Leveraging Self-Supervised Learning for Enhanced Medical Image Analysis: A Comparative Study of Barlow Twins Pre-training and ImageNet Fine-Tuning

Mohamed Elsherif, Simon Frank, Daniela Kemp, Gwent Krause & Tim Rebig
Department of Computer Science, Eberhard Karls University of Tübingen.

Abstract

Medical image classification is a challenging and time-consuming task due to the availability of labeled large-scale medical imaging datasets. To address this challenge, we employ Barlow Twins pre-training on a ResNet-18 backbone as a self-supervised learning approach to generate transferable representations for medical image classification tasks. We attach a linear readout head to probe the feature vectors produced by the backbone and train the combined model in conjunction with two different fine-tuning strategies, namely, surgical and full fine-tuning. We evaluate the performance of our Barlow Twins model in comparison to a ResNet model pre-trained on ImageNet. Our experiment results demonstrate the superiority of our pre-trained features compared to the generic ImageNet based model on two different medical domains, showcasing the strength of utilizing Barlow Twins for enhancing performance in medical imaging classification tasks, particularly in the context of liver tumors and colorectal adenocarcinomas.

1 Introduction

In recent years, the field of computer-aided diagnosis and analysis in healthcare has seen significant progress due to advancements in deep learning and the availability of large-scale medical imaging data sets. However, obtaining labeled data for training traditional supervised learning methods can be challenging and time-consuming in the medical domain as it requires medical expertise. To address this, self-supervised learning (SSL), a subset of unsupervised learning, has emerged as a promising approach (Bagnell and Hebert, 1990). In SSL, two essential tasks/phases characterize the learning pipeline: the pretext task (unsupervised learning) and the downstream task (supervised learning) (Noroozi and Favaro, 2016). During the pretext task, a model is trained in a supervised manner using unlabeled data, enabling it to learn meaningful representations by utilizing the data itself as a supervisory signal without the need for explicit annotations (Holmberg et al., 2020). Commonly used models include pre-trained CNN architectures like ResNet (He et al., 2016), DenseNet (Huang et al., 2017), VGG (Simonyan and Zisserman, 2014), or EfficientNet (Tan and Le, 2019) ordinarily trained on the ImageNet dataset (Deng et al., 2009). The first phase involves employing a method from a diverse set of approaches, such as contrastive learning, predictive learning, generative learning, and temporal/viewpoint transformation, to create auxiliary tasks that facilitate effective representation learning (Liu et al., 2021). Following the pretext task, the pre-trained model, trained on a large amount of data during the pretext step, is adapted/fine-tuned for a downstream task (e.g image classification, segmentation, object detection or disease diagnosis) where the data is potentially scarce, a concept known as transfer learning (Ge and Yu, 2017; Chopra et al., 2018). During this transfer learning phase, a fine-tuning strategy is utilized to preserve the pre-trained knowledge (Goodfellow et al., 2016). Among those strategies specifically aligned with image-related downstream tasks, which have shown to maximize the utility of pre-trained features and yield significant benefits compared to commonly used end-to-end fine-tuning that is still largely the norm, are the full fine-tuning (all layers of the pre-trained model are updated using the target task's labeled data) and the surgical fine-tuning (only a subset of layers is fine-tuned, while the remaining layers are kept frozen and unchanged) strategies (Lee et al., 2022; Khan and Fang, 2023).

In the context of medical image classification, contrastive learning methods, exemplified by SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), and SwAV (Caron et al., 2020), have gained prominence in the pretext task by maximizing agreement between augmented versions of the same input (positive pairs) and minimizing agreement between augmented versions of different inputs (negative pairs) using a contrastive loss function. Here augmented refers to distortions applied to the samples e.g. image transformations. While these approaches offer flexibility in pretext task design and can handle data imbalance and noisy or corrupted data, they can be computationally expensive, especially for large-scale data sets, due to the need for negative pairs and large batch sizes (Chen et al., 2020b; Zhang et al., 2020).

In contrast, the Barlow Twins method (He et al., 2021) minimizes an objective function which reduces the discrepancy between the identity matrix and a cross-correlation matrix constructed from two batch embeddings. The embedding vectors result from distorting every sample in two random ways before passing them through a backbone (encoder), typically a neural network, followed by upscaling them via a MLP (Multilayer Perceptron) as projection head to a higher dimension. Barlow Twins achieves informative and non-redundant representations that capture fine-grained details and essential features in the data, making them suitable for downstream tasks such as image classification. Moreover, Barlow Twins uses positive pairs only, reducing the need for batch sizes of several thousand samples, making it more memory-efficient and easier to implement. Even though the method scales well to large data sets, generalizes to unseen data, and performs effectively with limited labeled data (He et al., 2021), and has shown promising results in other domains, its performance in downstream medical image classification tasks remains under-explored and merits further exploration (Krishnan et al., 2022).

In our experiment, we sought to investigate the application of Barlow Twins for medical imaging classification, focusing specifically on liver tumor and colorectal cancer domains. To evaluate the effectiveness of the Barlow Twins approach in context of classification of medical imaging, we compare it with traditional transfer learning using ImageNet pre-training as a baseline (Russakovsky et al., 2015). Additionally, we explore two fine-tuning strategies in the pretext phase in conjunction with a linear readout head. The replacement of a model's head up to a chosen layer with a linear classifier and freezing all lower layers for training, also called probing in the literature (Alain and Bengio, 2018), allows for investigating the learned features of specific layers. Compared to fine-tuning the whole model, probing has shown to be less prone to performance drops in the presence of distribution shifts (Kumar et al., 2022) which are of critical concern for practical medical diagnostic (Park et al., 2021).

