ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Investigating the Impact of Imbalanced Medical Data on the Performance of Self-Supervised Learning Approaches

Manuel Laufer Felicitas Brokmann Dominik Mairhöfer
Erhardt Barth
Thomas Martinetz

University of Lübeck - Institute for Neuro- and Bioinformatics
Ratzeburger Allee 160, 23562 Lübeck - Germany

Abstract. In clinical practice, a substantial amount of data is generated on a daily basis for diagnostic purposes. Since expensive expert knowledge is required for data annotation in order to use this data for supervised learning, large amounts of data often remain unused. Self-supervised learning methods are well suited for using unlabeled data by pre-training networks to solve pretext tasks. As medical data follow an underlying uneven distribution of occurring diseases, they are inherently imbalanced. This could introduce an unwanted bias during pre-training, ultimately leading to negative consequences that may inhibit the benefits of fine-tuning. In this work we investigate the impact of the imbalance of 2D and 3D medical datasets used for pre-training, as well as the importance of the type and size of the dataset used for pre-training and the pretext task. Our findings indicate that the size of the dataset used for pre-training has greater impact on the final tasks than its balance.

1 Introduction

Since medical imaging often is the basis for diagnosis and therapy, the amount of data acquired is growing on a daily basis. This data can be used for the supervised training of deep networks that can support the physician in his decisions. However, this training requires labels that are assigned by medical experts and are therefore cost-intensive. This results in labeled medical datasets being relatively small compared to datasets of e.g. natural images, despite the large amount of data available. In order to still be able to benefit from these unlabeled data, self-supervised learning methods (SSL) can help. These SSL approaches use so-called pretext tasks to extract prior knowledge from the unlabeled data itself in the form of representative features (pre-training), which then can be used to initialize downstream, problem-specific tasks (fine-tuning). The SSL principle can thus address the dependency on large amounts of labeled data successfully. Due to the availability of large amounts of data of which only a few are labeled, SSL methods are the subject of recent research [1].

However, using unlabeled data for pre-training comes with a potential downside: the distribution of the data is unknown. Especially medical data are inherently imbalanced, due to the distribution of diseases in the population. While the negative consequences of imbalanced class distributions in the training data

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

have already been extensively researched in the field of supervised learning [2], there is little work that has dealt with the robustness of SSL methods against the imbalance of data used for pre-training. In order to investigate whether this imbalanced pre-trainings data distribution induces a potentially unwanted bias that has a negative impact on the benefits of fine-tuning, and to provide recommendations for the application of SSL methods with imbalanced medical data in realistic medical scenarios, we addressed the following questions: *What influence do the type, size and the imbalance of datasets used for pre-training have on the performance of a downstream classification task? How does the pretext task affect improvements of classification performance due to fine-tuning?*

For this, we varied both the size and the class label distribution of the following three 2D and 3D medical datasets: EyePACS [3], LUNA2016 [4] and STOIC [5]. In this paper, we used Relative Patch Location (RPL) [6] as a pretext task because of its simplicity. In order to evaluate the RPL pretext task with regard to its performance and its usefulness for generating meaningful features, a pre-training with randomly selected labels was used for comparison.

2 Related Work

Due to the different characteristics of medical images compared to natural images, there is a lot of research on how well existing pretext tasks can be applied to medical datasets [7, 8] and the development of new pretext tasks tailored to medical applications [9]. In addition to comparing different pretext tasks, Zhang et al. [8] also investigated the influence of the imbalance of datasets used for fine-tuning. They found that SSL pre-training has a negative impact on downstream tasks when the dataset used for fine-tuning is highly imbalanced. Using a theoretical argument, they also postulate this negative influence on a potential imbalance in the dataset used for pre-training. Liu et al. [10] showed that contrastive SSL methods are more robust to pre-training data imbalance in case of natural images than supervised approaches. However, the comparison is made with supervised pre-training, which is rarely possible in a medical context. Nevertheless, a small negative influence of the imbalanced data used for pre-training is visible in the contrastive SSL methods. However, to the best of our knowledge, there is currently no research that addresses whether the inherent imbalance of medical datasets used for pre-training has an impact on the performance of a downstream task for a realistic clinical scenario.

