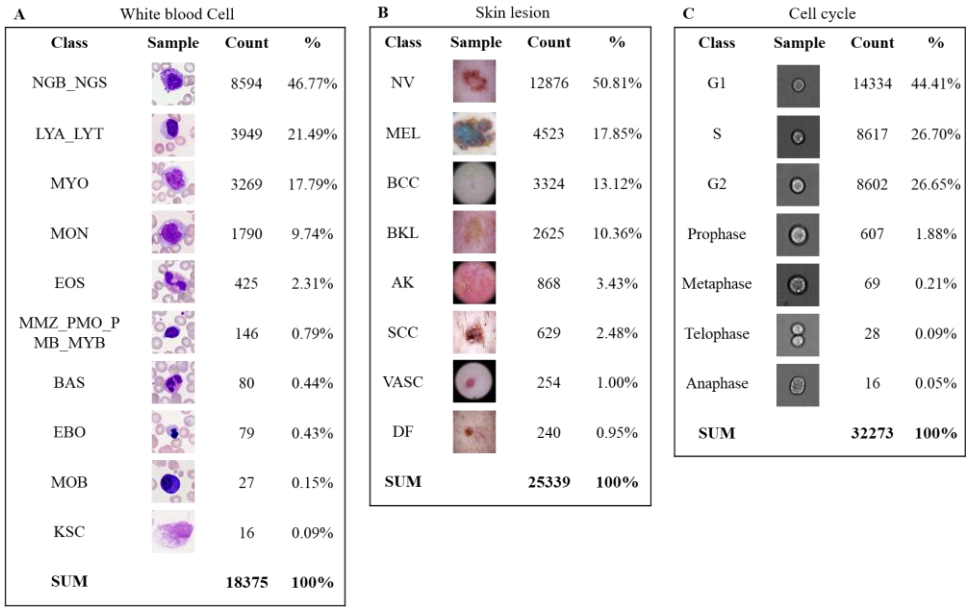
Deep learning image classification algorithms typically require large annotated datasets. In contrast to real world images where labels are typically cheap and easy to get, biomedical applications require experts' time for annotation, which is often expensive and scarce. Therefore, identifying methods to maximize performance with a minimal amount of annotation is crucial. A number of active learning algorithms address this problem and iteratively identify most informative images for annotation from the data. However, they are mostly benchmarked on natural image datasets and it is not clear how they perform on biomedical image data with strong class imbalance, little color variance and high similarity between classes. Moreover, active learning neglects the typically abundant unlabeled data available.

In this paper, we thus explore strategies combining active learning with pre-training and semi-supervised learning to increase performance on biomedical image classification tasks. We first benchmarked three active learning algorithms, three pre-training methods, and two training strategies on a dataset containing almost 20,000 white blood cell images, split up into ten different classes. Both pre-training using self-supervised learning and pre-trained ImageNet weights boosts the performance of active learning algorithms. A further improvement was achieved using semi-supervised learning. An extensive grid-search through the different active learning algorithms, pre-training methods and training strategies on three biomedical image datasets showed that a specific combination of these methods should be used. This recommended strategy improved the results over conventional annotation-efficient classification strategies by 3% to 14% macro recall in every case. We propose this strategy for other biomedical image classification tasks and expect to boost performance whenever scarce annotation is a problem.

1. Introduction

Recent success of deep learning methods rely heavily on large amounts of well-annotated training data [1]. Especially for biomedical images, annotations are scarce as they crucially depend on the availability of trained experts whose time is often expensive and limited. Active learning algorithms are designed to address this issue by finding the most informative images for annotation [2][3][4] but are mostly benchmarked on natural image datasets such as ImageNet [5][6][7]. Biomedical images however differ in their characteristics from natural images. They are typically not as diverse in terms of color range and often they are classified by only small feature variations, e.g. in texture and size [8][9]. Moreover, biomedical image datasets are often imbalanced, containing rare classes, which can significantly influence the diagnosis. **Active learning** has been shown to work in biomedical image classification tasks [3][10] and image segmentation [11]. However, it is not clear which particular active learning algorithm will be the most suitable for different biomedical image data and how the performance can be improved by combining it with other deep learning methods.

Pre-training methods such as transfer learning and self-supervised pre-training show a great potential for being used as the network's initial weights to improve the network performance on classification tasks involving low number of labeled images [12][13][14]. Here, a network uses representation from another, ideally similar dataset (i.e. transfer learning), or it learns a representation without incorporating any labels (self-supervised learning)[16]. The most common transfer learning method is to use pre-trained ImageNet weights. This method has been used in many biomedical applications to initialize deep learning models [17][18]. However, Raghu and Zhang et al. [19] showed that in several biomedical imaging applications, transfer learning from ImageNet does not lead to better results. Furthermore, self-supervised learning has recently been shown to be effective for improving classification performance on biomedical images [24].