By exploring the performance of Barlow Twins in conjunction with these different fine-tuning strategies, we aim to assess its effectiveness as a self-supervised learning approach in generating transferable representations for medical image classification tasks.

2 Related Work

During the last few years, self-supervised not only managed to catch up to the performance of supervised learning in image classification tasks, but to even eventually surpass previous methods in performance on large datasets with few labels (Zhang and Gu; He et al.; Chen et al.; Chen et al.; Hénaff et al.; Xu; Shurrab and Duwairi; Huang et al.; Chen et al.). While early approaches (up until 2020) in self-supervised learning tried to train their models on related tasks, their accuracy could seldom match the performance of supervised models (Zhang and Gu; Haghighi et al.; Xu; Chen et al.). Such tasks included prediction of falsified or removed patches, ordering of patches or scaled versions, as well as discrimination of modified image pairs against negative samples (Zhang and Gu; Hénaff et al.; Xu; Shurrab and Duwairi; Huang et al.; Chen et al.). These approaches aim to learn the general structure of the image (Zhang and Gu). Methods leveraging innate knowledge found in images to perform generative or prediction tasks on their targets, where often computationally expensive. To improve such models' performance, several tasks could be combined to get generalized features (Haghighi et al.; Tamé et al.), the number of layers or the batch size could be increased (Chen et al., 2020a; Hénaff et al.), the image size of the data set scaled up (Hénaff et al.; Ciga et al.) or the number of training samples increased (Haghighi et al.; Zhang et al.; Grill et al.; Hénaff et al.).All those methods also significantly raise requirements for computation power and time, making the early approaches only feasible for large amounts of data with few labels – such as medical images. Advances in the year 2020 led to constrastive methods overtake supervised learning. MoCo (Momentum Contrast) versions v1 & v2, are still widely used methods, utilizing a queue of features for the negative samples with a weighted momentum average to influence the distance optimization of the positive samples, reducing both processing time and memory needs (He et al., 2020; Chen et al., b). SimCLR (Simple framework for Contrastive Learning of visual Representation) introduces a nonlinear transformation for both representations of an image where the agreement is maximized. Large amounts of time and data were needed to further improve the model (Chen et al., 2020a). BYOL (Bootstrap your own latent) uses two adjusted image embeddings, where the target network predicts the image composition of the other network's embedding. The second network is slowly updated with the moving average of the target network, since BYOL does not use negative pairs, the amount of optimizations is drastically reduced, making the approach both faster and less prone to overfitting and thereby better (Grill et al., 2020).

Since 2020, many approaches are based on contrastive learning with positive pairs. This serves to reduce computing time and can be combined with other methods to further improve performance. With these advances, SSL was increasingly used in medical image classification (Zhang et al.; Srinidhi et al.; Ciga et al.). A further improvement was the usage of diversified image augmentations, to have more possible positive samples. While some tried to only include medically sound augmentations (Qin et al.) such as color changes, adjustments such as rotation and cropping did improve the performance for every model that uses them (Zbontar et al.; Qin et al.; Ciga et al.). More diverse approaches (Srinidhi et al.) use pseudolabels and constant fine-tuning of classification layers using updated pseudolabels, for the price of potentially overfitting if mistakes happen early in the classification.

Some methods try to optimize the loss function for redundancy reduction like Ermolov et al. and Barlow Twins (He et al., 2021). Since the latter was published in 2021, it has been used in diverse ML tasks, such as solving Atari games (Cagatan), but even though it is well-suited for large feature spaces and little data, there was scarce usage in the medical imaging field. In a arxiv preprint from 2021 by Zhang et al., they trained a Barlow Twin model on histopathological breast cancer images and found in their experiments, that while Barlow Twins' features removed an inherent data source bias, training sets containing 1000s of samples performed worse with this method than with fully fine-tuned pre-trained supervised ResNet-50 models. Methodical, our choice of training parameters differs significantly from their work leading to contrasting experimental results and conclusions.

3 Experiment

3.1 Datasets

We conducted our experiment using the MIMeta dataset, which is comprised of 17 publicly available image datasets from different medical domains (Group, 2023). We chose the MIMeta dataset because it can provide a consistent interface for loading different medical datasets, and all images are standardized to 224x224 pixel size, thus allowing for seamless cross-domain pre-training and consistent image handling across different domains during pre-training. In addition, MIMeta dataset provides a predefined data split for each dataset which we diligently adhered to throughout the whole training procedure.

From the MIMeta dataset, we chose two subset medical domains for our experiment:

3.1.1 Liver Tumor Domain

As the first subset of the MIMeta dataset, we chose the liver tumor domain. This subset encompasses a total of 4,935 contrast-enhanced CT scan images. These images represent axial, coronal, and sagittal slices, amounting to 1,645 images each. The central objective of this dataset revolves around the identification of 11 distinctive labels corresponding to 8 distinct organs. These organs include the heart, lungs, liver, spleen, pancreas, kidneys, bladder, and femoral head. Notably, for bilateral organs, namely the lungs, kidneys, and femoral head, the dataset accounts for laterality by specifying "Right" or "Left".

