Systematic comparison of incomplete-supervision approaches for biomedical image classification

Sayedali Shetab Boushehri[*1,2,4,6] ALI.BOUSHEHRI@ROCHE.COM
Ahmad Bin Qasim[*1,2,5] AHMAD.QASIM@TUM.DE
Dominik Waibel[1,2,3] DOMINIK.WAIBEL@HELMHOLTZ-MUENCHEN.DE
Fabian Schmich[6] FABIAN.SCHMICH@ROCHE.COM
Carsten Marr[†1,2] CARSTEN.MARR@HELMHOLTZ-MUENCHEN.DE

[1] *Institute of AI for Health, Helmholtz Munich, Germany*
[2] *Institute of Computational Biology, Helmholtz Munich, Germany*
[3] *Technical University of Munich, School of Life Sciences, Germany*
[4] *Technical University of Munich, Department of Mathematics, Germany*
[5] *Technical University of Munich, Department of Informatics, Germany*
[6] *Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Germany*

Editors: Under Review for MIDL 2022

Abstract

Deep learning based image classification often requires time-consuming and expensive manual annotation by experts. Incomplete-supervision approaches including active learning, pre-training, and semi-supervised learning have thus been developed and aim to increase classification performance with a limited number of annotated images. Up to now, these approaches have been mostly benchmarked on natural image datasets, which differ fundamentally from biomedical images in terms of color, contrast, image complexity, and class imbalance. We therefore analyzed the performance of combining seven active learning, three pre-training, and two semi-supervised methods on exemplary, fully annotated biomedical image datasets covering various imaging modalities and resolutions. For each method combination, the training started with using only 1% of labeled data. We increased the labeled training data by 5% iteratively, evaluating the performance with 4-fold cross-validation in each cycle. The results showed that the pre-training methods ImageNet and SimCLR in combination with pseudo-labeling as the training strategy dominate the best performing combinations, while no particular active learning algorithm prevailed. For three out of four datasets, these combinations reached over 90% of the fully supervised results by only adding 25% of labeled data. An ablation study showed that pre-training and semi-supervised learning contributed up to 25% increase in macro F1-score in each cycle. In contrast, state-of-the-art active learning algorithms contributed less than 5% increase of macro F1-score in each cycle. Based on the result of our study, we suggest employing pre-training and an appropriate incomplete-supervision training strategy for biomedical image classification when a limited number of annotated images is available. We believe that our study is an important step towards annotation-scarce and resource-efficient model training for biomedical classification challenges.

Keywords: incomplete-supervision, biomedical imaging, deep learning, active learning, pre-training, transfer learning, self-supervised learning, semi-supervised learning.

[*] Contributed equally
[†] Corresponding author

1. Introduction

Recent successes of deep learning methods rely on large amounts of well-annotated training data (Tan et al., 2018). However, annotations for biomedical images are often scarce as they crucially depend on the availability of trained experts, whose time is expensive and limited. Many biomedical image classification tasks can be categorized as incomplete-supervision approaches, where labeled data is limited while unlabeled data is abundant (Zhou, 2017; Blasi et al., 2016). From this perspective, there are three directions to take when facing an incomplete-supervision problem:

(i) Active learning algorithms address the issue by finding the most informative instances for further annotation (Joshi et al., 2009; Ren et al., 2020; Settles, 2009) and have been benchmarked extensively on natural image datasets (Ash et al., 2019; Ducoffe and Precioso, 2015; Gal et al., 2017; Holub et al., 2008; Killamsetty et al., 2020; Sener and Savarese, 2017; Wei et al., 2015; Yoo and Kweon, 2019a,b).

(ii) Pre-training methods such as transfer learning and self-supervised learning can help to optimize the network performance with small number of labeled images (Chen et al., 2020a; Newell and Deng, 2020; van den Oord et al., 2018; Sagheer and Kotb, 2019). In transfer learning, a neural network uses the representation from another model, ideally trained on a similar dataset. A common transfer learning approach, also used in many biomedical applications, is to initialize a model with pre-trained ImageNet weights (Rajpurkar et al., 2017; Wang et al., 2017). In self-supervised learning, a representation without any labels is learned (Jing and Tian, 2020). Pre-training based on self-supervised learning for medical image analysis recently has been studied by (Ericsson et al., 2020; Taher et al., 2021), where a variety of methods are compared.

(iii) Semi-supervised learning leverages unlabeled data in addition to labeled data during training, to increase the performance as well as the stability of predictions (Sohn et al., 2020; Tarvainen and Valpola, 2017).

Methodological improvements of these three approaches are mostly benchmarked on natural image datasets. Biomedical image datasets however differ from natural images in a couple of important characteristics: They are often strongly imbalanced, typically less diverse in terms of shapes and color range, and classes are often distinguished by only small feature variations, e.g., in texture and size (Esteva et al., 2017; Matek et al.). Besides that, biomedical images are different among different domains as well as experiments.

In this paper, we address the following questions: Do approaches that perform well on natural image datasets show the same performance on biomedical image datasets? Which one of the three incomplete-supervision approaches work best on biomedical images? What is the best combination of these approaches which work on biomedical images?

Thus we performed a systematic comparison on different incomplete-supervision approaches, including seven active learning algorithms (plus random sampling), three pre-training methods (plus random initialization), and two training strategies (plus supervised learning), on four exemplary biomedical imaging datasets. We compared each approaches as well as their combinations on each dataset. Then we analyzed the contribution of each approach for the top combinations. Finally, we recommended a combination of approaches for dealing with similar biomedical classification tasks.