A White blood Cell

Class	Sample	Count	%
NGB_NGS		8594	46.77%
LYA_LYT		3949	21.49%
MYO		3269	17.79%
MON		1790	9.74%
EOS		425	2.31%
MMZ_PMO_PMB_MYB		146	0.79%
BAS		80	0.44%
EBO		79	0.43%
MOB		27	0.15%
KSC		16	0.09%
SUM		**18375**	**100%**

B Skin lesion

Class	Sample	Count	%
NV		12876	50.81%
MEL		4523	17.85%
BCC		3324	13.12%
BKL		2625	10.36%
AK		868	3.43%
SCC		629	2.48%
VASC		254	1.00%
DF		240	0.95%
SUM		**25339**	**100%**

C Cell cycle

Class	Sample	Count	%
G1		14334	44.41%
S		8617	26.70%
G2		8602	26.65%
Prophase		607	1.88%
Metaphase		69	0.21%
Telophase		28	0.09%
Anaphase		16	0.05%
SUM		**32273**	**100%**

Figure 1. Biomedical image datasets used in this study are exhibiting strong class imbalance, little color variance and high similarity between classes. (A) White blood cell: A dataset with 18357 images (128x128 pixel) of white human blood cells with ten expert labeled classes from blood smears of 100 patients diagnosed with Acute Myeloid Leukemia (AML) and 100 individuals which show no symptoms of the disease [8][24][25]. (B) Skin lesion: A dataset with 25339 dermoscopy images (128x128 pixel) of skin lesions with eight skin cancer classes [26][27][28]. (C) Cell cycle: A dataset comprising 32272 images (64x64 pixel) of Jurkat cells in seven different cell cycle stages created by imaging flow cytometry [29].

Finally, **semi-supervised learning** uses unlabeled data to increase the performance as well as the stability of predictions [21][22]. In the field of biomedical imaging, many applications leverage high-throughput technology [23] to generate large quantities of unlabeled data, whereas, as discussed, annotations are typically scarce. Thus, the paradigm of semi-supervised learning is particularly appealing in this domain.

In this paper, first we compare different active learning algorithms on a challenging biomedical image dataset. We improve the results of the best algorithm by adding pre-training and semi-supervised learning. To prove that whether this combination of active learning algorithm, pre-training and training strategy always works, We perform an extensive grid-search on three active learning algorithms plus random sampling (baseline), three pre-training methods plus random initialization (baseline), and two training strategies including supervised and semi-supervised learning on three exemplary biomedical image data sets. As the result of this investigation, we find an optimal strategy for incomplete-supervision biomedical image data.

2. Datasets

We evaluate the efficiency and performance of combinations of active learning algorithms, pre-training methods, and training strategies on three fully annotated datasets from the biomedical imaging field (Figure 1).

3. Methods

In this section, we define the active learning algorithms, pre-training methods and training strategies evaluated throughout this paper. We consider that there exists a labeled subset of our data, L, such that $L = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)...(x_N, y_N)\}$, with x_i being an image and y_i the corresponding label. Also, a subset of unlabeled images U exists, where $U = \{u_1, u_2, u_3...u_K\}$ and $K>>N$. By definition, we consider $D = L \cup U$, where D is the whole dataset. We define a model as f_Θ with parameters Θ, and a stochastic augmentation function a. The function a consists of multiple augmentation steps such as cropping, flipping, rotating, random noise etc.

3.1. Active learning algorithms

The performance of a model f_Θ with parameters Θ can be increased by labeling images from U, and thus adding pairs of images and corresponding labels (x_i, y_i) to L. The labeling of unlabeled image is carried out in iterations, which consist of the selection of s images $S \subseteq U$ with $|S| = s$ for annotation, after the performance of the model converges with the updated labeled set L. Active learning algorithms aim on selecting images in U for annotation, such that the addition of these images to L results in a maximum increase in the

evaluation metrics M. The main difference between active learning algorithms is how images are chosen for labeling. The algorithms evaluated in this paper are based on model uncertainty δ. The s images $S \subseteq U$ with $|S| = s$ with the highest uncertainty are selected for labeling in each iteration. In this work, we compare three different active learning algorithms:

Random sampling: During each active learning iteration each image in $S \subseteq U$ is chosen arbitrarily. Random sampling acts as a baseline. Hence, all other algorithms are expected to perform better than random sampling.

Entropy-based sampling: Entropy measures the average amount of information or "bits" required for encoding the distribution of a random variable [2]. Here, entropy is used as criteria for active learning [2] to select the s images $S \subseteq U$, whose predicted outcomes have the highest entropy, assuming that high entropy of predictions mean high model uncertainty δ. By definition, entropy focuses on the whole predicted distribution rather than only on the highest probability outcomes of the model [2].

Augmentation-based sampling: Let a be a function that performs stochastic data augmentation, such as cropping, horizontal flipping, vertical flipping or erasing on a given image. Each unlabeled image $u_i \in U$ is transformed using a and this process is repeated J times to obtain the set U_i with $|U_i| = J$. The random transformations are followed by a forward-pass through the model f_Θ. This results in J predictions $\hat{Q}_i = \{\hat{q}_{1i}, \hat{q}_{2i}, \hat{q}_{3i} \ldots \hat{q}_{Ji}\}$, where $\hat{q}_i = \text{argmax}\, P_\Theta(\hat{y}_i|u_i)$ is the most probable class according to the model output for each set U_i of perturbed copies of an unlabeled image $u_i \in U$. The model uncertainty δ can be estimated by keeping a count of the most frequently predicted class (mode) for each image. The idea behind this approach is that if the model is certain about an image then it should output the same prediction for randomly augmented versions. So the lower the frequency of the mode, the higher the uncertainty δ[3]. During each active learning iteration, the images with the lowest frequency of the most frequently predicted class are annotated and added to the labeled set L.