3.1.2 Colorectal Cancer Domain

Constituting the second subset within the MIMeta dataset, the domain of interest pertains to colorectal cancer (CRC). This segment encompasses a substantial collection of 107,180 images derived from hematoxylin and eosin stained tissue slices extracted from the colon and rectum. The primary objective underlying this dataset is the classification of nine distinct stromal and non-stromal constituents of the tumor microenvironment: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM).

3.2 Proposed Methods

3.2.1 Pre-training

We adopted the Barlow Twin model and adhered to the hyperparameters outlined in the original Barlow Twin paper (He et al., 2021), featuring a learning rate of 3×10^{-2}, the Adam optimizer, and a batch size of 256.

To lay the groundwork for our model's foundation, we initiated two pre-training runs: In the initial run, we employed a cross-domain strategy, leveraging the images from the liver tumor domain as detailed earlier. Given the MIMeta toolbox's constraint of loading one dataset at a time, we introduced a wrapper. This wrapper facilitated the concurrent loading of all three subsets of the liver tumor dataset, streamlining our operations; In the subsequent pre-training run, we applied Barlow Twins to the colorectal cancer domain. This choice would enable us to delve into our approach using a significantly larger volume of data.

Given the substantial reliance of medical images on textural features (Castellano et al., 2004) (Kiani, 2022), we employed the ResNet-18 architecture which contains 11.2 million trainable parameters. Our choice is underpinned by the inherent texture bias present, for instance, in CNNs trained on ImageNet data (Geirhos et al., 2022). For effective feature extraction, we incorporated a projection head of size 2048 composed of three fully connected linear layers. The initial two layers are followed by batch normalization layers and rectified linear units, collectively contributing an additional 9.4 million trainable parameters.

3.2.2 Fine-tuning

We established a comparative framework by employing fine-tuning to the ImageNet pre-trained and our Barlow Twins pre-trained ResNet-18 models. These models generate a 512-dimensional intermediate feature vector representation after their respective last convolutional layers. This feature vector serves as the basis for subsequent classification tasks. To address the classification challenge, we employed a linear readout head. This readout head encompasses a linear layer followed by a softmax activation function. The pivotal 512-dimensional feature vector obtained from the final convolutional layer serves as the input for this readout head. Our exploration encompassed two distinct strategies for fine-tuning. In both scenarios, the models underwent training tailored to the specific dataset's downstream task. This training harnessed the available training set, consisting of image-label pairs. The first approach involved preserving the weights of the backbone architecture while exclusively training the parameters of the readout head, amounting to 5.6k parameters. This configuration enables linear feature probing. Alternatively, the second fine-tuning approach embraced comprehensive training of the entire model, encompassing both the backbone and the readout head. Comparing both strategies will allow us, as we use a single linear layer as readout head, to determine the training influence of labeled ground truth data on the features obtained during pre-training. In a bid to discern the impact of supervised labels on performance, we manipulated the quantity of data utilized from the training set. This variance enabled us to systematically evaluate the influence of varying label proportions on the experimental outcomes.

3.3 Feature clustering analysis

Besides performance evaluation, we also visually inspected representations of the feature vectors produced by the last layer of our backbone network on which we applied T-SNE (t-Distributed Stochastic Neighbor Embedding) (van der Maaten and Hinton, 2008) for dimensionality reduction. The resulting two-dimensional clustering maps of the dataset classes provide intuition for the discriminatory power of the respective features underlying the classification task and indicate which classes will be easier or harder to separate in higher dimensions.

4 Results & Discussion

4.1 Liver Tumor Domain

4.1.1 Accuracy Analysis

The accuracy results for various fine-tuning strategies on the sagittal organ slice dataset, as outlined in Section 3.2.2, are presented in Figure 1. Specifically, when only the readout head is trained, our Barlow Twin model achieves accuracies of 63.41%, 71.34%, 77.44%, and 79.27% for fine-tuning proportions of 10%, 25%, 50%, and 100%, respectively. In contrast, the ImageNet model attains accuracies of 54.27%, 48.78%, 47.56%, and 50% using the same fine-tuning approach.

When fine-tuning involves updating all parameters, our model achieves accuracies of 68.9%, 75%, 78.66%, and 78.05%, while the ImageNet model achieves 62.2%, 71.95%, 79.88%, and 76.22% using the aforementioned fine-tuning proportions.

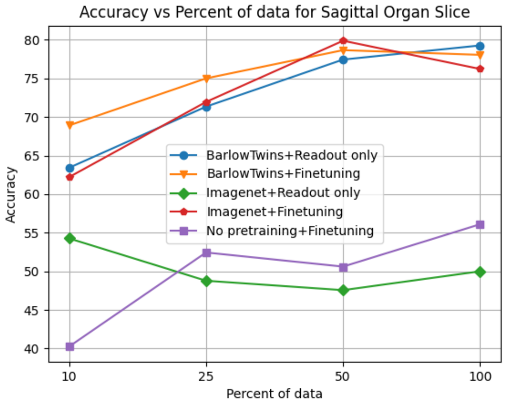

Figure 1: Accuracy on the sagittal organ slice dataset with different percentages of data used with different fine-tuning strategies

Similar trends are observed for the axial and coronal organ slice datasets. The detailed accuracy values can be found in Figures 9 and 10 in Appendix A.1. Notably, the highest accuracy is achieved on the axial slice dataset. This could be attributed to the axial view's provision of spatial information about the organs in relation to a virtual mid-line dividing the slice image into right and left halves. As a result, this axial perspective enhances the model's capacity to accurately differentiate and classify organ features.

