Leveraging BYOL for Image Classification in Limited-Labeled Medical Domains (Computer Vision)

Brendan P. Murphy
bigsur@stanford.edu

Abstract

Medical image interpretation is crucial for diagnosing and treating diseases, particularly in low-resource settings with limited access to medical expertise. However, the scarcity of labeled medical imaging data poses challenges for traditional supervised learning. Self-supervised learning methods like Bootstrap Your Own Latent (BYOL) leverage unlabeled data to enhance model performance on limited labeled data. In this study, I evaluate BYOL's effectiveness across different medical imaging modalities. Initially, a baseline ImageNet pre-trained ResNet-18 and standard BYOL pre-trained on ChestMNIST were explored. Performance evaluation was conducted on out-of-distribution datasets: PneumoniaMNIST, BreastMNIST, and DermaMNIST. I then enhanced BYOL with 3 additional augmentation techniques, this "BYOL+" model showed promise on PneumoniaMNIST, achieving an AUC of 0.86 compared to the supervised base model's 0.82. Furthermore, BYOL+ outperformed the base BYOL and supervised fine-tuned models in linear probing and few-shot learning on PneumoniaMNIST. These findings highlight the resilience of BYOL+ to distribution shifts within the same modality and its proficiency in few-shot learning scenarios. Overall, BYOL demonstrates better performance in similar modalities and comparable performance to baseline models on different tasks, showcasing the potential of self-supervised learning, specifically BYOL, in enhancing robustness across distribution shifts in medical image classification. My study expands understanding of these methods in the medical field, facilitating more accurate diagnoses, especially in resource-constrained settings.

1 Introduction

Medical image interpretation is of paramount importance in the field of medical image analysis as it enables the diagnosis and treatment of a wide range of diseases. This significance is magnified in low resource settings where access to medical expertise may be limited. However, the development of high-performance models heavily relies on training them with large medical imaging datasets. Unfortunately, obtaining large amounts of labeled data can be difficult due to the cost of expert annotation, posing a challenge for traditional supervised learning approaches.

To address this challenge, self-supervised learning methods have emerged as a promising solution for medical imaging tasks. These methods leverage extensive amounts of unlabeled data to pre-train a network and enhance its performance on limited labeled data, surpassing traditional supervised models that are typically pre-trained on datasets like ImageNet [Iyer et al., 2022, Van Uden et al., 2023]. Although no self-supervised model has consistently demonstrated top performance across all tasks, Bootstrap Your Own Latent (BYOL) has shown promising results in both medical and non-medical imaging tasks[Anton et al., 2022, Ericsson et al., 2021, Lee and Lee, 2022]. BYOL

CS230: Deep Learning, Spring 2023, Stanford University, CA.

learns a representation of the data by predicting the features of a differently augmented version of the same data[Grill et al., 2020]. However, its application in the medical field remains relatively unexplored.

In this study, my aim is to evaluate the effectiveness of BYOL in maintaining performance across different imaging modalities. To accomplish this, I will pretrain the BYOL model using the unlabeled ChestMNIST dataset. Subsequently, I will assess its performance on three out-of-distribution datasets, namely PneumoniaMNIST, BreastMNIST, and DermaMNIST. These datasets originate from distinct imaging modalities or equipment, encompass diverse patient populations, and represent various geographical regions.

By demonstrating that BYOL, when pretrained on medical images, can achieve comparable performance to benchmark models on these diverse datasets, I hope to showcase its robustness across distribution shifts. This research will contribute to expanding the knowledge and understanding of self-supervised learning methods in the medical field, and help pave the way for improved medical image interpretation and facilitate more accurate diagnoses, particularly in resource-constrained settings.