2. Biomedical Imaging Datasets

We have selected four exemplary, publicly available, and fully annotated datasets from the biomedical imaging field to evaluate the efficiency and performance of active learning algorithms, pre-training methods, and training strategies (see Figure 1). For more information about the datasets, refer to Appendix A:

(a) The white blood cell dataset contains 18,395 microscopic images of single stained human leukocyte cells in ten classes (Matek et al., 2019b,a).
(b) The skin lesion dataset contains 25,339 dermoscopy images from eight skin cancer classes, which can be used for melanoma diagnosis (Codella et al., 2017; Combalia et al., 2019; Tschandl et al., 2018).
(c) The cell cycle dataset comprises 32,273 images of Jurkat cells in seven different cell cycle stages created by imaging flow cytometry (Eulenberg et al., 2017).
(d) The diabetic retinopathy dataset consists of 3,672 color fundus retinal photography images classified into five stages of diabetic retinopathy (APTOS, 2019)

3. Results

3.1. Experimental setup

We randomly selected 1% of data from each dataset as our initial annotated set and trained a ResNet18 (He et al., 2016). In each cycle, we added 5% of annotated data as suggested by one of the seven active learning algorithms (including BADGE (Ash et al., 2019), learning loss (Yoo and Kweon, 2019a), augmentation-based (Sadafi et al., 2019), Monte Carlo dropout (Gal et al., 2017), entropy-based (Settles, 2009), margin confidence (Zhou and Sun, 2014) and least confidence (Culotta and McCallum, 2005)) or randomly sampled 5% as a baseline. This process was repeated eight times leading to eventually adding 40% (and using 41%) of annotated data in total. We combined active learning with three different pre-training methods (ImageNet (Raghu et al., 2019b), autoencoder (Goodfellow et al., 2016) and SimCLR (Chen et al., 2020a)) and random initialization as baseline and two different training strategies (FixMatch (Sohn et al., 2020) and pseudo-labeling (van Engelen and Hoos, 2020)) with supervised learning as baseline resulting in $4 \times 8 \times 4 \times 3 \times 4 \times 9 = 13,824$ independent experiments (see Figure 2). We performed a 4-fold cross-validation in each cycle and calculated macro F1-score, accuracy, precision, and recall. The macro F1-score was used as our main metric of comparison, defined as the average F1-score over all classes, thus accounting for the imbalanced nature of the datasets. To quantitatively compare different combinations, we looked at the average macro F1-score across all cycles. Moreover, every combination is reported in the form of "active learning algorithm + pre-training method + training strategy". (For more information about the methods, refer to Appendix B)

3.2. Experiments

Learning loss + SimCLR + pseudo-labeling on the white blood cell dataset (see Figure 3a) achieved the highest average macro F1-score of 0.71±0.07 (mean±standard deviation on n=8 cycles). BADGE + ImageNet + pseudo-labeling on the skin lesion dataset achieved the