Our Barlow Twin model consistently outperforms the ImageNet pre-trained model across all 3 datasets, regardless of the proportion of fine-tuning data used, when only the readout head is trained. Moreover, our pre-trained model displays an enhanced capability to improve accuracy with more fine-tuning data, while the ImageNet model occasionally experiences a decline in accuracy with increased fine-tuning data. The discrepancies in performance might be due to shifts in training data distribution and the test set's limited size. With a small test set, classifiers' decision boundaries can fluctuate based on hyperparameters and weight initialization, particularly for weak features. This highlights the resilience of our domain-specific Barlow Twin pre-trained features in contrast to the generic ImageNet based features.

4.1.2 Class-Specific Performance

Normalized confusion matrices for the coronal dataset, utilizing all data for fine-tuning the readout head, are presented in Figure 2 for our Barlow Twin model and Figure 3 for the ImageNet pre-trained model. Our model achieves 100% accuracy for heart, left lung, right lung, liver, left femoral head, and right femoral head. For pancreas and right kidney, accuracy exceeds 90%, while it surpasses 80% for spleen, left kidney, and bladder.

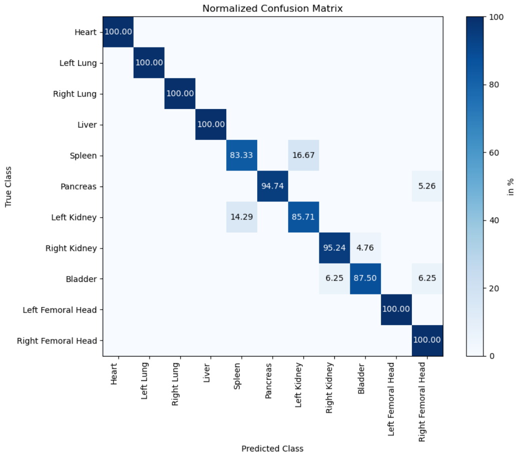

Figure 2: Normalized confusion matrix with Barlow twin pre-training and readout head fine-tuning of coronal organ slices

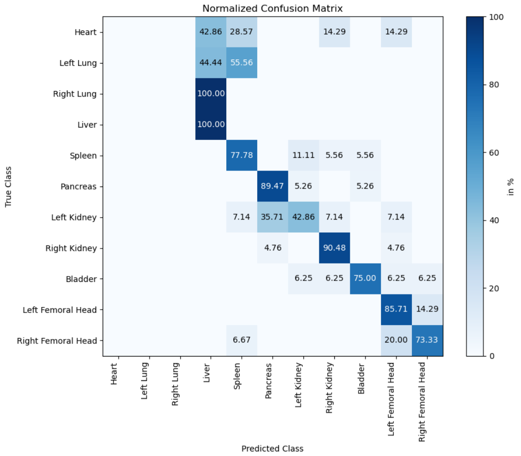

Figure 3: Normalized confusion matrix with ImageNet pre-training and readout head fine-tuning of coronal organ slices

Conversely, the ImageNet pre-trained model records zero accuracy for heart, right or left lung. Accuracy for liver, pancreas, right kidney, and left femoral head surpasses 80%, while it exceeds 70% for spleen, bladder, and right kidney. Accuracy for left kidney is only 42.86%.

While our model identifies all classes correctly, the ImageNet model only identifies a subset of classes. Moreover, our Barlow Twin pre-trained model significantly outperforms the ImageNet model across all organ classes, as evidenced by class-specific accuracy in the confusion matrix.

4.2 Colorectal Cancer Domain

4.2.1 Accuracy Analysis

The outcomes concerning the colorectal cancer dataset are visualized in Figure 4. The Barlow Twins pre-trained model demonstrates performance metrics of 91.25%, 93.26%, 95.78%, 96.29%, 96.43%, and 96.64% accuracy for 0.5%, 1%, 10%, 25%, 50%, and 100% training data, respectively, when only a readout head is trained. With an additional update of the CNN parameters, the model achieves accuracies of 88.93%, 92.03%, 96.34%, 96.42%, 97.34%, and 97.97% on the same training subsets. For the ImageNet pre-trained model, accuracy scores of 85.91%, 88.66%, 92.40%, 93.11%, 93.49%, and 93.61% are achieved, with only a trained readout head. Whole-model fine-tuning leads to accuracies of 89.30%, 93.00%, 95.98%, 94.57%, 97.74%, and 96.86% on the aforementioned training subset sizes. In general, our approach yields the best results across different training subsets. The only exception is observed for 50% of the training data, where the fully fine-tuned ImageNet pre-trained model is slightly better. Importantly, the Barlow Twins pre-trained model significantly outperforms the ImageNet pre-trained model when only a readout head is trained, which can be attributed to superior feature quality following pre-training. The performance differences decrease when training the entire models. Notably, fine-tuning the entire models only slightly surpasses our pre-training method's performance, when training only the readout head, for training subset sizes of 10% and larger. Fine-tuning the entire model even reduces test set accuracy for smaller subset sizes, further emphasizing our feature space's quality. Increasing training samples leads to overall higher test set accuracy. Notable inconsistencies arise in the training runs of the ImageNet pre-trained model when fine-tuning the entire model, as training data increases from 10% to 25%, and subsequently from 50% to 100%. This phenomenon stems from the vast difference in trainable parameters between a small readout head and the entire model. As a result, the model's sensitivity to hyperparameter choices increases. Although efforts are made to tune hyperparameters, the resulting models do not reach their anticipated peak performance levels.