Monte Carlo (MC) dropout: Dropout is a commonly used technique for model regularization, which randomly ignores a fraction of neurons during training to mitigate the problem of overfitting. It is typically disabled during test time. MC-dropout involves the assessment of uncertainty in neural networks using dropout at test time [30][31] and thus estimates the uncertainty of the prediction of an image. MC-dropout generates non-deterministic prediction distributions for each image. The variance of this distribution can be used as an approximation for model uncertainty δ [32]. During each active learning iteration, the images with the highest variance are annotated and added to the labeled set L. This has been shown to be an effective selection criterion during active learning [5].

3.2. Pre-training methods

Network initialization can increase the performance of neural networks [33]. It is considered to be even more essential when the amount of annotated data is not considerably large [20]. In this work, we utilize three different pre-training methods plus random initialization (baseline):

Random initialization was shown to perform poorly compared to more sophisticated initialization measures [34]. We use Kaiming He initialization [35] as a baseline random initialization method.

ImageNet weights are obtained by training a feature extraction network on the ImageNet dataset. After training on ImageNet data, the weights of the feature extractor network can be used for initialization of models which are to be trained on other datasets [19]. This has become a standard pre-training for classification tasks as it often helps the network converge faster than with random initialization. It also has been shown to be beneficial in low-data biomedical imaging regimes [19].

Autoencoders are a class of neural networks used for feature extraction [36]. The objective of the autoencoders is to reconstruct the input. An encoder network e encodes the input x into its latent representation $e(x)$. The encoder typically includes a bottleneck layer with relatively few nodes. The bottleneck layer forces the encoder to represent the input data in a compact form. This latent representation is then used as an input to a decoder network d which tries to output a reconstruction $d(e(x))$ of the original input. Hence autoencoders do not require labels for training and the whole dataset can be used for training an autoencoder architecture. For pre-training the encoder is used as a feature extraction network while the decoder is generally discarded. This has been shown to significantly improve network initialization on biomedical image datasets [37].

SimCLR is a framework for contrastive learning of visual representations [12]. It learns representations in a self-supervised manner by using an objective function that minimizes the difference between representations of the model f_Θ on a pairs of differently augmented copies of the same image. Let a be a function that performs stochastic data augmentations (such as cropping, adding color jitter, horizontal flipping and gray scale) on a given image. Each image $x \in D$ in a mini-batch of size B is passed through the stochastic data augmentation function a twice to obtain $X_i = \{x_{1i}^{'}, x_{2i}^{'}\}$. These pairs can be termed as positive pairs as they originate from the same image x_i. A neural network encoder e extracts the feature vectors from the augmented images. A multi-layer perceptron with one hidden layer is used as a projection head for projecting the feature vectors h to the projection space where then, a contrastive loss is applied. The contrastive loss function is a softmax loss function applied on a similarity measure between positive pairs

against all the negative examples in the batch and is weighted by the temperature parameter τ that controls the weight of negative examples in the objective function. SimCLR training does not require labels and the whole dataset can be used for training. Using SimCLR as a pre-training method shows significant improvement on ImageNet classification [12].

3.3. Training strategies

Large amounts of unlabeled data are typically available in biomedical applications. Ideally, this unlabeled data is not only used for network initialization but also during training. Thus, we compare the performance of training the model only using the existing labeled data a.k.a. supervised learning versus a semi-supervised approach, which incorporates the unlabeled data in the training process.

In **supervised learning** we are looking for a model f_Θ with parameters Θ to learn a mapping $\hat{Y} = f_\Theta(L)$ such that the objective function $Loss\ (\hat{y}_i, y_i)$ is minimized. Supervised learning uses only labeled data. The performance of the model can be evaluated using an evaluation metric M such as accuracy, recall etc. The objective function used in this paper is the multi-class cross-entropy loss function, $Loss = -\sum_i^N \sum_j^C y_{ij}\ \log(\hat{y}_{ij})$ with C being the total number of classes in the dataset and N being the size of L.

For **Semi-supervised learning,** we use FixMatch [21], a combination of consistency regularization [38] and pseudo-labeling [39]. Given the set of unlabeled images $U = \{u_1, u_2, u_3 ... u_K\}$ with $|U| = K$, consistency regularization tries to maximize the similarity between model outputs, obtained by passing stochastically augmented versions of the same image through the model $f_\Theta(a(x))$. Pseudo-labeling refers to using pseudo-labels for unlabeled images. Pseudo- labels are obtained by passing the unlabeled images through the model f_Θ, i.e. $\hat{Y} = f_\Theta(U)$ and using the outcome with maximum probability in the predicted distribution $\hat{q}_i = \text{argmax}\ P_\Theta(\hat{y}_i|x_i)$ as the pseudo-label if the maximum probability value $\hat{q}_i$ is above a threshold τ. Using the pseudo-labels, the unlabeled images are added to the set of labeled images L temporarily.