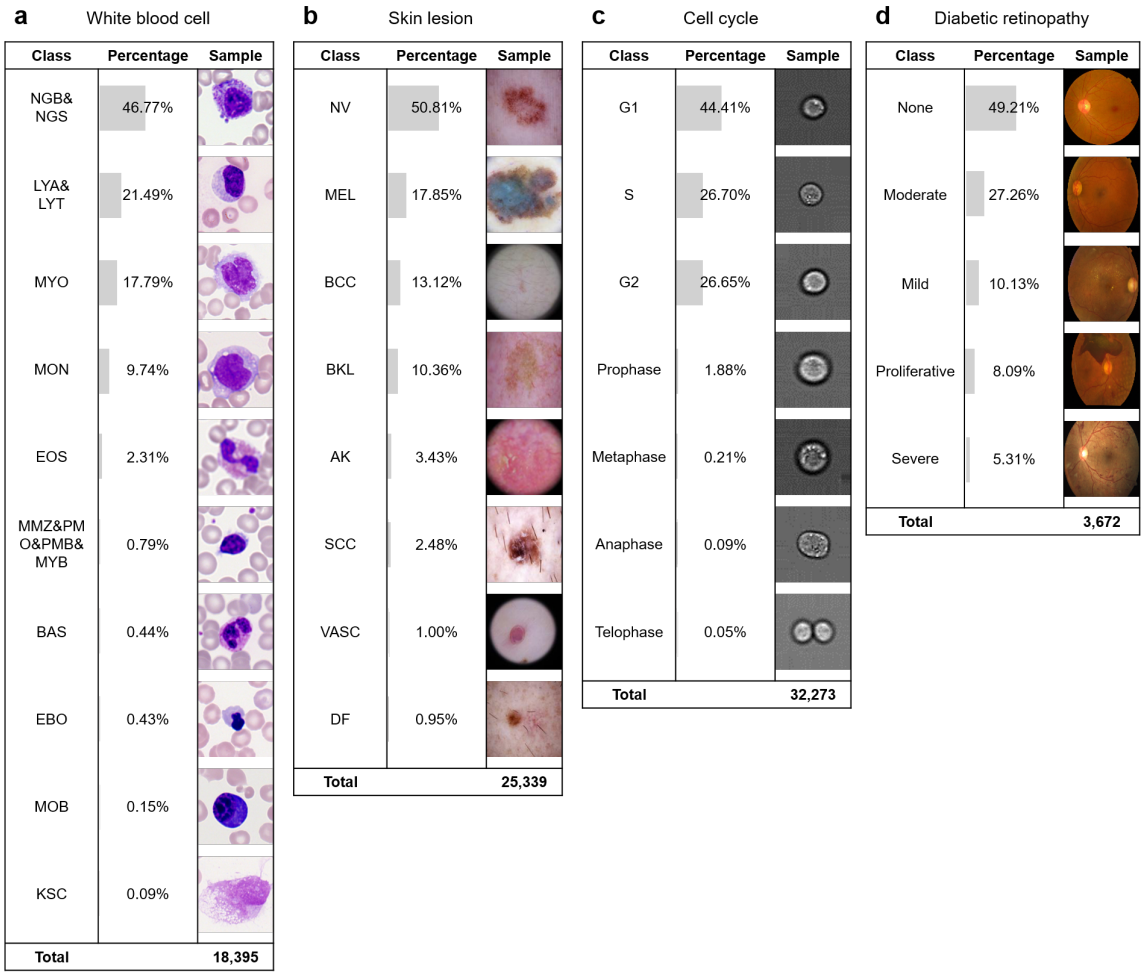

Figure 1: The four selected biomedical image datasets exhibit strong class imbalance, little color variance and high similarity among classes.

highest average macro F1-score (0.56±0.09). BADGE + ImageNet + pseudo-labeling on the cell cycle dataset achieved the highest average macro F1-score (0.54±0.08). Augmentation-based + ImageNet + FixMatch on the diabetic retinopathy dataset achieved the highest average macro F1-score (0.54±0.08).(see Figure 3).

In almost all cases (17 out of 20), the top-5 combinations were the ones that performed well from the first cycle where no active learning is involved. ImageNet and SimCLR pre-training, as well as pseudo-labeling, were always in the top combinations. Furthermore, no active learning algorithm showed up in the best combinations consistently. Finally, BADGE+ImageNet+pseudo-labeling was the top combination on two different dataset.

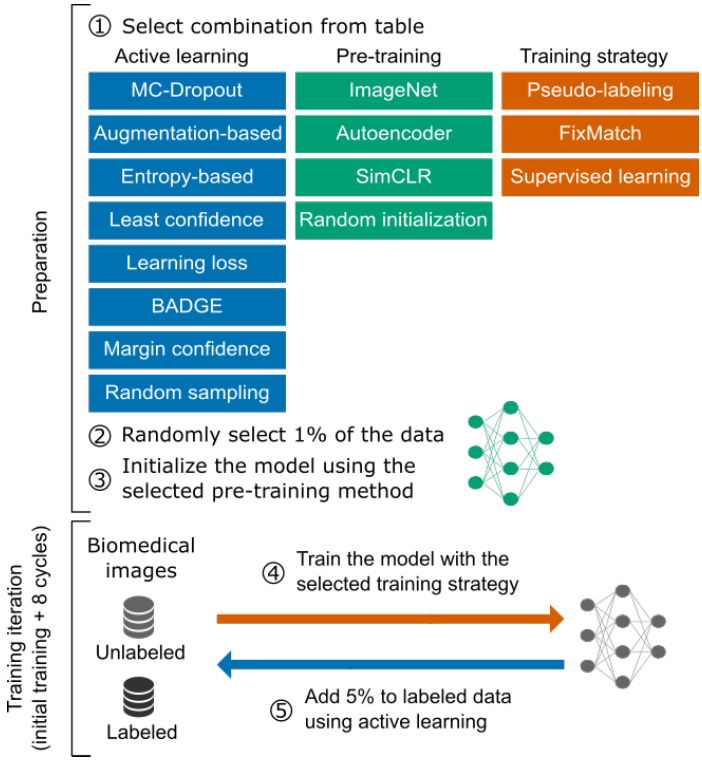

Figure 2: We systematically compared combinations of different incomplete-supervision approaches on biomedical imaging datasets. Specifically, we ran $4 \times 8 \times 4 \times 3 \times 4 \times 9 = 13,824$ independent experiments (4 datasets, 7 active learning algorithms + 1 baseline, 3 pre-training methods + 1 baseline, 2 training strategies + 1 baseline, 4-fold cross-validation and 1 initial step + 8 active learning cycles) to identify the best out of 96 possible combinations.

3.3. Ablation study

To better understand each approach's contribution to the performance, we selected the top combination for each dataset (see Figure 3) and conducted a systematic ablation study. We define the contribution to the performance of each incomplete-supervision approach by calculating the difference in F1-score if that approach was substituted with its baseline: active learning algorithms were substituted with random sampling, pre-training methods with the random initialization, and training strategies with supervised learning. The analysis showed that the contribution of pre-training and semi-supervised learning can reach up to 25% increase in macro F1-score. In contrast, the active learning algorithms contributed up to only 5% increase in macro F1-score in each cycle (see Figure 4).