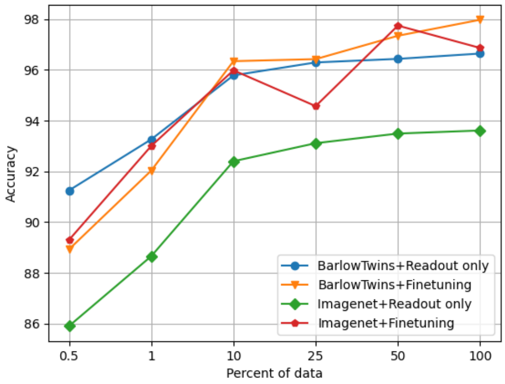

Figure 4: Accuracy on the colorectal cancer dataset with different percentages of data used with different fine-tuning strategies

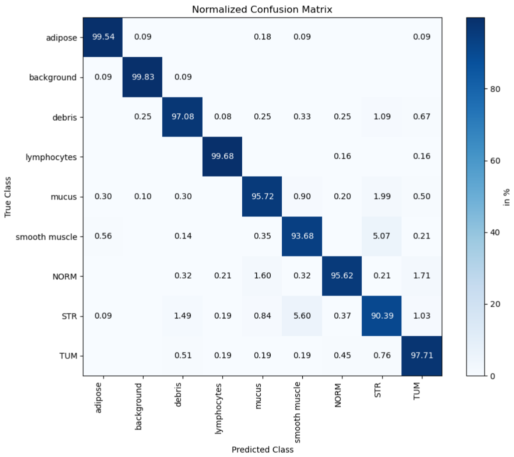

Figure 5: Normalized confusion matrix with Barlow Twins pre-training and readout head fine-tuning on 100 percent of colorectal cancer data

4.2.2 Class-Specific Performance

Further insight into the quality of the Barlow Twins' pre-trained feature space is gained by examining differences in confusion matrices, particularly when exclusively training a readout head. The confusion matrix for 100% of the data is presented in Figure 5. Classes such as adipose, background, and lymphocytes achieve near-perfect predictions with accuracies of 99.54%, 99.83%, and 99.68%, respectively. Debris and TUM classes attain accuracies of 97.08% and 97.71%, while mucus and NORM classes achieve 95.72% and 95.62%, respectively.

Accuracy values slightly diminish for the STR and smooth muscle classes, reaching 90.39% and 93.68%, respectively. The model shows a tendency to predict the STR class when the true class is smooth muscle, occurring in 5.07% of cases. Conversely, when the true class is STR, the model leans towards predicting smooth muscle in 5.60% of cases. The high recall accuracies for adipose, background, and lymphocytes indicate effective linear separation in the feature space for these classes. Although debris, mucus, NORM, and TUM classes exhibit slightly lower accuracy, they are not significantly afflicted by major confusions. Notably, distinguishing between STR and smooth muscle poses a significant challenge.

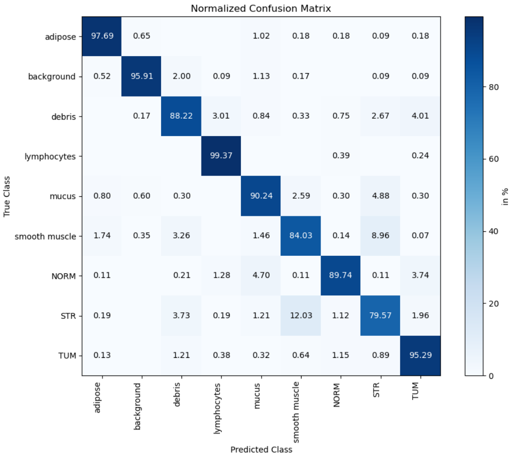

Figure 6: Normalized confusion matrix with Barlow Twins pre-training and readout head fine-tuning on 0.5 percent of colorectal cancer data

The confusion matrix for training on a 0.5% subset of the data is illustrated in Figure 6. Distinct class predictions are accurate: adipose (97.69%), background (95.91%), debris (88.22%), lymphocytes (99.37%), mucus (90.24%), smooth muscle (84.03%), NORM (89.74%), STR (79.57%), and TUM (95.29%). It is noteworthy that the model tends to predict smooth muscle when the ground truth is STR (8.96% of cases), and vice versa (12.03% of cases). Adipose, background, lymphocytes, and TUM classes only experience marginal accuracy decreases compared to training on the entire dataset, indicating robust linear separability. The choice of samples has a more pronounced impact on other classes, particularly in distinguishing between smooth muscle and STR, which heavily relies on a comprehensive sample representation along boundary regions. Importantly, the confusion between STR and smooth muscle is not exclusive to the Barlow Twins-derived feature space; it persists across all settings, as detailed in Appendix A.1.1. This phenomenon can be attributed to tissue morphology similarities reflected in analogous textures. Achieving linear differentiation between both tissue types evidently presents a more intricate challenge.