The FixMatch loss consists of a supervised loss term i.e. the multi-class cross-entropy loss and the unsupervised loss term. The unsupervised loss term is calculated by passing the unlabeled dataset through a stochastic weak augmentation function a_{weak} (e.g. rotation and translation) and then applying pseudo-labeling on the output prediction distribution with threshold. Another set of pseudo-labels is obtained by passing the unlabeled dataset through stochastic strong augmentation function a_{strong} (e.g. color distortion, random noise and random erasing). After calculating the two sets of pseudo-labels for unlabeled images, consistency regularization is applied by calculating cross-entropy between the pseudo-labels. The loss function contains the weighting parameter λ, which weighs the unsupervised loss term:

$$L_{fixmatch} = L_{supervised} + \boldsymbol{\lambda} \cdot L_{unsupervised} \quad (1)$$

Significant performance improvement has been observed over supervised training in a low-data regime [21].

4. Results

In this study for each experiment, we use randomly selected 1% of data as our initial annotated set. Then in each iteration, we add 5% of data as annotated using the algorithms in section 3.1. This process is repeated 4 times which leads to adding 20% and in total 21% of labeled data. Moreover, we perform a 4-fold cross-validation in each iteration and calculate macro accuracy, precision, recall, and F1-score. We use the macro recall, defined as the average of recall per class, as our main metric of comparison, to account for the imbalanced nature of the datasets and the existence of rare classes.

We use ResNet18 [40] as the fixed architecture for training. For each dataset, we pre-trained the ResNet18 using an autoencoder or SimCLR [12]. For the autoencoder pre-training, we used a feature extractor network consisting of a ResNet18 encoder and a decoder with transposed convolutional layers. After training the autoencoder, the ResNet18 encoder is used as a feature extractor network while the decoder is discarded.

4.1. Comparison of active learning algorithms on white blood cell data

We first compared the performance of different annotation-efficient approaches on the white blood cell dataset (Figure 2A). We started the training with random initialization of the network and used labeled data for training in an iterative fashion. The augmentation-based sampling outperforms the other active learning algorithms (see Table 1) in almost all iterations (see Figure 2A). When 20% of the dataset is added as annotated images, augmentation-based sampling reaches a macro recall of 0.72±0.03 (mean±standard deviation from 4-fold cross-validation), entropy-based sampling a macro recall of 0.72±0.02, MC-dropout a recall of 0.66±0.04 and random sampling a recall of 0.68±0.02.

4.2. Pre-training on white blood cell images further improves performance

We next tried to improve the best performing active learning algorithm (augmentation-based sampling) by incorporating pre-training (Figure 2B). We repeated the experiment using augmentation-based sampling with 3 pre-trained networks using weights from ImageNet, an

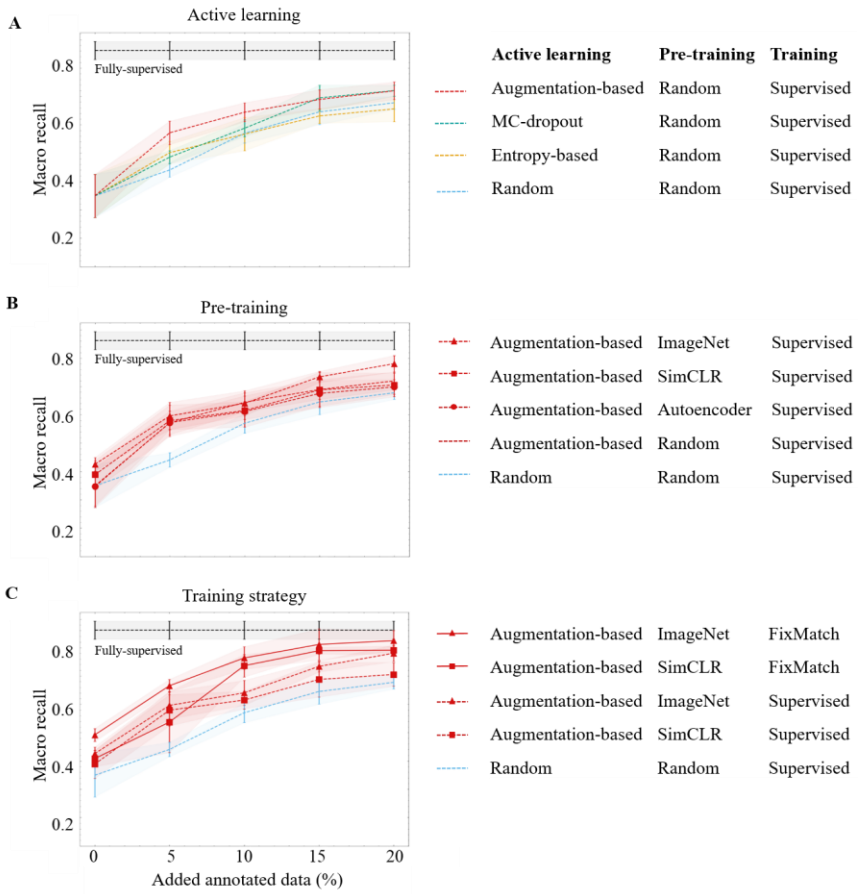


Figure 2. On the white blood cell dataset, combining augmentation-based sampling, ImageNet pre-training and semi-supervised learning via FixMatch converges to the performance of fully-supervised learning. (A) We compute the macro recall for three different active learning algorithms including augmentation-based sampling (dashed red line) entropy-based sampling (dashed green line), and MC-dropout (dashed yellow line) and compare it to random sampling (dashed blue line). We used 1% of the data as our initial labeled set. In each iteration, we added 5% to the labeled set. We show mean ± standard deviation of the macro recall from 4-fold cross-validation. (B) We chose augmentation-based sampling (dashed red line, as in A) as the best active learning algorithm and now compared different pre-training methods including ImageNet weights (triangle), SimCLR (square), and autoencoder (circle) with random initialization (dashed blue line). (C) To study the effect of semi-supervised learning, we repeated the best performing experiments from B using FixMatch. Two combinations of augmentation-based active learning, ImageNet pre-training and FixMatch (solid red line with triangle) as well as augmentation-based sampling, SimCLR pre-training and FixMatch (solid red line with square) were implemented.