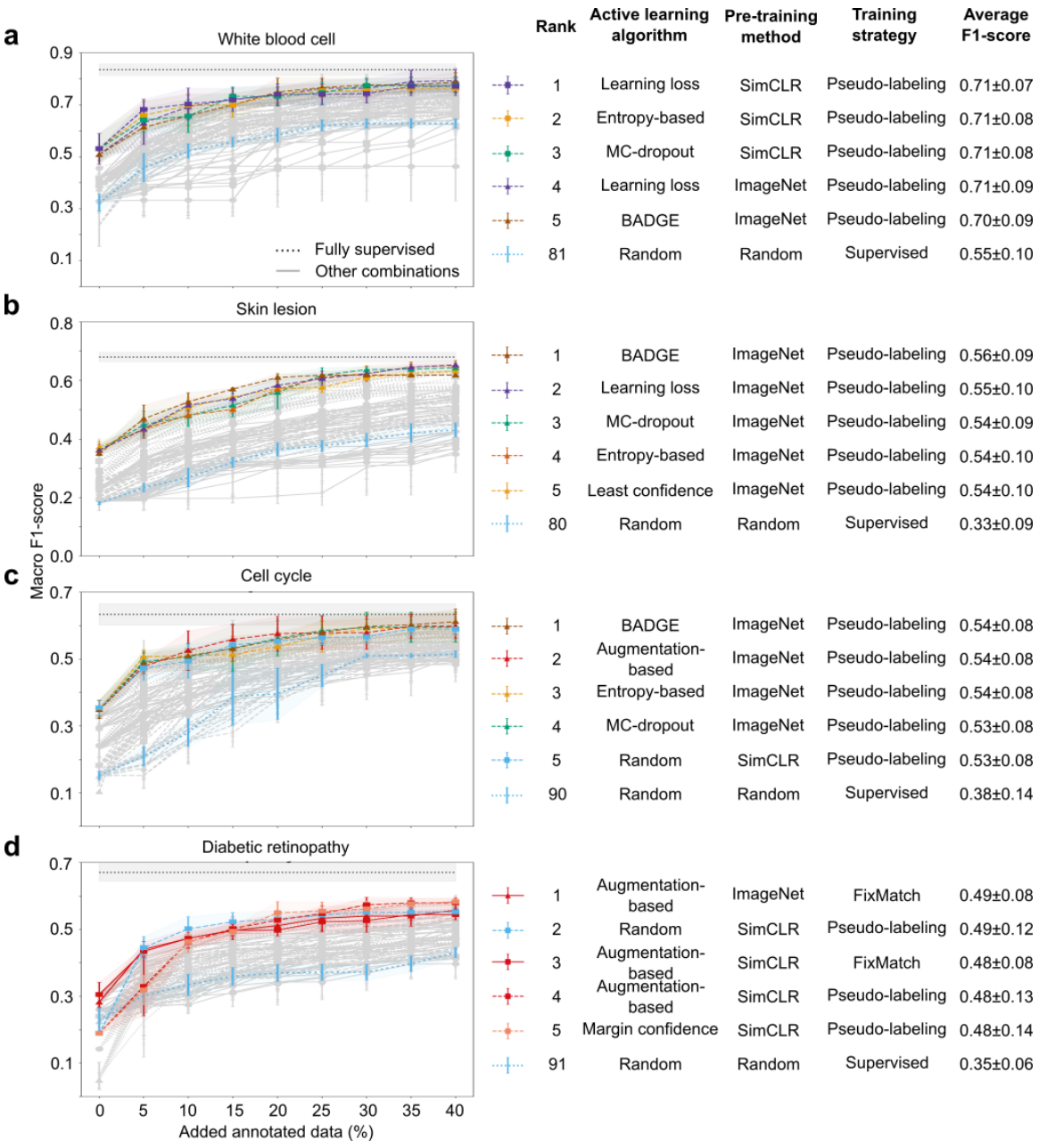

Figure 3: ImageNet and SimCLR as pre-training methods and pseudo-labeling as the training strategy dominate the best performing combinations, while no particular active learning algorithm prevails. In each panel (a-d) the upper bound of performance is fully supervised learning (black dotted line).The grey lines are combinations which did not achieve the top-5 rank. The baseline is plotted in light blue.

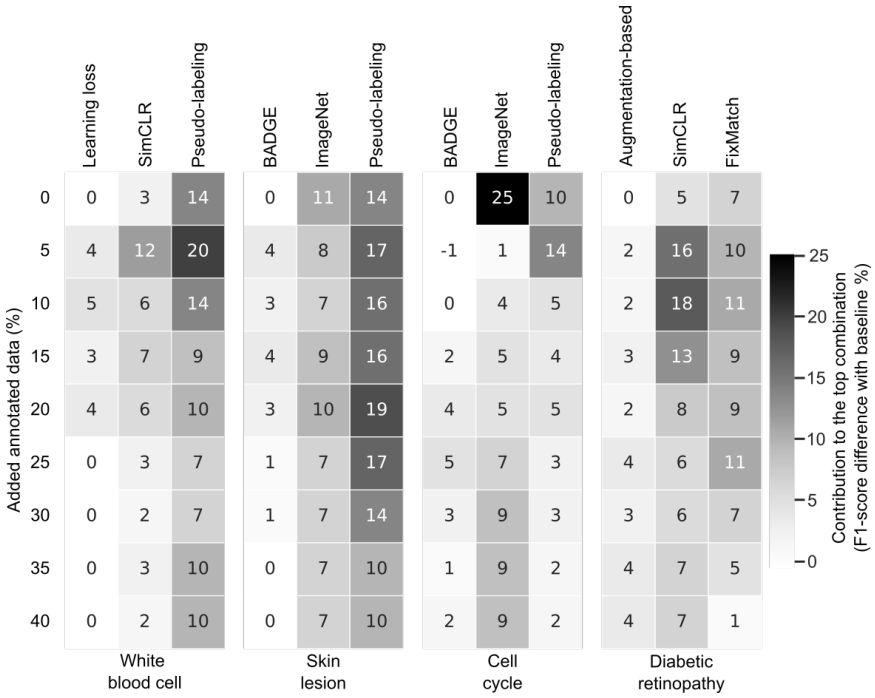

Figure 4: Semi-supervised learning and pre-training contribute stronger to the top performing combination in comparison to active learning. For every dataset, the top combination of active learning algorithm, pre-training method, and training strategy is used (see Figure 3). The contribution to performance of each approach is calculated by substituting it with its baseline and subtracting the obtained macro F1-score from the original.

4. Discussion

We have investigated how incomplete-supervision approaches can increase performance on sparsely labeled biomedical datasets. With a systematic study over seven active learning algorithms, three pre-training methods, and two training strategies, as well as their baselines, we have studied how these approaches work on biomedical imaging datasets.