Contrasting our results with the usage of Barlow Twins on histopathological breast cancer data by Zhang et al. we explain the different outcomes given equal batch sizes with the short pre-training and fine-tuning time in their experimental set-up. Our 1000 epochs adopted from He et al. are crucial for attaining our feature capabilities compared to their 100 pre-training epochs, especially because their larger projection head of 4096 and ResNet-50 should potentially have allowed for better performance of the Barlow Twins method.

4.3 Visual Feature Analysis

The application of T-SNE dimensionality reduction to the feature vectors derived from the colorectal cancer dataset reveals a noticeable differentiation between the backbone trained using Barlow Twins and the ImageNet pre-trained backbone from the pretext task. This distinction is evident in Figures 7 and 8. Notably, the Barlow Twins' features exhibit reduced dispersion and greater concentration around cluster centers for a majority of classes. Importantly, an enhanced separation between clusters is observed across almost all cases. However, there are exceptions such as the background and adipose tissue classes, which are effectively distinguished by both backbones. Conversely, the smooth muscle and STR classes display overlap and reduced concentricity with both approaches, aligning with our earlier performance findings.

It is worth mentioning that the TUM class (colorectal adenocarcinoma epithelium), which holds crucial significance in medical diagnosis, appears less scattered within other classes.

The application of additional supervised fine-tuning bridges the advantage of Barlow Twins' features over those produced by ImageNet, provided labeled data is available. However, we acknowledge that the smaller liver tumor organ slices dataset prevents conclusive visual inferences. The T-SNE representations of the test set encompass only approximately 165 samples across 11 classes, resulting in a less coherent and interpretable depiction. For further details, refer to Appendix A.2.

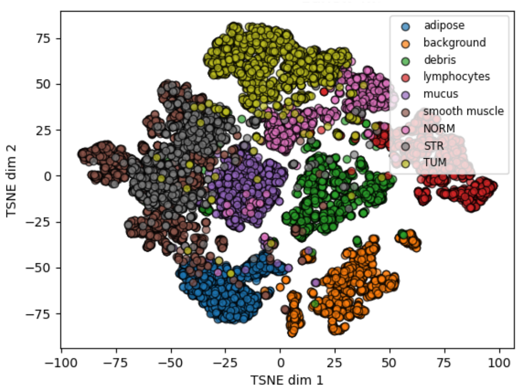

Figure 7: T-SNE projection of colorectal cancer feature vectors by ResNet-18 pre-trained with Barlow Twins.

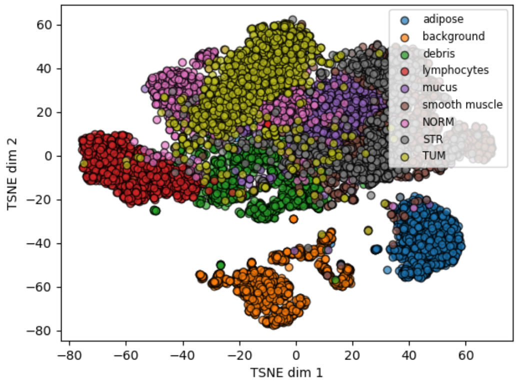

Figure 8: T-SNE projection of colorectal cancer feature vectors by ResNet-18 pre-trained on ImageNet.

5 Limitations

Our study acknowledges the presence of several limitations. Firstly, we encountered hardware constraints that have limited us to a maximum batch size of 256 during the pre-training phase. Those constraints were in the form of limited memory capacity of Nvidia RTX 4090's 24GB of VRAM. Despite these constraints, our chosen batch size was still expected to yield meaningful results based on existing evidence from various studies showing that batch sizes in the range of 256 can yield competitive results in image classification tasks (Keskar et al., 2016; He et al., 2018). However, larger batch sizes may optimize the convergence and thus the performance of the Barlow Twins method even further (Smith, 2017).

Our usage of ResNet pretrained on ImageNet may be another limitation. ImageNet classification is between a multitude of classes (Shurrab and Duwairi), and often focused on the center of the image (Haghighi et al.; Ganin and Lempitsky, 2015) which influence the architecture of the network, leading to a domain gap between general object classes and medical images. The latter often exhibit unique characteristics that deviate from everyday objects, potentially affecting the transferability of features (Ganin and Lempitsky, 2015).

Lastly, our study solely relied on image-based features and decisions for medical image classification. While image-based approaches are valuable, they inherently lack the contextual insights that clinical data can provide (Li et al., 2019). The absence of clinical data, including patient history, demographic information, and other medical context, in our study, could curtail the model's capacity to make decisions that align with real-world clinical practice, wherein such contextual information is pivotal for accurate diagnosis and decision-making. Hence, addressing these limitations would lead to a more precise interpretation of the results and their implications.