2 Prior Work

2.1 Self-supervised learning in medical imaging

Self-supervised learning has emerged as a powerful approach for medical image classification, surpassing the performance of supervised pretraining methods.[Azizi et al., 2021] Comparative studies have shown that SimCLR outperforms BYOL for pretraining backbones in left ventricle echocardiography segmentation[Saeed et al., 2022], while another study identified BYOL as the best overall performer among various self-supervised methods in medical imaging.[Anton et al., 2022] Moreover, BYOL has demonstrated its efficacy in grading diabetic retinopathy on the MedMNIST dataset [Lee and Lee, 2022]. In a recent review paper, the most commonly utilized frameworks for self-supervised learning in medical imaging were found to be SimCLR, MoCo, and BYOL, with 13, 8, and 3 papers, respectively[Huang et al., 2023]. Despite their prevalence, BYOL remains relatively unexplored, and there is no clear consensus on the best-performing self-supervised method for medical imaging. Further exploration and evaluation of BYOL's potential are warranted.

2.2 Image augmentation in self-supervised methods

Image augmentation is crucial in enhancing the quality of representations learned through self-supervision with Siamese networks. Recent studies have demonstrated the effectiveness of incorporating random contrast and brightness adjustments in generating powerful representations. These learned representations exhibit robustness to out-of-distribution data, surpass the classification accuracy of fully supervised models for various disease labels, and demonstrate generalization capabilities to unseen conditions Van der Sluijs et al. [2023] Notably, BYOL has shown greater robustness to the choice of image augmentations compared to contrastive methods, which can be attributed to its independence from negative pairs [Grill et al., 2020]. However, there is a lack of studies exploring the integration of cut out, contrast, and brightness augmentation within BYOL specifically for medical imaging. To address this gap, I will investigate the impact of these augmentation techniques on the generalization of self-supervised learning in medical imaging, particularly focusing on out-of-domain datasets.

3 Dataset

3.1 Overview

In this project, I leveraged four medical imaging datasets from MedMNIST v2, a standardized and lightweight collection specifically designed for diverse classification tasks involving biomedical images [Yang et al., 2023]. The selected datasets encompass ChestMNIST, BreastMNIST, PneumoniaMNIST, and DermaMNIST. ChestMNIST originates from the NIH-ChestXray14 dataset, which was collected in Bethesda, United States. It comprises anterior-posterior (AP) views of adult patients [Wang et al., 2017]. On the other hand, PneumoniaMNIST is derived from a hospital in China and focuses on chest X-ray (CXR) images of children aged 1 to 5 years old to evaluate the presence of pneumonia [Kermany et al., 2018]. By incorporating datasets with distinct patient populations and geographical origins, I aim to assess the resilience of the proposed approach, BYOL, against distribution shifts.

To expand the scope of the main experiment, I include two additional datasets: DermaMNIST and BreastMNIST. DermaMNIST represents a multiclass dermatology dataset with seven labels collected from Austria and Australia [Tschandl et al., 2018]. In contrast, BreastMNIST encompasses ultrasound images of adult females collected in Egypt [Al-Dhabyani et al., 2020]. These datasets offer diversity in terms of image modalities, patient populations, and geographical locations. Incorporating such variations enables us to thoroughly assess the adaptability and generalization capabilities of BYOL when confronted with out-of-domain datasets. By incorporating these diverse datasets into the study, my goal is to comprehensively evaluate the efficacy and versatility of BYOL in addressing various medical imaging scenarios.

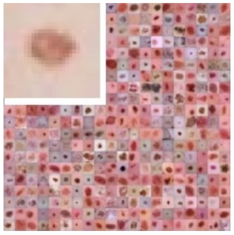


Figure 1: DermaMNIST

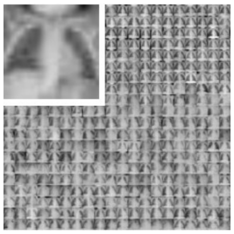


Figure 2: PneumoniaMNIST

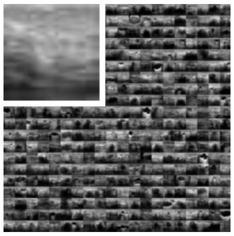


Figure 3: BreastMNIST

3.2 Pre-processing

I pre-processed the images with the torchvision transorms Grayscale function using 3 output channels since ResNet-18 expects RGB. I converted the grayscale images to 3 channels for BreastMNIST, ChestMNIST and PneumoniaMNIST datasets. However, the DermaMNIST dataset was already in RGB, so no conversion was necessary. Additionally, all image tensors were pre-processed with an element-wise normalization using mean and standard deviation of .5 for all channels to standardize the data and improve convergence and stability during training.