Our analysis showed that combining active learning algorithms, pre-training methods, and semi-supervised learning strategies leads to superior performance as compared to their baselines for every biomedical imaging datasets. We found that the contribution of pre-training and semi-supervised learning can reach up to 25% increase in macro F1-score. In contrast, we observed that the state-of-the-art active learning algorithms contribute up to 5% increase in macro F1-score in each cycle. Therefore, we recommend investing in time and resources on semi-supervised learning strategy and pre-training methods as the identified best approaches, instead of finding the appropriate active learning while working on biomedical imaging datasets.

In addition, we found that high performance on natural images does not guarantee the same quality on biomedical images. This can be due to the fact that algorithms such as

FixMatch, carry implicit assumptions about the data distribution, and need to be adapted to the new domain during the implementation phase.

In terms of implementation of active learning algorithms, all of them were easy to implement by following definition or provided public codes, except learning loss, as it brought changes in the architecture, loss function, and training. For pre-training methods, the same applies. However, SimCLR needed large batch sizes (> 2048), which led to memory problems during the execution. Regarding the training strategies, tuning the optimal hyperparameters for FixMatch was difficult and the default parameters did not work. Also, the only combinations which showed lower performance than the baseline were the ones which included FixMatch as the training strategy. In terms of run-time, pseudo-labeling required slightly more than supervised learning, but FixMatch took at least three times more than supervised learning in every case (see Appendix C).

Based on the implementation and numerical considerations, we suggest that the combination of BADGE (active learning), ImageNet initialization (pre-training), and pseudo-labeling (training strategy) can be considered as a good choice for dealing with problems where annotated data is limited. For three datasets, this combination reached more than 90% of the fully supervised results by only using 25% of the labeled data (see Figure 5).

While the selected approaches might seem somewhat arbitrary, they are selected in such a way that they cover a wide range of methodologies. Among active learning algorithms there are different ways of estimating uncertainty (in BADGE via gradients of the network and clustering; in learning loss via separate loss function; in Augmentation-based via input perturbation; in MC-dropout via model perturbation, Margin sampling, and least confidence; and in entropy based via softmax output). For the pre-training methods, the weights are calculated differently in each method (ImageNet: pre-training on natural images; autoencoder: learning efficient codings of unlabeled data by attempting to regenerate the input; SimCLR: using positive and negative samples to train a network with a contrastive loss function). For the semi-supervised learnings the same applies (FixMatch: using strong and weak augmentations and using a contrastive loss function; pseudo-labeling: using the softmax values for generating pseudo-labels)

Due to the computational costs, we used a fixed architecture and a fixed set of parameters. While this choice might not lead to the best fully supervised performance for each dataset (e.g., compared to much bigger architectures or series of ensemble learners used for white blood cell [(Matek et al., 2019b) $ROC = 0.99$], cell cycle [(Eulenberg et al., 2017) $accuracy = 0.99$], and diabetic retinopathy [APTOS 2019 $\kappa^2 = 0.93$]), it provides a framework to systematically analyze the combination of incomplete-supervision approaches. Based on the work of (Chen et al., 2020b), we also suggest testing bigger architectures to figure out if there is a correlation between the architecture size and the performance for biomedical data. Finally, based on the recent findings (Taher et al., 2021), another question to answer in upcoming works would be to compare self-supervised learning pre-training on natural images vs. biomedical images.

We believe that our study is an important step towards helping bioinformaticians working on annotation-scarce and resource-efficient model training of biomedical image classification challenges.

References

APTOS. Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection, 2019. Accessed: 2021-11-28.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. June 2006.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. June 2019. arXiv:1906.03671.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big Self-Supervised models advance medical image classification. January 2021. arXiv:2101.05224.

Thomas Blasi, Holger Hennig, Huw D Summers, Fabian J Theis, Joana Cerveira, James O Patterson, Derek Davies, Andrew Filby, Anne E Carpenter, and Paul Rees. Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nat. Commun.*, 7:10256, January 2016.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. February 2020a. arXiv:2002.05709.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised models are strong Semi-Supervised learners. June 2020b. arXiv:2006.10029.

Noel C F Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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). October 2017. arXiv:1710.05006.

Marc Combalia, Noel C F Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. August 2019. arXiv:1908.02288.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In Manuela M Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 746–751. AAAI Press / The MIT Press, 2005.

Melanie Ducoffe and Frederic Precioso. QBDC: Query by dropout committee for training deep supervised architecture. November 2015. arXiv:1511.06412.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? *CoRR*, abs/2011.13377, 2020. URL https://arxiv.org/abs/2011.13377.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, January 2017.

Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E Carpenter, Paul Rees, Fabian J Theis, and F Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nat. Commun.*, 8(1):463, September 2017.

Mafalda Falcão Ferreira, Rui Camacho, and Luís F Teixeira. Using autoencoders as a weight initialization method on deep neural networks for disease detection. *BMC Med. Inform. Decis. Mak.*, 20(Suppl 5):141, August 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, June 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. March 2017. arXiv:1703.02910.

X Glorot and Y Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference*, 2010.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016.

Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 571–581. Curran Associates, Inc., 2018.

K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Olle G Holmberg, Niklas D Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, Johannes Schiefelbein, Siegfried Priglinger, Karsten U Kortuem, and Fabian J Theis. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2(11):719–726, November 2020.

A Holub, P Perona, and M C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2008.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, May 2020.

A J Joshi, F Porikli, and N Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, June 2009.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. December 2020. arXiv:2012.10630.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

C. Matek, S. Schwarz, C. Marr, and K. Spiekermann. A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls (AML-Cytomorphology_LMU). https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61080958. Accessed: 2019-10-29.