3 Datasets and Experiments

For our experiments, we used the following datasets: EyePACS [3] consists of 35126 2D retinal images for diabetic retinopathy (DR) classification. This size is rare for a medical dataset. Each image is labeled based on the severity of the DR in a range from 0 (no DR) to 4 (proliferative DR). LUNA2016 [4] and STOIC [5] are both 3D chest CT datasets. While LUNA2016, consisting of 888 CTs, was labeled for lung nodule classification, the 2000 CTs from STOIC were

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

labeled for classification of COVID-19. Both are binary classification problems.

Since contrastive self-supervised learning approaches frequently used in the natural image domain, such as MoCo [11], usually require larger batch sizes, we have used Relative Patch Location (RPL) [6] as a predictive SSL approach, as it is easier to train and achieves similar results on medical datasets [7, 8]. In our case, RPL divides the image/CT into 9/27 patches of equal size and passes a randomly selected patch to the network alongside the center patch. The task of the network as pretext task is to predict the position of the randomly selected patch as a position index relative to the center patch. In this way, the focus of learning should be on semantic image content, such as structural and contextual relationships [6].

To investigate the influence of the size and imbalance of the dataset used for pre-training on the fine-tuning, the number of elements in the dataset used for pre-training was adjusted as well as the ratio of the rare and the frequent class (see Equation 1).

$$ratio = \frac{\#rare\_class}{\#frequent\_class} \tag{1}$$

Following Zhang et al. [8], in the case of a multi-class problem, the data used for pre-training was divided into healthy and non-healthy. Where the dataset size allowed, ratios were formed between *0* (only frequent class), *1* (class-balanced) and *inf* (only rare class). Note that it was not always possible to obtain the ratio *inf* with a large dataset size. To address the influence of the pretext task, we conducted pre-trainings with random labels in addition to RPL. Random Label Choice (RLC) passes patches of an image/CT to the network with a random label that is in the same range, as in the later fine-tuning task. Since this label can change after each epoch, it contrasts directly with a pretext task that is supposed to learn contextual relationships and thus serves as a further baseline. The pre-training was performed on all three datasets individually, whereby EyePACS was only subsequently fine-tuned in-domain on EyePACS and both 3D datasets were subsequently fine-tuned only on LUNA2016. In the in-domain 3D experiment (pre-training and fine-tuning on LUNA2016), due to the small size of the dataset, the dataset used for pre-training was randomly sampled from the complete dataset used for fine-tuning while preserving the original class label distribution. This is a realistic scenario, as if only a small amount of data is available, all the data would usually be used for pre-training. However, for EyePACS and STOIC, a direct separation of datasets used for pre-training and fine-tuning was possible due to the size and the type of the dataset. A split of 0.9, 0.05, and 0.05 for training, validation and test dataset respectively was performed for all datasets used for pre-training. The dataset used for fine-tuning training, validation and test split for the 3D fine-tuning on LUNA2016 was done exactly as in Zhang et al. [8] using subsets. In order to simulate a realistic medical scenario, only 1000 labeled images were randomly drawn from the EyePACS dataset for fine-tuning, but with the same original data distribution. The data split here was 0.8, 0.1, 0.1 for training, validation and test dataset, respectively. All pre-training and fine-tuning experiments were

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Table 1: Fine-tuning results for the 2D EyePACS dataset, pre-trained on EyePACS, with different pretext tasks (RPL, RLC), different imbalance ratios (nat. represents original dataset distribution), and different dataset sizes (N).