3.3 Dataset Breakdown

Breakdown of training, validation, and test sets with splits. Class Distribution Histograms are shown in the Appendix.

Dataset	Train	Validation	Test	Splits
DermaMNIST	7007	1003	2005	70/10/20
PneumoniaMNIST	4708	524	624	82/9/9
BreastMNIST	546	78	156	77/11/22

Table 1

4 Methods

4.1 Overview

Models and data manipulation were built using Torch [Paszke et al., 2019], Numpy [Harris et al., 2020], PyTorch Lightning [Falcon et al., 2020], and MedMNIST [Yang et al., 2023]. Architectures and models were adapted from Resnet [He et al., 2015] and BYOL [Grill et al., 2020], [Odom, 2021] and [Lucidrains, 2021]. Image augmentations relied on the Kornia library[Braga et al., 2020] as well as custom transforms. All experiments were trained on an NVIDIA A10G.

4.2 Models

1. **BYOL**

   Similar to a siamese network, BYOL utilizes two Resnet-18 same encoder networks: the "online" and "target" networks. The online network receives an input image and generates a representation while the target network, which is a moving-average of the online network's parameters, also generates a representation for the same input image. Additionally, there is an MLP "projector" that projects the encoded features to a lower-dimensional space. This projector consists of two linear layers with batch normalization and a ReLU activation function. Also of note is the stop-gradient operation (sg) which prevents the updating of the target networks weights, thus ensuring its weights are only updated from the online networks weights via the moving average process.

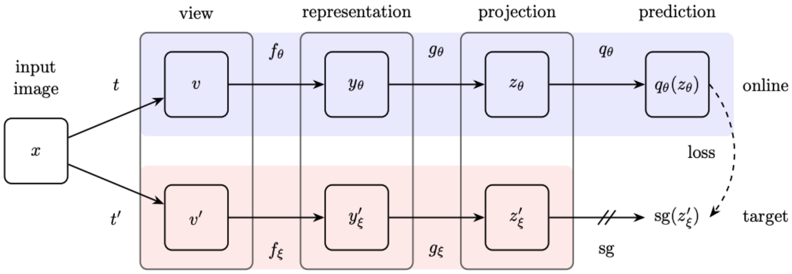


Figure 4: [Grill et al., 2020]BYOL

2. **BYOL+**

   This model is the same as above but with 3 additional augmentations: cutout, contrast, and brightness. I added these methods specifically as they have recently been shown to be effective, but have yet to been tested with BYOL.

3. **Base Model**

   The base model is an ImageNet pre-trained Resnet-18, which was separately fine-tuned on each of the three target datasets as a baseline for the experiments.

4. **Fine-Tune**

   For the primary experiment, I fine-tuned the Base and BYOL pre-trained models on each of the target datasets separately.

5. **Linear Probing**

   Tuned only the final linear layer of the BYOL pre-trained models.

6. **Few Shot**

   Fine-tuned all layers of the BYOL pre-trained models but only on a small subset of samples.

4.3 Loss Functions

The BYOL loss function is composed of a consistency loss between the two augmented views of the same image, as well as a target prediction loss between the predicted target and target projections. The consistency loss encourages the models representations of augmented views of the same image to be consistent with each other, thus learning invariant features. The target prediction loss encourages the model to predict the target representation given an augmented view. These two loss values are calculated using a normalized mean square error(MSE), they are then summed to obtain the overall BYOL loss:

```
normalized_mse(x: Tensor, y: Tensor):
    x = f.normalize(x, dim=-1)
    y = f.normalize(y, dim=-1)
    return 2 - 2 * (x * y).sum(dim=-1)

loss = torch.mean(normalized_mse(pred1,targ2) + normalized_mse(pred2,targ1))
```

By minimizing this loss, the model learns to produce consistent representations for augmented views and accurately predict the target representation, leading to improved representation learning and generalization capabilities.