C Matek, S Schwarz, K Spiekermann, and C Marr. Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. *bioRxiv*, 2019a.

Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell*, 1(11):538–544, November 2019b.

G J McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *J. Am. Stat. Assoc.*, 70(350):365–369, June 1975.

Alejandro Newell and Jia Deng. How useful is Self-Supervised pretraining for visual tasks? March 2020. arXiv:2003.14323.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In H Wallach, H Larochelle, A Beygelzimer, F dAlché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3347–3357. Curran Associates, Inc., 2019a.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. February 2019b. arXiv:1902.07208.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-Level pneumonia detection on chest X-Rays with deep learning. November 2017. arXiv:1711.05225.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. August 2020. arXiv:2009.00236.

Ario Sadafi, Niklas Koehler, Asya Makhro, Anna Bogdanova, Nassir Navab, Carsten Marr, and Tingying Peng. Multiclass deep active learning for detecting red blood cell subtypes in brightfield microscopy. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 685–693. Springer International Publishing, 2019.

Alaa Sagheer and Mostafa Kotb. Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. *Sci. Rep.*, 9(1):19038, December 2019.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep Semi-Supervised learning. June 2016. arXiv:1606.04586.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A Core-Set approach. August 2017. arXiv:1708.00489.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised learning with consistency and confidence. January 2020. arXiv:2001.07685.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(56):1929–1958, 2014.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. *CoRR*, abs/2108.05930, 2021. URL https://arxiv.org/abs/2108.05930.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279. Springer International Publishing, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. March 2017. arXiv:1703.01780.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*, 5:180161, August 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. July 2018. arXiv:1807.03748.

Jesper E van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Mach. Learn.*, 109(2):373–440, February 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on Weakly-Supervised classification and localization of common thorax diseases. May 2017. arXiv:1705.02315.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1954–1963, Lille, France, 2015. PMLR.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019a.

Donggeun Yoo and In So Kweon. Learning loss for active learning. May 2019b. arXiv:1905.03677.

Jin Zhou and Shiliang Sun. Improved margin sampling for active learning. In *Pattern Recognition*, pages 120–129. Springer Berlin Heidelberg, 2014.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *Natl Sci Rev*, 5(1):44–53, August 2017.

Appendix A. Datasets

The datasets in this study originate from different fields of medicine. The nature of image sources and the technology used to record them are different.

- The white blood cell dataset contains 18,395 images (128x128x3 pixels) of single human leukocyte cells. To ensure a meaningful test set, neutrophils (segmented and band), lymphocytes (typical and atypical) and immature leukocytes (myeloblasts, promyelocytes, promyelocytes-bilobed, and myelocytes) are merged based on the class definitions (Matek et al., 2019b,a).

- The skin lesion dataset contains 25,339 dermoscopy images (128x128x3 pixels) from eight skin cancer classes (Codella et al., 2017; Combalia et al., 2019; Tschandl et al., 2018), which can be used for melanoma diagnosis. The dataset has been used in the ISIC 2018 challenge as an effort to improve melanoma diagnosis.

- The cell cycle dataset comprises 32,273 images of Jurkat cells (64x64x3 pixels) in seven different cell cycle stages created by imaging flow cytometry (Eulenberg et al., 2017). For better visualization, only the bright-field channel is shown.

- The diabetic retinopathy dataset consists of 3,672 color fundus retinal photography images (2095x2095x3) classified into five stages of diabetic retinopathy APTOS 2019 Blindness Detection. For computational reasons, the size of the images were reduced from 2095x2095x3 to 128x128x3 pixels. The dataset has been used in the APTOS 2019 Blindness Detection challenge (APTOS, 2019).

Appendix B. Methods

B.1. Architecture

We use ResNet18 (He et al., 2016) as the training architecture. For each dataset, we pretrain the ResNet18 using an autoencoder or SimCLR (Chen et al., 2020a). For more information about the implementation consideration, refer to our GitHub.

B.2. Active learning algorithms

The performance of a model f^θ with parameters θ can be increased by labeling images from the set of unlabeled images U, thus adding pairs of images and corresponding labels (x_i, y_i) to the set of labeled images L. The labeling of unlabeled images is carried out in cycles, in which s images $S \subseteq U$ with $|S| = s$ are selected for annotation and added to L, after the performance of the model converges with the previous 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 in the U are prioritized for labeling. The algorithms evaluated in this paper are based on uncertainty δ. Uncertainty δ is a scalar value which is attributed to each image in U. The s images $S \subseteq U$ with $|S| = s$ with the highest uncertainty are selected for labeling in each cycle.

B.2.1. Monte Carlo dropout (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 (Kendall and Gal, 2017; Srivastava et al., 2014) 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 (Gal and Ghahramani, 2016). During each active learning cycle, the s 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 (Gal et al., 2017).

B.2.2. 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 $U_i = \{u_{1_i}, u_{2_i}, u_{3_i}, ..., u_{J_i}\}$ with $|U_i| = J$. The random transformations are followed by a forward-pass through the model f_θ. This results in J predictions $\hat{Q}_i = \{\hat{q}_{1_i}, \hat{q}_{2_i}, \hat{q}_{3_i}, ..., \hat{q}_{J_i}\}$, where $\hat{q}_i = argmax[P^\theta(\hat{y}_i|u_i)]$ 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, it should output the same prediction for randomly augmented versions. Thus, the lower the frequency of the mode, the higher the uncertainty δ (Sadafi et al., 2019). During each active learning cycle, the images with the lowest frequency of the most frequently predicted class are annotated and added to the labeled set L.