Pret. Task	Ratio	N	Kappa [%]	ACC [%]	AUC [%]	MAE
None	-	-	47.45±10.80	61.00±4.58	65.18±2.98	0.56±0.01
RPL	0.0	9048	49.17±4.86	62.67±3.51	63.28±2.60	0.54±0.05
RPL	0.3	9048	50.97±7.70	64.00±1.73	64.84±5.71	0.55±0.03
RPL	0.6	9048	50.08±3.98	62.33±8.50	64.02±3.46	0.55±0.03
RPL	1.0	9048	45.60±2.39	65.00±1.00	63.96±0.69	0.51±0.01
RPL	inf	9048	43.50±2.94	60.00±2.65	60.94±0.26	0.56±0.02
RPL	nat.	9048	51.82±1.54	59.00±3.46	63.39±3.05	0.58±0.07
RLC	nat.	9048	32.01±15.65	54.67±2.08	58.81±4.26	0.66±0.03
RPL	0.0	18096	53.75±2.74	60.00±6.56	63.99±3.31	0.53±0.05
RPL	0.3	18096	53.48±3.37	65.00±3.00	64.25±2.11	0.48±0.02
RPL	0.6	18096	52.34±4.31	64.00±2.65	65.15±3.27	0.53±0.03
RPL	1.0	18096	52.77±5.24	62.67±5.77	63.72±1.07	0.52±0.09
RPL	nat.	18096	52.58±2.63	61.33±4.16	64.46±2.19	0.54±0.04
RLC	nat.	18096	41.43±6.83	56.33±5.69	57.48±4.65	0.62±0.05

conducted three times with different seeds.

Following Zhang et al. [8] we used a 2D or 3D U-Net [12] Encoder as backbone for pre-training and fine-tuning. However, the classification head was adapted for the corresponding tasks and dimensions: 9 or 27 class classification for RPL pre-training in 2D and 3D respectively, regression and binary classification for fine-tuning on 2D and 3D datasets respectively.

4 Results and Discussion

Since the five classes of the EyePACS dataset follow a severity order, we modeled the problem as a regression. Next to the Mean Absolute Error (MAE), we measured Area Under Curve (AUC), Accuracy (ACC) and quadratic weighted Kappa, as frequently used metrics for this dataset. In Table 1 the results of the fine-tuning on the EyePACS dataset for different pretext tasks, sizes, and imbalances of the EyePACS dataset used for pre-trainings are shown. The from-scratch baseline, without any pre-training, is marked as pretext task *None*. The results show that there are indeed pre-training combinations that perform worse than no pre-training. In fact, none of the fine-tuned models has a higher AUC than the from-scratch model. However, this may be an outlier, as all other metrics are usually outperformed. Nevertheless, the imbalance of EyePACS used for pre-training does not seem to be the reason for this, as no direct correlation between imbalance and poor performance is apparent. The performance with balanced datasets used for pre-training does not deviate significantly from the performance of imbalanced datasets used for pre-training. Instead, it becomes

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Table 2: Fine-tuning results for the 3D LUNA2016 dataset, based on LUNA2016 and STOIC pre-training, with different pretext tasks (RPL, RLC), different imbalance ratios (nat. represents original dataset distribution), and different dataset sizes (N).

Pret. Task	Ratio	Pre-trained on LUNA2016			Pre-trained on STOIC		
		N	AUC [%]	ACC [%]	N	AUC [%]	ACC [%]
None	-	-	94.72±0.83	98.53±0.22	-	94.72±0.83	98.53±0.22
RPL	0.0	298	96.88±0.54	98.38±0.11	590	98.05±1.00	98.59±0.42
RPL	0.3	298	97.55±1.02	97.75±0.79	590	98.02±0.20	99.07±0.14
RPL	0.6	298	96.58±0.62	98.09±1.00	590	97.88±0.63	98.75±0.22
RPL	1.0	298	96.49±0.49	98.39±0.63	590	97.62±1.01	97.71±0.87
RPL	inf	298	97.01±0.65	98.16±0.09	590	98.13±0.21	98.88±0.15
RPL	0.0	590	97.37±0.51	98.99±0.30	1200	97.83±0.86	98.54±0.37
RPL	0.3	590	97.43±0.73	98.68±0.94	1200	97.24±0.62	98.89±0.40
RPL	0.6	590	97.59±1.46	99.06±0.16	1200	97.96±0.77	98.47±0.66
RPL	1.0	590	98.24±0.37	98.77±0.25	1200	97.55±0.64	98.71±0.52
RPL	nat.	888	98.65±0.32	99.01±0.23	2000	98.15±0.34	98.60±0.37
RLC	nat.	888	98.66±0.88	99.05±0.27	2000	98.38±1.25	98.90±0.27

clear, that both the fine-tuning with native imbalance and the averaged values of the ratios (ΔKappa=4.13 pp) demonstrate that a pre-training with 18096 elements performs better than one with 9048 elements. A comparison of the two pretext tasks shows that, RLC is significantly inferior to RPL. However, a closer look at the training curves showed that the RLC training had not yet fully converged, so that longer training times could improve the results.