The loss functions for the downstream classification tasks were PyTorch's Binary Cross Entropy with Logits and Cross Entropy Loss.

5 Experiment

5.1 Self Supervised Pre-training

For the BYOL implementation I utilized a pre-trained Resnet-18 for the encoders, Adam optimization, 1e-4 for the learning rate since I was training on a fairly large dataset, a moving average decay of .99, and a fixed batch size of 256. I trained it on the unlabelled ChestMNIST dataset for 1000 epochs which took approximately 10-12 hours. The image augmentations included random applications of color jitter, gaussian blur, resized crops, grayscale, and horizontal flipping.

For my improved BYOL+ model, I added 3 more augmentations: cutout, contrast, and brightness as these methods have recently shown to be effective,Van der Sluijs et al. [2023] but have yet to be integrated with BYOL.

5.2 Fine-tuning the Pre-trained Model

The primary experiment consisted of fine-tuning the Baseline, BYOL and BYOL+ models for an additional 25 epochs on each of the three target data sets: PneumoniaMNIST, BreastMNIST, and DermaMNIST while keeping the hyper-parameters consistent to compare their performance. I then tuned the BYOL models on the PneumoniaMNIST and BreastMNIST datasets only, as the DermaMNIST results showed little benefit from the self-supervised pre-training.

5.3 Linear Probing and Few Shot Learning

Finally, I evaluated the improved BYOL models with Linear Probing and Few Shot Learning. Linear probing highlighted the transferability of the learned representations to other medical image classification tasks. The Linear probing tasks were trained for 50 epochs and the few shot tasks were trained for 25-50 epochs on 20 samples.

5.4 Hyperparameter Tuning

The learning rate for the BYOL SSL pre-training was .0001 since I was training a complex network on a fairly large dataset. I tested pre-training with batch sizes of 32, 128, and 256 and ultimately selected 256. For the fine-tuning tasks I chose a moderately small learning rate of .001 as the datasets were fairly sized. For the few shot tasks I selected a larger learning rate of .01 as the training sample size was very small. I selected a relatively high learning rate of .01 for linear probing since i was only training the last layer. I consistently used L2 regularization by setting the weight decay to 1e-6.

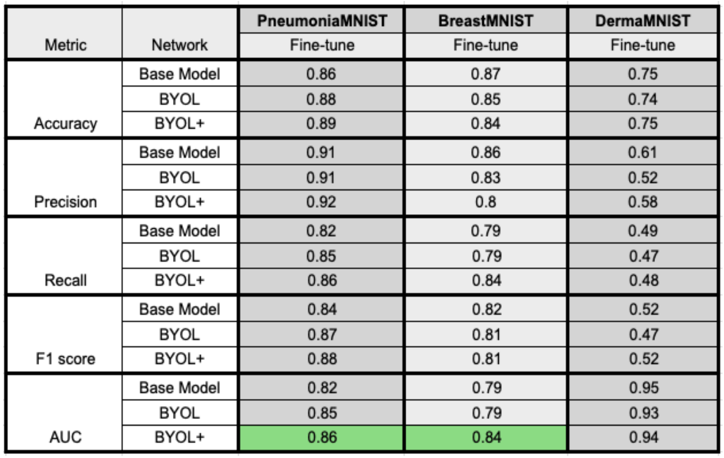
6 Performance Evaluation

In this study, I evaluated different parameter and hyperparameter settings using the BYOL, BYOL+, and downstream task models. My primary performance metric, AUC, was chosen for its robustness to class imbalance and threshold independence. The results revealed that BYOL+ consistently outperformed all other models, including the supervised fine-tuned model, on both the PneumoniaMNIST and BreastMNIST datasets. This performance advantage was evident across various parameter and hyperparameter settings.