B.2.3. Entropy-based sampling

Entropy measures the average amount of information or "bits" required for encoding the distribution of a random variable. Here, entropy is used as a criterion for active learning to select the s images $S \subseteq U$, whose predicted outcomes (softmax layer) have the highest entropy, assuming that high entropy of predictions means high model uncertainty δ. By definition, entropy focuses on taking the complete predictive distribution into account (Settles, 2009).

B.2.4. Least confidence

Least confidence sampling is the simplest and most common form of uncertainty sampling. The difference between the most confident prediction out of all class predictions (the highest softmax value) and 100% confidence is used as a metric. Hence, by selecting the s images ($S \subseteq U$) which the model is least confident about, the model performance is optimized (Culotta and McCallum, 2005).

B.2.5. Learning loss

Learning loss includes a second network, called loss prediction module, which can be added to an active learning network, namely the target model. It is trained to predict the losses of the target model on unlabeled inputs, simultaneously with the training of the target model. For the next active learning cycle, this module can be used to select images for which the target model is likely to produce a wrong prediction (Yoo and Kweon, 2019a).

B.2.6. BADGE

BADGE (Ash et al., 2019) is an active learning algorithm that selects diverse samples which incur a large magnitudinal shift in the gradient space. The model is considered to be uncertain about an image if knowing the label of the image results in a large gradient of the loss with respect to model parameters. As the labels are not known, BADGE considers the predicted labels as true labels. Secondly, to ensure that a diverse batch of images

is selected, BADGE uses the $kMEANS++$ algorithm (Arthur and Vassilvitskii, 2006). Hence, BADGE trades off between uncertainty and diversity of the s images $S \subseteq U$ which are selected for active learning.

B.2.7. Margin confidence sampling

Margin confidence sampling is similar to least confidence sampling. The difference between the most confident prediction and the second most confident prediction is used as the metric only for margin confidence sampling. The main idea is that the smaller the difference is, the higher the model uncertainty on an image. As a result, the s images $S \subseteq U$ with the least difference are selected (Zhou and Sun, 2014).

B.2.8. Random sampling (baseline)

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

B.3. Pre-training methods

Network initialization can increase the performance of neural networks (Hanin and Rolnick, 2018). It is shown to be even more essential when the amount of annotated data is not considerably large (Holmberg et al., 2020). In this work, we utilize three different pre-training methods plus random initialization (baseline).

B.3.1. ImageNet weights

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 to initialize the models, which are to be trained on other datasets (Raghu et al., 2019a). This has become a standard pre-training for classification tasks as it often helps the network to converge faster than with random initialization. Additionally, it has been shown to be beneficial in biomedical imaging (Raghu et al., 2019b).

B.3.2. Autoencoders

Autoencoders are a class of neural networks used for feature extraction (Goodfellow et al., 2016). The objective of the autoencoders is to reconstruct the input. An encoder network enc encodes the input x into its latent representation $enc(x)$. 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 dec, which aims to output a reconstruction $dec(enc(x))$ 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 imaging datasets (Ferreira et al., 2020).

B.3.3. SimCLR

SimCLR is a framework for contrastive learning of visual representations (Chen et al., 2020a). 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 pairs of differently augmented copies of the same image. Let a function perform stochastic data augmentations (such as cropping, adding color jitter, horizontal flipping, and grayscale) on a given image. Each image $x \in D$ in a mini-batch of size B is passed through the stochastic data augmentation function twice to obtain $x_i = \{x_{1_i}, x_{1_i}\}$. These pairs can be termed as positive pairs as they originate from the same image x_i. A neural network encoder $enc(x)$ extracts the feature vectors h 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. Using SimCLR as a pre-training method shows significant improvement in ImageNet classification (Chen et al., 2020a).

B.3.4. Random initialization (baseline)

It has been shown that complete random initialization performs poorly compared to more sophisticated initialization measures (Glorot and Bengio, 2010). We thus use Kaiming He initialization (He et al., 2015) (which has been shown to boost the performance) as a baseline random initialization method.

B.4. 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 two semi-supervised strategies, which incorporates the unlabeled data in the training process.

B.4.1. FixMatch

Fixmatch is a semi-supervised learning strategy which combines consistency regularization (Sajjadi et al., 2016) and pseudo-labeling (Lee, 2013). 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 or translation) and then applying pseudo-labeling on the output prediction distribution with a threshold. Another set of pseudo-labels is obtained by passing the unlabeled dataset through a strong stochastic augmentation function a_{strong} (e.g. color distortion, random noise, or 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:

$$L_{fixmatch} = L_{supervised} + \lambda L_{unsupervised} \tag{1}$$

Using FixMatch, a significant performance improvement has been observed compared to supervised training in a low-annotation regime (Sohn et al., 2020).

B.4.2. Pseudo-labeling

Pseudo-labeling is a semi-supervised learning strategy (van Engelen and Hoos, 2020). It involves training a base learner on $L = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ with $|L| = N$ as well as $U = \{u_1, u_2, ..., u_K\}$ with $|U| = K$, for which the labels are acquired through pseudo-labeling (McLachlan, 1975). The training process involves two steps. First, the base learner is trained on L as well as the pseudo-labeled set from previous cycles and predictions ($\hat{y}$). Second, the unlabeled images, for which the base learner outputs predictions with a high confidence, are assigned the corresponding predicted label and added to the training set as pseudo-labeled images for the next cycle.

B.4.3. Supervised learning (baseline)

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 model's performance 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,

$$Loss = -\sum_{l}^{N}\sum_{j}^{C} log(\hat{y}_{ij}) \tag{2}$$

with C being the total number of classes in the dataset and N being the size of L.