6 Conclusion & Future Work

Our experiment results demonstrate the effectiveness of our domain-specific Barlow Twin pre-trained features compared to the generic ImageNet based features. Our Barlow Twin model consistently outperforms the ImageNet model on various datasets and fine-tuning strategies, showcasing the strength of the learned features. Further supported by visual analysis and the probing approach, these well-separated features exhibit robustness and improved accuracy. The success of our approach indicates the potential of utilizing this self-supervised pre-training methods for enhancing performance in medical image classification tasks, especially in the presence of label scarcity. The demonstrated performance gap can only be bridged by large amounts of labeled data.

As future work, provided access to hardware resources with sufficient VRAM capacity, experiments with larger batch sizes and higher dimensional projection heads can be conducted. Moreover, extensive hyperparameter search and tailoring Barlow Twins' sample augmentation methods specifically to the textures present in the medical image domain at hand can enhance the performance.

Additionally, the surgical fine-tuning approach could be investigated in more detail by analyzing each convolution layer's impact on predictive performance to leverage the texture bias even further.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

J. Andrew Bagnell and Martial Hebert. 1990. A self-supervised learning approach for object detection on a mobile robot. In *Proceedings of the 1990 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Omer Veysel Cagatan. BarlowRL: Barlow twins for data-efficient reinforcement learning.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *International Conference on Machine Learning (ICML)*.

G. Castellano, L. Bonilha, L.M. Li, and F. Cendes. 2004. Texture analysis of medical images. *Clinical Radiology*, 59(12):1061–1069.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. a. Self-supervised learning for medical image analysis using image context restoration. 58:101539.

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. b. Improved baselines with momentum contrastive learning.

Sumit Chopra, Srinath Balakrishnan, and Raghuraman Gopalan. 2018. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML*.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. 7:100198.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. 2020. Whitening for self-supervised representation learning. *CoRR*, abs/2007.06346.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*.

Weidi Ge and Yizhou Yu. 2017. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1086–1095.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2022. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning.

Medical Image Analysis Group. 2023. Mimeta dataset. `https://www.l2l-challenge.org/data.html`. Accessed on August 15, 2023.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. DiRA: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20792–20802. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Ross Girshick, Maryna Zelenyuk, and Piotr Dollár. 2021. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, Mu Li, Xiaodong Liu, and Jian Sun. 2018. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

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. 6(1):74.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Muhammad Osama Khan and Yi Fang. 2023. Revisiting fine-tuning strategies for self-supervised medical imaging analysis.

Faeze Kiani. 2022. Texture features in medical image analysis: a survey.

Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. 2022. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.

Yujin Lee, Alice S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2022. Surgical fine-tuning improves adaptation to distribution shifts. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Haibo Li, Guorong Wu, and Qian Wang. 2019. Multimodal medical image fusion: A survey. In *Fusion in Computer Vision*, pages 53–78. Springer.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876.

Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak N. Patel. 2021. Reliable and trustworthy machine learning for health using dataset shift detection. *CoRR*, abs/2110.14019.

Wenkang Qin, Shan Jiang, and Lin Luo. Pathological image contrastive self-supervised learning. In Xinxing Xu, Xiaomeng Li, Dwarikanath Mahapatra, Li Cheng, Caroline Petitjean, and Huazhu Fu, editors, *Resource-Efficient Medical Image Analysis*, volume 13543, pages 85–94. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet: A large-scale hierarchical image database. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. 8:e1045.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Chetan L. Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. 75:102256.

Iván de Andrés Tamé, Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. Self-supervised curricular deep learning for chest x-ray image classification.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Jiashu Xu. 2021. A review of self-supervised learning methods in the field of medical image analysis. *International Journal of Image, Graphics and Signal Processing*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction.

Chuyan Zhang and Yun Gu. Dive into self-supervised learning for medical image analysis: Data, models and tasks.

Lantian Zhang, Mohamed Amgad, and Lee A D Cooper. A histopathology study comparing contrastive semi-supervised and fully supervised learning.

Zhirong Zhang, Yoshua Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2020. A comprehensive overview of self-supervised learning. *arXiv preprint arXiv:2006.10029*.

A Appendix

A.1 Performance results

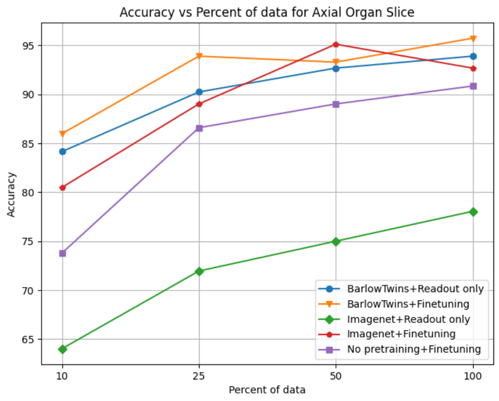

Figure 9: Accuracy on the axial organ slice dataset with different percentages of data used with different fine-tuning strategies

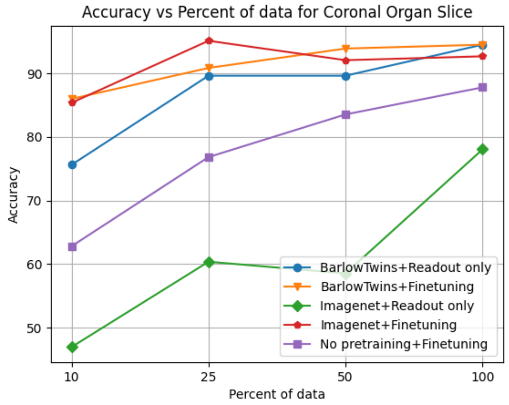

Figure 10: Accuracy on the coronal organ slice dataset with different percentages of data used with different fine-tuning strategies

A.1.1 Confusion Matrices for the CRC Domain

Additional confusion matrices for the missing models for fine-tuning on 0.5 percent of the training data.