Specifically, when examining the PneumoniaMNIST dataset, BYOL+ demonstrated superior performance not only compared to the base BYOL model but also outperformed the supervised fine-tuned model when employing linear probing and few-shot learning techniques. These results indicate that BYOL+ not only exhibits resilience to distribution shifts within the same image modality but also excels in few-shot learning scenarios. Notably, the robust performance of BYOL+ is particularly significant considering the dataset's characteristics, representing a pediatric population from diverse geographic locations.

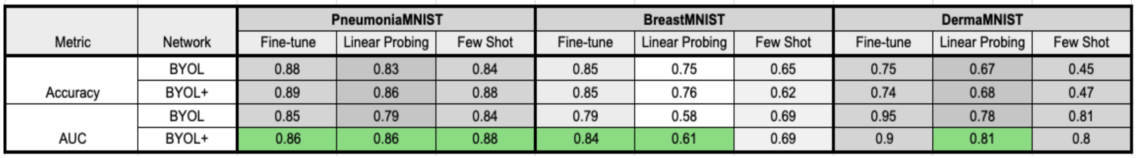
However, the robustness of BYOL+ was not consistently observed when confronted with different modalities, especially in more complex tasks like multi-class classification. In such cases, the base supervised model slightly outperformed both BYOL models with fine-tuning. Nevertheless, it is worth noting that BYOL+ still exhibited improved performance when fine-tuning was applied to the BreastMNIST dataset, representing a distinct modality. Additionally, when comparing BYOL and BYOL+ with linear probing, BYOL+ showed better performance, although not surpassing the performance achieved with fine-tuning.

These findings underscore the effectiveness of BYOL+ within the same image modality, particularly in scenarios involving distribution shifts and few-shot learning, as evidenced by the results obtained on the PneumoniaMNIST dataset. Further research is necessary to enhance its performance in more complex tasks with different modalities, such as multi-class classification.



Metric	Network	PneumoniaMNIST	BreastMNIST	DermaMNIST
		Fine-tune	Fine-tune	Fine-tune
Accuracy	Base Model	0.86	0.87	0.75
	BYOL	0.88	0.85	0.74
	BYOL+	0.89	0.84	0.75
Precision	Base Model	0.91	0.86	0.61
	BYOL	0.91	0.83	0.52
	BYOL+	0.92	0.8	0.58
Recall	Base Model	0.82	0.79	0.49
	BYOL	0.85	0.79	0.47
	BYOL+	0.86	0.84	0.48
F1 score	Base Model	0.84	0.82	0.52
	BYOL	0.87	0.81	0.47
	BYOL+	0.88	0.81	0.52
AUC	Base Model	0.82	0.79	0.95
	BYOL	0.85	0.79	0.93
	BYOL+	0.86	0.84	0.94

Figure 5: Fine-tuning of the pre-trained Base, BYOL, and BYOL+ models.



Metric	Network	PneumoniaMNIST			BreastMNIST			DermaMNIST		
		Fine-tune	Linear Probing	Few Shot	Fine-tune	Linear Probing	Few Shot	Fine-tune	Linear Probing	Few Shot
Accuracy	BYOL	0.88	0.83	0.84	0.85	0.75	0.65	0.75	0.67	0.45
	BYOL+	0.89	0.86	0.88	0.85	0.76	0.62	0.74	0.68	0.47
AUC	BYOL	0.85	0.79	0.84	0.79	0.58	0.69	0.95	0.78	0.81
	BYOL+	0.86	0.86	0.88	0.84	0.61	0.69	0.9	0.81	0.8

Figure 6: Tuned BYOL, and BYOL+ models

7 Conclusion

In this study I tested the BYOL method of self-supervised pretraining on different medical imaging tasks and compared it to the baseline ImageNet pre-trained model on the same tasks. The tasks were of varying modality and complexity which characterized the models strengths and weaknesses. My results show that self-supervised pretraining can enhance the performance of downstream medical tasks of different modalities, though it didn't generalize well to more complex tasks such as color images and highly imbalanced multi-class problems.