The results for fine-tuning on LUNA2016 for different pretext tasks, sizes, imbalances, and types of the dataset used for pre-training are shown in Table 2. For the LUNA2016 downstream task, which is to solve a binary classification into benign and malignant lung nodules, ACC and AUC were calculated, with AUC being the more important metric, as the dataset is inherently highly imbalanced. The AUC metric indicates that any pre-trained model performs better than the from-scratch baseline. Similar to the experiments on EyePACS, more data used for pre-training has a positive effect on the benefits of fine-tuning, as shown e.g. by the AUC values for LUNA2016 at N=888 and for STOIC at N=2000. When comparing the results between pre-training on LUNA2016 and STOIC, it can be seen that they are both similar for N=590, which is probably due to the fact that both datasets consist of chest CTs, albeit containing different diseases. There is also no visible correlation between the dataset imbalance in pre-training and performance on downstream tasks for the 3D datasets. Note that pre-training with the RLC pretext task results in equal or better fine-tuning performances than using the RPL pretext task. This raises doubts as to why RPL actually works as a pretext task when it is apparently sufficient for similar or better performance that the model weights were trained on an unsolvable problem on data with random labels.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

5 Conclusion

In this paper, we investigated whether the imbalance, size and type of medical datasets used for pre-training affect the performance in downstream tasks. The results of our experiments on 2D and 3D data do not support the intuitive relationship postulated by Zhang et al. [8] between the imbalance of the dataset used for pre-training and the decline of the downstream task performance. Instead, we confirmed a clear correlation between the size of the dataset used for pre-training and performance on downstream tasks. Both findings imply that it is not recommended omitting existing pre-training data just to balance the classes. The comparatively smaller performance improvements between from-scratch models and pre-trained models for the 2D data compared to the 3D data may be due to the fact that the problem is more difficult to solve on EyePACS, as the from-scratch results show. Whether a correlation between the effectiveness of SSL and the difficulty of a downstream task exists, would be an interesting follow-up question. Furthermore, the results of our RPL and RLC experiments raise the question of the extent to which the RPL pretext task truly learns semantic relationships, which is also worth exploring in future work.

References

[1] Rayan Krishnan, Pranav Rajpurkar, et al. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 8 2022.

[2] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 12 2019.

[3] Emma Dugas, Jared, Jorge, et al. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle.

[4] Arnaud Arindra Adiyoso Setio, Alberto Traverso, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, 12 2017.

[5] Marie Pierre Revel, Samia Boussouar, et al. Study of thoracic CT in COVID-19: The STOIC project. *Radiology*, 301(1):E361–E370, 10 2021.

[6] Carl Doersch, Abhinav Gupta, et al. Unsupervised Visual Representation Learning by Context Prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[7] Aiham Taleb, Winfried Loetzsch, et al. 3D Self-Supervised Methods for Medical Imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020.

[8] Chuyan Zhang, Hao Zheng, et al. Dive into Self-Supervised Learning for Medical Image Analysis: Data, Models and Tasks. *Medical Image Analysis*, 9 2022.

[9] Yuting He, Guanyu Yang, et al. Geometric Visual Similarity Learning in 3D Medical Image Self-supervised Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, Vancouver, Canada, 6 2023.

[10] Hong Liu, Jeff Z. HaoChen, et al. Self-supervised Learning is More Robust to Dataset Imbalance. *ICLR 2022 - 10th International Conference on Learning Representations*, 10 2021.

[11] Kaiming He, Haoqi Fan, et al. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Olaf Ronneberger, Philipp Fischer, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, pages 1–8, 2015.