autoencoder and SimCLR (see Methods and Table 1). We find that both the starting point as well as the first two iterations show the highest improvement, with the macro recall increasing at least 12% in all cases. However, random initialization catches up with the SimCLR and autoencoder pre-training when 10% of annotated data is used. Interestingly, the combination of augmentation-based sampling with ImageNet pre-training is always outperforming the other pre-training methods with a noticeable difference, even at the last iteration when 20% of the dataset is added as labelled images. Here, the augmentation-based sampling with ImageNet weights reaches a macro recall of 0.78±0.03, random initialization reaches a macro recall of 0.72±0.03 and initialization with SimCLR pre-training reaches a recall of 0.71±0.04.

4.3. Semi-supervised learning further improves recall for white blood cells data

Now we investigate the effect of using unlabeled data during training. We choose augmentation-based sampling as the best performing active learning algorithm (Figure 2A) and the best two pre-training methods, i.e. ImageNet and SimCLR (Figure 2B), for training with FixMatch (Figure 2C). Clearly, adding semi-supervised learning improves performance for the initial step and the rest of iterations with more than 6% of macro recall increase. This combination outperforms supervised training in every iteration, reaching

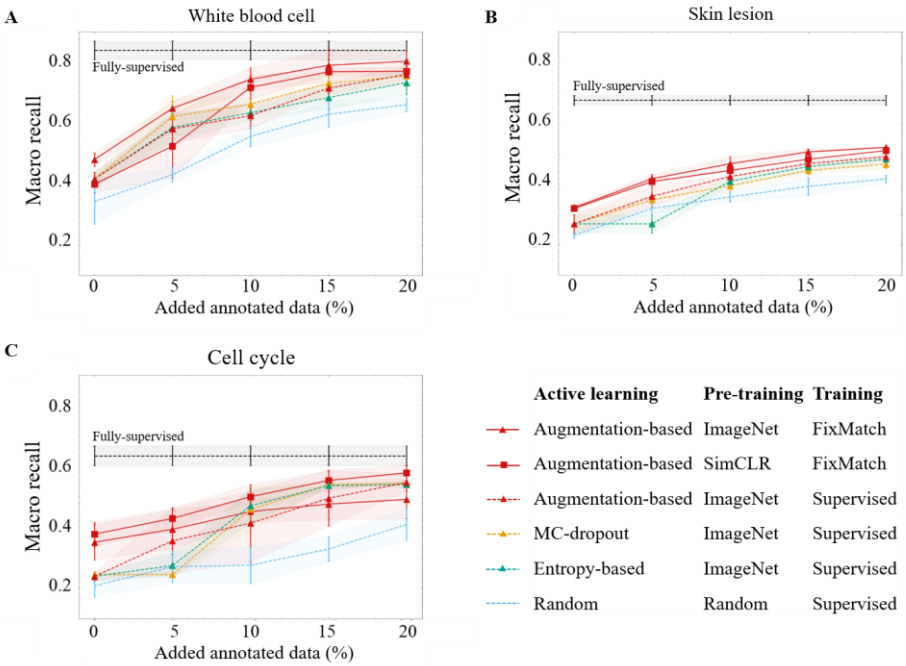


Figure 3. The combination of augmentation-based sampling, SimCLR or ImageNet pre-training and semi-supervised training with FixMatch is the optimal strategy on all three biomedical datasets. We show mean ± standard deviation of the macro recall from 4-fold cross-validation. **(**A) On the white blood cell dataset the optimal strategy with ImageNet initialization outperformed all other baseline methods for each active learning iteration by at least 3%. With only 20% of added annotated data, this combination performs almost as good as a fully supervised trained model. **(**B) On the skin lesion dataset the optimal strategies with ImageNet and SimCLR pre-training outperformed all other methods. During the initial step (no added data) and 5% added data (first iteration), both optimal strategies were at least 4% better than all baseline methods. **(**C) On the cell cycle dataset the optimal strategies with ImageNet and SimCLR pre-training were ~14% better than all baseline methods with no added data. Nonetheless, the optimal strategy with ImageNet pre-training did not improve as rapidly as the optimal strategy with SimCLR pre-training. The optimal strategy with SimCLR pre-training was ~3% better than all baseline methods and only 6% worse than the fully supervised trained model, however using only 20% of annotated data.

0.82±0.04 macro recall with 20% of the added annotated data in the last iteration. FixMatch also improves augmentation-based sampling with SimCLR pre-training, reaching 0.79±0.01 macro recall. Interestingly the macro recall is only 4% lower than using fully-supervised learning on the whole data.