The insights gained from this project provide the motivation for continued exploration of different image transforms while leveraging self-supervised learning to enhance model performance. In the future I would explore the use of augmentation techniques specifically suited to medical imaging tasks but would conduct those experiments on higher resolution datasets such as CheXpert as the resolution of the medMNIST dataset isn't suitable for clinical settings. Combining self-supervised pre-training and supervised fine-tuning has the potential to unlock new levels of accuracy, generalizability and effeciency in medical imaging classification tasks.

8 Code

Link to GitHub
https://github.com/csbrendan/cs230 [Murphy, 2023]

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Jonah Anton et al. How well do self-supervised models transfer to medical imaging? *Journal of Imaging*, 8(12):320, 2022.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.

Rafael Gomes Braga, Edgar Riba, Daniel Ponsa, and João Cartucho. Kornia: an open source differentiable computer vision library for pytorch, 2020. URL `https://arxiv.org/abs/2003.12786`.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5414–5423, June 2021.

William Falcon, Jirka Borovec, Adrian Wälchli, N Eggert, J Schock, J Jordan, N Skafte, V Bereznyuk, E Harris, T Murrell, et al. Pytorchlightning/pytorch-lightning: 0.7. 6 release. *Zenodo: Geneva, Switzerland*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

CR Harris, KJ Millman, SJ van der Walt, R Gommers, P Virtanen, D Cournapeau, E Wieser, J Taylor, and S Berg. Smith 474 nj. *Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'ıo JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature*, 585(7825):357–362, 2020.

K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. arxiv preprint arxiv: 151203385, 2015.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023.

Niveditha S Iyer, Aditya Gulati, Oishi Banerjee, Cécile Logé, Maha Farhat, Agustina Saenz, and Pranav Rajpurkar. Self-supervised pretraining enables high-performance chest x-ray interpretation across clinical distributions. *medRxiv*, pages 2022–11, 2022.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Joohyung Lee and Eung-Joo Lee. Self-supervised pre-training improves fundus image classification for diabetic retinopathy. In *Real-Time Image Processing and Deep Learning 2022*, volume 12102, pages 193–198. SPIE, 2022.

Lucidrains. byol-pytorch. `https://github.com/lucidrains/byol-pytorch`, 2021.

Brendan Murphy. cs230. `https://github.com/csbrendan/cs230`, 2023.

Frank Odom. Byol. `https://github.com/fkodom/byol`, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

Mohamed Saeed, Rand Muhtaseb, and Mohammad Yaqub. Is contrastive learning suitable for left ventricular segmentation in echocardiographic images? *arXiv preprint arXiv:2201.07219*, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

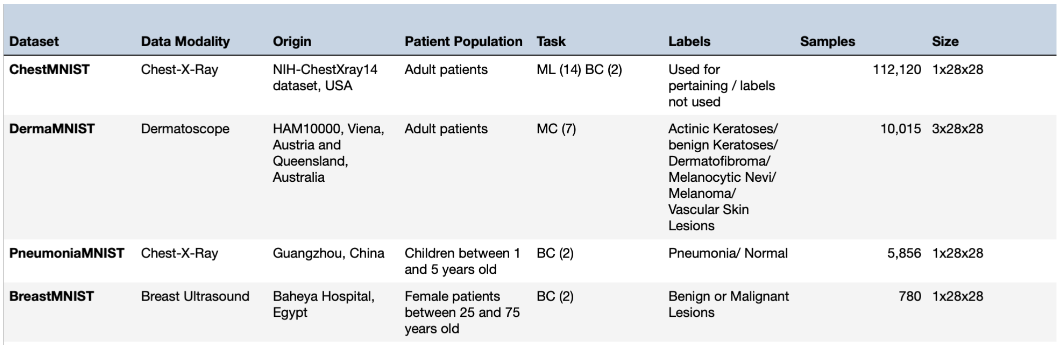
Rogier Van der Sluijs, Nandita Bhaskhar, Daniel Rubin, Curtis Langlotz, and Akshay Chaudhari. Exploring image augmentations for siamese representation learning with chest x-rays. *arXiv preprint arXiv:2301.12636*, 2023.