Appendix C. Approach considerations

C.1. Active learning

Our combinatorial-search showed that no single best active learning algorithm outperforms the others consistently. Even though they perform better than random sampling, the results of using learning loss, augmentation-based, BADGE, and MC-dropout are dataset dependent. In terms of implementation, all of the methods were straightforward except learning loss, as it brought changes in the architecture, loss function, and implementation.

C.2. Pre-training

Regarding pre-training methods, ImageNet and SimCLR led consistently to top results, while autoencoder pre-training did not prove to be effective. After close inspection of all combinations (Top-5 combination shown in Figure 2), we observed SimCLR to be more effective than ImageNet in combination with supervised learning. This observation is in alignment with recent papers where SimCLR and other self-supervised methods outperform ImageNet on biomedical applications (Raghu et al., 2019a; Azizi et al., 2021; Holmberg et al., 2020). However, in our analysis ImageNet and SimCLR pre-training performed comparatively similarly when being combined with a semi-supervised method. As semi-supervised

learning strategies incorporate unlabeled data in their training, using self-supervised methods is redundant. In terms of implementation, SimCLR implementation was straightforward but needed large batch sizes (> 2048), which was cumbersome during the execution.

C.3. Training strategy

Semi-supervised learning outperformed supervised learning in all cases as it exploits unlabeled data during training. In particular, pseudo-labeling was the top choice for all datasets, while FixMatch only performed well for the diabetic retinopathy dataset. The augmentations in FixMatch are not designed for biomedical images (see Methods), which could have worsened the performance. While the pseudo-labeling implementation was straightforward, tuning the optimal hyperparameters for FixMatch was difficult. In terms of run-time, pseudo-labeling required slightly more more training time than supervised learning, but FixMatch took at least three times more than supervised learning in every case (see Figure 5a).

C.4. Optimal combination

As a result of this work, we recommend an annotation and resource-efficient strategy for biomedical imaging active learning tasks. We propose that the combination of BADGE (active learning), ImageNet initialization (pre-training), and pseudo-labeling (training strategy) can be considered as a stable choice for dealing with problems where annotated data is limited. For three datasets, this combination reached at least 90% of the fully supervised results by only using 25% of the labeled data (see Figure 5b).

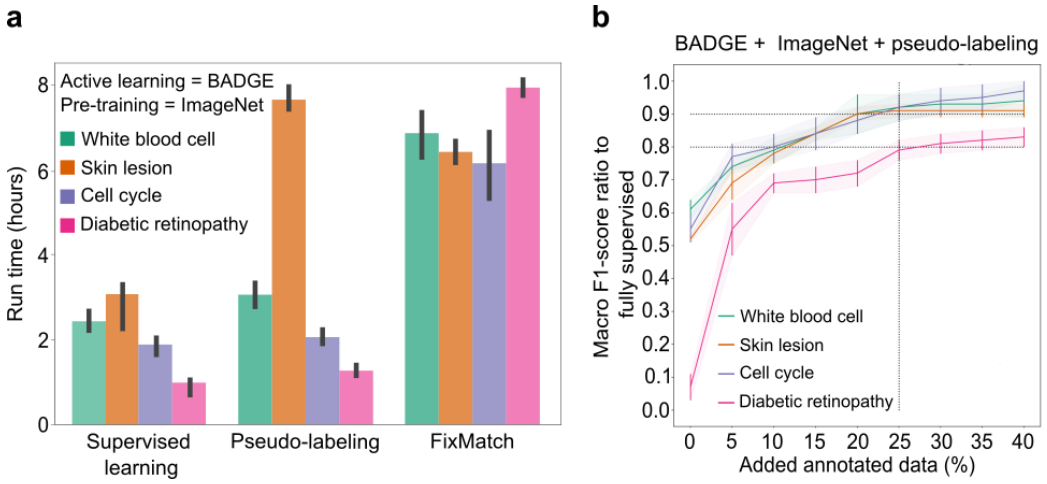

Figure 5: The annotation and resource-efficient combination, BADGE + ImageNet + pseudo-labeling, reaches above 90% of the fully supervised result in three out of four biomedical datasets by using only 25% of annotated data.

For three out of four datasets, the top combinations reached more than 90% in macro F1-score of the fully supervised approach (see Figure 3 and Figure 5b), with only 26% of the data

being labeled (1% randomly selected and 25% added in 5 active learning cycles). Notably, this was not the case for the diabetic retinopathy dataset, where the top combination still lacked 12% from the fully supervised results with using 41% of the labeled data. One reason might be image resolution: For computational reasons, we had to reduce the height and width of the images from 2095x2095 to 128x128 pixels, which might have contributed to misclassifications between the 'proliferate' and 'severe' classes (data not shown).

Appendix D. Additional Information

D.1. Acknowledgment

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

Data and Software availability: All scripts and how to access and process the data can be found here: https://github.com/marrlab/Med-AL-SSL.

Acknowledgment: We thank Björn Menze, Tingying Peng, Christian Matek, Melanie Schulz, Rudolf Matthias Hehr, Lea Schuh, Valerio Lupperger, and Ario Sadafi (Munich) for discussions and for contributing their ideas.

D.2. Author contributions

ABQ implemented code and conducted experiments with supervision of SSB and DW. SSB, ABQ, DW, and CM wrote the manuscript with FS. SSB created 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.

D.3. Funding

SSB has received funding by F. Hoffmann-la Roche LTD (No grant number is applicable) and supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS". CM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 866411).

D.4. Data and Software availability

All scripts and how to access and process the data can be found here: https://github.com/marrlab/Med-AL-SSL.