4.4. Grid-search identifies the best performing combinations for three biomedical datasets

To investigate whether the combination of augmentation-based sampling for active learning, ImageNet or SimCLR weights for pre-training, and FixMatch as the training strategy is always outperforming other combinations of the methods listed in Table 1 for three substantially different biomedical datasets (Figure 1), we performed a systematic grid-search. Specifically, we ran 3x4x4x2x4x5 = 1920 independent runs (3 datasets, 3 active learning algorithms plus random sampling, 3 pre-training methods plus random initialization, 2 training strategies, 4-fold cross-validation and 1 initial step plus 4 active learning iterations) to identify the best combination. We used the macro recall in the last iteration (using 20% of annotated data) as our criteria for performance.

Active learning algorithms	Pre-training methods	Training strategies
- Random (baseline) - MC-dropout - Augmentation-based - Entropy-based	- Random (baseline) - ImageNet - Autoencoder - SimCLR	- Supervised learning - Semi-supervised learning (FixMatch)

Table 1: We compared the combination of active learning algorithms, different network pre-training methods and training strategies on three biomedical image datasets (Figure 1).

We found that the combination of augmentation-based sampling with ImageNet or SimCLR pre-training and FixMatch consistently outperforms the rest (for comparing all the combinations, please refer to the supplementary materials).

For the white blood cell dataset, already at the initial step (1% labeled data) we see a 6% improvement using FixMatch with ImageNet initialization over conventional training with only labeled data (Figure 3A). This difference

A

	Active learning algorithm	Pre-training method	Training strategy	Accuracy	Precision	Recall	F1-score
Fully-supervised	-	ImageNet	Supervised learning	0.95±0.02	0.71±0.01	0.86±0.03	0.73±0.02
Recommended	Augmentation-based	ImageNet	FixMatch	**0.92±0.01**	0.78±0.03	**0.82±0.04**	**0.70±0.03**
Conventional (Baseline)	Augmentation-based	ImageNet	Supervised learning	0.92±0.02	0.78±0.07	0.78±0.03	0.67±0.05
Conventional (Baseline)	MC-dropout	ImageNet	Supervised learning	0.90±0.01	**0.81±0.03**	0.78±0.01	0.66±0.07
Conventional (Baseline)	Entropy-based	ImageNet	Supervised learning	0.91±0.01	0.8±0.05	0.75±0.04	0.64±0.03
No active learning	Random	Random	Supervised learning	0.89±0.02	0.74±0.03	0.68±0.02	0.58±0.03

B

	Active learning algorithm	Pre-training method	Training strategy	Accuracy	Precision	Recall	F1-score
Fully-supervised	-	ImageNet	Supervised learning	0.75±0.04	0.67±0.01	0.68±0.02	0.67±0.02
Recommended	Augmentation-based	ImageNet	FixMatch	**0.65±0.00**	0.44±0.01	**0.52±0.00**	**0.46±0.01**
Conventional (Baseline)	Augmentation-based	ImageNet	Supervised learning	0.59±0.01	0.39±0.01	0.49±0.01	0.42±0.01
Conventional (Baseline)	Entropy-based	ImageNet	Supervised learning	0.58±0.03	**0.39±0.03**	0.48±0.01	0.41±0.03
Conventional (Baseline)	MC-dropout	ImageNet	Supervised learning	0.56±0.01	0.36±0.02	0.46±0.01	0.38±0.02
No active learning	Random	Random	Supervised learning	0.50±0.02	0.32±0.02	0.42±0.01	0.33±0.02

C

	Active learning algorithm	Pre-training method	Training strategy	Accuracy	Precision	Recall	F1-score
Fully-supervised	-	ImageNet	Supervised learning	0.74±0.03	0.71±0.02	0.64±0.03	0.61±0.03
Recommended	Augmentation-based	SimCLR	FixMatch	**0.74±0.01**	0.58±0.00	**0.58±0.03**	**0.51±0.03**
Conventional (Baseline)	Augmentation-based	ImageNet	Supervised learning	0.69±0.02	0.67±0.05	0.55±0.03	0.45±0.06
Conventional (Baseline)	MC-dropout	ImageNet	Supervised learning	0.73±0.02	**0.77±0.01**	0.54±0.01	0.50±0.01
Conventional (Baseline)	Entropy-based	ImageNet	Supervised learning	0.72±0.01	0.68±0.08	0.54±0.01	0.49±0.01
No active learning	Random	Random	Supervised learning	0.69±0.02	0.77±0.04	0.41±0.05	0.36±0.06

Table 2. Comparing the results of the last iteration, our recommended strategies outperform conventional annotation-efficient learning. (A) On the white blood cell dataset, the combination of augmentation-based sampling, ImageNet pretraining and FixMatch training brings an improvement of 4% on macro recall and 3% on F1-score over the highest baseline. With using only 20% of added labeled data, this strategy is only 4% lower in recall and 3% lower with respect to the F1-score as compared to fully-supervised training. (B) On the skin lesions dataset, the recommended strategy brings an improvement of 3% on macro recall, 5% improvement on precision and 6% on F1-score. The high recall difference to the fully-supervised results shows that the amount of labeled data was not enough and more iterations were needed. (C) On the cell cycle dataset, the recommended strategy brings an improvement of 3% on recall and 6% on F1-score.

seems to be consistent in all the iterations, resulting in 4% improvement in total compared to the best results using only labeled data for training. For the skin lesions dataset we see the same trend (Figure 3B). The initial step using semi-supervised learning with either ImageNet or SimCLR initialization is at least 5% better than every conventional supervised learning strategy. While in the next iterations conventional methods get closer, there is always a performance difference. Finally, for the cell cycle dataset (Figure 3C), combining SimCLR and FixMatch gives a drastic boost with more than 16% improvement compared to conventional methods at the start. While this improvement gets less after adding 10% of the labeled data, there is still a considerable difference between the methods. Using 20% of the labeled data, we still see a 3% improvement.