Cara Van Uden, Jeremy Irvin, Mars Huang, Nathan Dean, Jason Carr, Andrew Ng, and Curtis Langlotz. How to train your chexdragon: Training chest x-ray models for transfer to novel tasks and healthcare systems. *arXiv preprint arXiv:2305.08017*, 2023.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

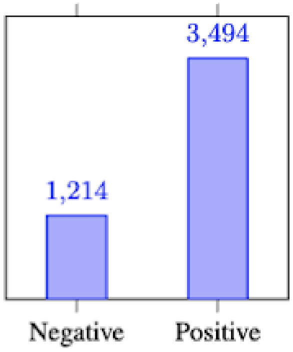
Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

A Appendix



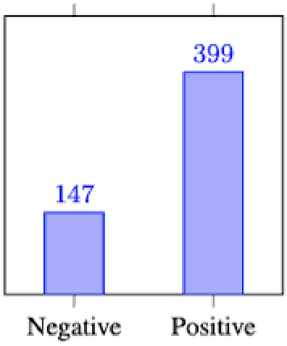
Dataset	Data Modality	Origin	Patient Population	Task	Labels	Samples	Size
ChestMNIST	Chest-X-Ray	NIH-ChestXray14 dataset, USA	Adult patients	ML (14) BC (2)	Used for pertaining / labels not used	112,120	1x28x28
DermaMNIST	Dermatoscope	HAM10000, Viena, Austria and Queensland, Australia	Adult patients	MC (7)	Actinic Keratoses/ benign Keratoses/ Dermatofibroma/ Melanocytic Nevi/ Melanoma/ Vascular Skin Lesions	10,015	3x28x28
PneumoniaMNIST	Chest-X-Ray	Guangzhou, China	Children between 1 and 5 years old	BC (2)	Pneumonia/ Normal	5,856	1x28x28
BreastMNIST	Breast Ultrasound	Baheya Hospital, Egypt	Female patients between 25 and 75 years old	BC (2)	Benign or Malignant Lesions	780	1x28x28

Figure 7: MedMNIST dataset details including Modality, Origin, Patient Population, Task and the Labels, etc.



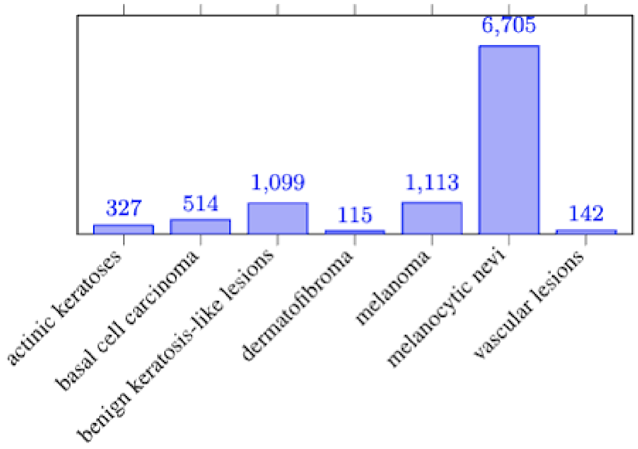
(a) PneumoniaMNIST

Figure 8



(b) BreastMNIST

Figure 9



(c) DermaMNIST

Figure 10

TN	FP
FN	TP

Figure 11: Confusion Matrix

167	67
13	387

Figure 12: BYOL+ Finetuned for Pneumonia

34	8
14	100

Figure 13: BYOL+ Finetuned for BreastMNIST