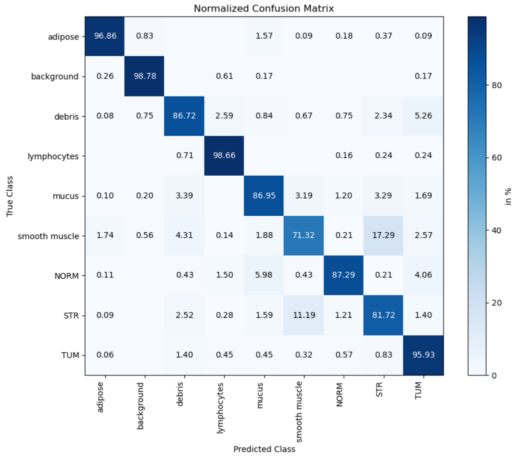

Figure 11: Normalized confusion matrix with ImageNet pre-training and fine-tuning the whole model of 0.5 percent of colorectal cancer data

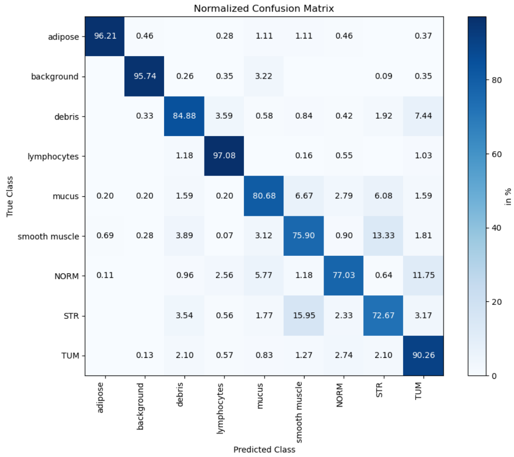

Figure 12: Normalized confusion matrix with ImageNet pre-training and only readout head fine-tuning of 0.5 percent of colorectal cancer data

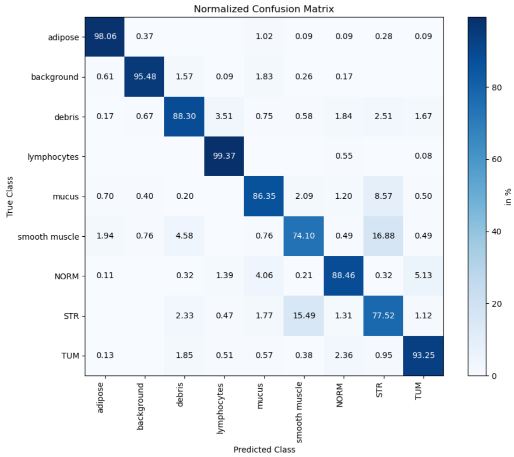

Figure 13: Normalized confusion matrix with Barlow Twins pre-training and fine-tuning the whole model of 0.5 percent of colorectal cancer data

A.2 Visual feature analysis

T-SNE feature maps with limited interpretability due to small test set size belonging to the liver tumor organ slices dataset.

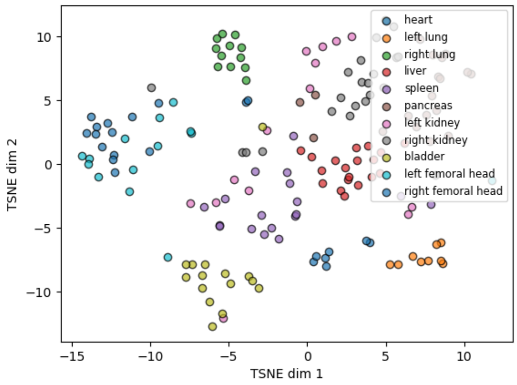

Figure 14: TSNE projection of axial organ slice feature vectors obtained from Barlow Twins pre-trained ResNet-18 with fine-tuned readout head.

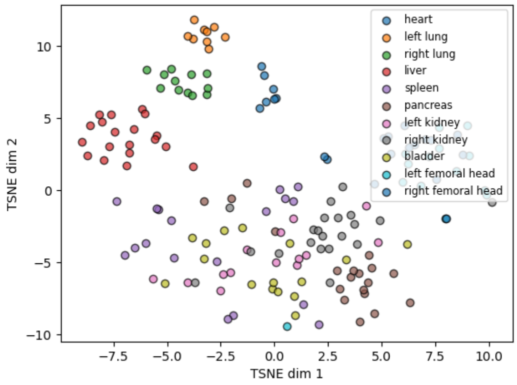

Figure 15: TSNE projection of axial organ slice feature vectors obtained from ImageNet based pre-trained ResNet-18 with fine-tuned readout head.