Looking at the final iteration (using 21% of the whole data as labelled images) for the white blood cells and the cell cycle dataset reveals that we can reach a performance similar to fully-supervised learning, which incorporates the fully annotated dataset, with only a ⅕ of labels (Figure 3). This observation does not hold for the skin lesions dataset however, which apparently requires more labeled data for training (Figure 3).

4.5. Recommended strategy

As a result of the previous sections, we have identified the optimal combination of augmentation-based sampling, ImageNet/SimCLR pre-training and FixMatch to show the best results on three biomedical datasets. As illustrated in Figure 3, the ImageNet pre-training works better for white blood cells and the skin lesions from the initial step. SimCLR pre-training seems to work best on the cell cycle data. Therefore, our recommended strategy is to find the best pre-training method on the initial step and combine it with augmentation-based sampling and FixMatch during training. The results of our recommended strategy improves macro recall by 4% for white blood cells data, 3% on skin lesions data and 3% for cell cycle data on the last iteration, with respect to the best conventional active learning method for each dataset.

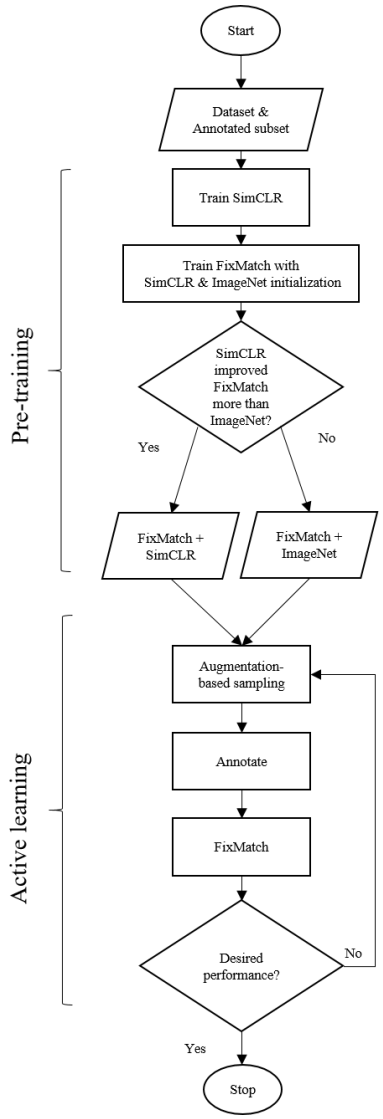


Figure 4. Recommended strategy for annotation-efficient classification of biomedical image data involves SimCLR or ImageNet pre-training, FixMatch as the semi-supervised algorithm for training and augmentation-based sampling during active learning until the desired performance is reached.

5. Conclusion

In this paper, we have investigated the performance of different annotation-efficient learning strategies for biomedical image classification. First, we showed that for classifying white blood cells into 10 different classes, active learning could boost macro recall. Second, we showed using ImageNet and SimCLR, pre-training could increase the performance further. However, their contribution is dataset dependent: While for white blood cell and skin lesion dataset, ImageNet weight led to better performance, SimCLR performed better for classifying cell cycles (Figure 3). This might be due to the nature of images: Cell cycle data is captured by fluorescent imaging, which follows a very different color distribution than other technologies such as dermoscopy cameras, which are closer to natural images. Therefore, ImageNet pre-training might not be the preferred way for such data.

We also showed that by incorporating unlabeled data in the training process in a semi-supervised manner, one can improve the performance of the classification noticeably. Finally, by doing a grid-search over all the possible algorithms and strategies (Table 1), we found out that the combination of ImageNet or SimCLR pre-training, FixMatch semi-supervised learning and augmentation-based sampling can improve existing methods for every dataset. The reason for this is probably the fact that while training FixMatch, the network faces many different augmentations for each image and learns to make a robust prediction. Augmentation-based sampling relies on the same idea for finding those images where predictions were not robust enough.

As a result of this study, we propose an annotation-efficient strategy for biomedical imaging active learning tasks where unlabeled data is abundant (Figure 4). We split our strategy into two parts including pre-training and active learning. First, we suggest to pre-train the network using SimCLR. Then compare FixMatch initialized with ImageNet weights to SimCLR pre-training. By comparing the results, select the best pre-training method. Eventually for the active learning part, we recommend to train FixMatch along with the best pre-training method and augmentation-based sampling to obtain optimal results.

Although our work shows potential for improvement of annotation-efficient learning for three biomedical image classification datasets, the methodology should be tested on more datasets to gain insights into correlations between dataset characteristics and the performance of the applied methods. Due to the computational costs, we used a fixed architecture and a fixed set of parameters. As the next step, we will try different architectures and parameters and evaluate the results accordingly. In addition, a variety of active learning, semi-supervised and self-supervised learning methods should be added to the work to find the optimal strategy. Finally, to make our findings relevant to the biomedical deep learning field, implementations of the combined methods that allow for quick and easy application need to be provided in an open source implementation.

Authors' contributions

The idea of this work was generated by SSB and DW. ABQ implemented the code and conducted experiments with supervision of SSB and DW. SSB, ABQ, DW and CM wrote the manuscript with FS. SSB created the figures with ABQ and the main storyline with CM. FS helped with the

manuscript narrative and editing. CM supervised the study. All authors have read and approved the manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862811 (RSENSE). CM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 866411). SSB is a member of the Munich School for Data Science (MUDS).

Acknowledgements

We thank Björn Menze, Tingying Peng, Christian Matek, Rudolf Matthias Hehr and Ario Sadafi (Munich) for discussions and contributing their ideas.

Software availability

The code and data used in this study can be found here https://github.com/marrlab/Med-AL-SSL

References

[1] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision. 2017. pp. 843–852.

[2] Settles B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences; 2009. Available: https://minds.wisconsin.edu/handle/1793/60660

[3] Sadafi A, Koehler N, Makhro A, Bogdanova A, Navab N, Marr C, et al. Multiclass Deep Active Learning for Detecting Red Blood Cell Subtypes in Brightfield Microscopy. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing; 2019. pp. 685–693.

[4] Joshi AJ, Porikli F, Papanikolopoulos N. Multi-class active learning for image classification. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 2372–2379.

[5] Gal Y, Islam R, Ghahramani Z. Deep Bayesian Active Learning with Image Data. arXiv [cs.LG]. 2017. Available: http://arxiv.org/abs/1703.02910

[6] Ducoffe M, Precioso F. QBDC: Query by dropout committee for training deep supervised architecture. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1511.06412

[7] Holub A, Perona P, Burl MC. Entropy-based active learning for object recognition. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2008. pp. 1–8.

[8] Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat Mach Intell. 2019;1: 538–544.

[9] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017. Available: http://www.nature.com/doifinder/10.1038/nature21056

[10] Smailagic A, Costa P, Young Noh H, Walawalkar D, Khandelwal K, Galdran A, et al. MedAL: Accurate and Robust Deep Active Learning for Medical Image Analysis. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018. pp. 481–488.

[11] Yang L, Zhang Y, Chen J, Zhang S, Chen DZ. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. Springer International Publishing; 2017. pp. 399–407.

[12] Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2002.05709

[13] van den Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1807.03748

[14] Sagheer A, Kotb M. Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. Sci Rep. 2019;9: 19038.

[15] Newell A, Deng J. How Useful is Self-Supervised Pretraining for Visual Tasks? arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2003.14323

[16] Jing L, Tian Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. IEEE Trans Pattern Anal Mach Intell. 2020;PP. doi:10.1109/TPAMI.2020.2992393

[17] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1705.02315

[18] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1711.05225

[19] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding Transfer Learning for Medical Imaging. arXiv [cs.CV]. 2019. Available: http://arxiv.org/abs/1902.07208

[20] Holmberg OG, Köhler ND, Martins T, Siedlecki J, Herold T, Keidel L, et al. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nature Machine Intelligence. 2020;2: 719–726.

[21] Sohn K, Berthelot D, Li C-L, Zhang Z, Carlini N, Cubuk ED, et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2001.07685

[22] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv [cs.NE]. 2017. Available: http://arxiv.org/abs/1703.01780

[23] Blasi T, Hennig H, Summers HD, Theis FJ, Cerveira J, Patterson JO, et al. Label-free cell cycle analysis for high-throughput imaging flow cytometry. Nat Commun. 2016;7: 10256.

[24] Matek, C., Schwarz, S., Marr, C., & Spiekermann, K. A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls (AML-Cytomorphology_LMU). In: The Cancer Imaging Archive (TCIA) [Internet]. [cited 29 Oct 2019]. Available: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61080958

[25] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26: 1045–1057.

[26] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018;5: 180161.

[27] Codella NCF, Gutman D, Emre Celebi M, Helba B, Marchetti MA, Dusza SW, et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1710.05006

[28] Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: Dermoscopic Lesions in the Wild. arXiv [eess.IV]. 2019. Available: http://arxiv.org/abs/1908.02288

[29] Eulenberg P, Köhler N, Blasi T, Filby A, Carpenter AE, Rees P, et al. Reconstructing cell cycle and disease progression using deep learning. Nat Commun. 2017;8: 463.

[30] Kendall A, Gal Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. pp. 5574–5584.

[31] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15: 1929–1958.

[32] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. International Conference on Machine Learning. 2016. pp. 1050–1059.

[33] Hanin B, Rolnick D. How to Start Training: The Effect of Initialization and Architecture. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2018. pp. 571–581.

[34] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference. 2010. Available: http://www.jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf?source=post_page----------

[35] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision. 2015. pp. 1026–1034.

[36] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[37] Ferreira MF, Camacho R, Teixeira LF. Using autoencoders as a weight initialization method on deep neural networks for disease detection. BMC Med Inform Decis Mak. 2020;20: 141.

[38] Sajjadi M, Javanmardi M, Tasdizen T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. arXiv [cs.CV]. 2016. Available: http://arxiv.org/abs/1606.04586

[39] Lee D-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. Workshop on challenges in representation learning, ICML. 2013. Available: https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf

[40] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. pp. 770–778.