

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 agree-

ment 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; Shurab 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; Haghghi 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; Shurab 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 (Haghghi 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 (Haghghi 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 contrastive methods overtake supervised learning. MoCo (Momentum Contrast) versions v1 & v2, are still widely used methods, uti-

lizing 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

283 muscle, normal colon mucosa (NORM), cancer-
284 associated stroma (STR), and colorectal adenocar-
285 cinoma epithelium (TUM).

286 **3.2 Proposed Methods**

287 **3.2.1 Pre-training**

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

293 To lay the groundwork for our model’s founda-
294 tion, we initiated two pre-training runs: In the
295 initial run, we employed a cross-domain strategy,
296 leveraging the images from the liver tumor domain
297 as detailed earlier. Given the MIMeta toolbox’s
298 constraint of loading one dataset at a time, we in-
299 troduced a wrapper. This wrapper facilitated the
300 concurrent loading of all three subsets of the liver
301 tumor dataset, streamlining our operations; In the
302 subsequent pre-training run, we applied Barlow
303 Twins to the colorectal cancer domain. This choice
304 would enable us to delve into our approach using a
305 significantly larger volume of data.

306 Given the substantial reliance of medical images
307 on textural features (Castellano et al., 2004) (Kiani,
308 2022), we employed the ResNet-18 architecture
309 which contains 11.2 million trainable parameters.
310 Our choice is underpinned by the inherent texture
311 bias present, for instance, in CNNs trained on Im-
312 ageNet data (Geirhos et al., 2022). For effective fea-
313 ture extraction, we incorporated a projection head
314 of size 2048 composed of three fully connected
315 linear layers. The initial two layers are followed
316 by batch normalization layers and rectified linear
317 units, collectively contributing an additional 9.4
318 million trainable parameters.

319 **3.2.2 Fine-tuning**

320 We established a comparative framework by em-
321 ploying fine-tuning to the ImageNet pre-trained and
322 our Barlow Twins pre-trained ResNet-18 models.
323 These models generate a 512-dimensional inter-
324 mediate feature vector representation after their
325 respective last convolutional layers. This feature
326 vector serves as the basis for subsequent classifica-
327 tion tasks. To address the classification challenge,
328 we employed a linear readout head. This read-
329 out head encompasses a linear layer followed by
330 a softmax activation function. The pivotal 512-
331 dimensional feature vector obtained from the fi-
332 nal convolutional layer serves as the input for this

333 readout head. Our exploration encompassed two
334 distinct strategies for fine-tuning. In both scenar-
335 ios, the models underwent training tailored to the
336 specific dataset’s downstream task. This training
337 harnessed the available training set, consisting of
338 image-label pairs. The first approach involved pre-
339 serving the weights of the backbone architecture
340 while exclusively training the parameters of the
341 readout head, amounting to 5.6k parameters. This
342 configuration enables linear feature probing. Alter-
343 natively, the second fine-tuning approach embraced
344 comprehensive training of the entire model, encom-
345 passing both the backbone and the readout head.
346 Comparing both strategies will allow us, as we use
347 a single linear layer as readout head, to determine
348 the training influence of labeled ground truth data
349 on the features obtained during pre-training. In a
350 bid to discern the impact of supervised labels on
351 performance, we manipulated the quantity of data
352 utilized from the training set. This variance en-
353 abled us to systematically evaluate the influence
354 of varying label proportions on the experimental
355 outcomes.

356 **3.3 Feature clustering analysis**

357 Besides performance evaluation, we also visually
358 inspected representations of the feature vectors pro-
359 duced by the last layer of our backbone network on
360 which we applied T-SNE (t-Distributed Stochas-
361 tic Neighbor Embedding) (van der Maaten and
362 Hinton, 2008) for dimensionality reduction. The
363 resulting two-dimensional clustering maps of the
364 dataset classes provide intuition for the discrimi-
365 natory power of the respective features underlying
366 the classification task and indicate which classes
367 will be easier or harder to separate in higher dimen-
368 sions.

369 **4 Results & Discussion**

370 **4.1 Liver Tumor Domain**

371 **4.1.1 Accuracy Analysis**

372 The accuracy results for various fine-tuning strate-
373 gies on the sagittal organ slice dataset, as out-
374 lined in Section 3.2.2, are presented in Figure 1.
375 Specifically, when only the readout head is trained,
376 our Barlow Twin model achieves accuracies of
377 63.41%, 71.34%, 77.44%, and 79.27% for fine-
378 tuning proportions of 10%, 25%, 50%, and 100%,
379 respectively. In contrast, the ImageNet model
380 attains accuracies of 54.27%, 48.78%, 47.56%,
381 and 50% using the same fine-tuning approach.

382 When fine-tuning involves updating all parameters, our model achieves accuracies of 68.9%, 75%,
 383 78.66%, and 78.05%, while the ImageNet model
 384 achieves 62.2%, 71.95%, 79.88%, and 76.22% us-
 385 ing the aforementioned fine-tuning proportions.
 386

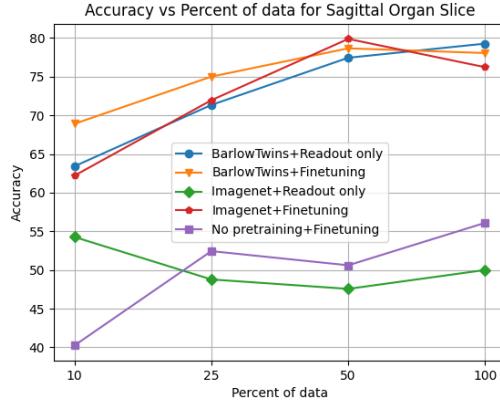


Figure 1: Accuracy on the sagittal organ slice dataset with different percentages of data used with different fine-tuning strategies

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

398 Our Barlow Twin model consistently outper-
 399 forms the ImageNet pre-trained model across all 3
 400 datasets, regardless of the proportion of fine-tuning
 401 data used, when only the readout head is trained.
 402 Moreover, our pre-trained model displays an
 403 enhanced capability to improve accuracy with more
 404 fine-tuning data, while the ImageNet model occa-
 405 sionally experiences a decline in accuracy with
 406 increased fine-tuning data. The discrepancies in
 407 performance might be due to shifts in training data
 408 distribution and the test set’s limited size. With
 409 a small test set, classifiers’ decision boundaries
 410 can fluctuate based on hyperparameters and weight
 411 initialization, particularly for weak features. This
 412 highlights the resilience of our domain-specific Bar-
 413 low Twin pre-trained features in contrast to the
 414 generic ImageNet based features.

4.1.2 Class-Specific Performance

415 Normalized confusion matrices for the coronal
 416 dataset, utilizing all data for fine-tuning the readout
 417 head, are presented in Figure 2 for our Barlow Twin
 418 model and Figure 3 for the ImageNet pre-trained
 419 model. Our model achieves 100% accuracy for
 420 heart, left lung, right lung, liver, left femoral head,
 421 and right femoral head. For pancreas and right kid-
 422 ney, accuracy exceeds 90%, while it surpasses 80%
 423 for spleen, left kidney, and bladder.
 424

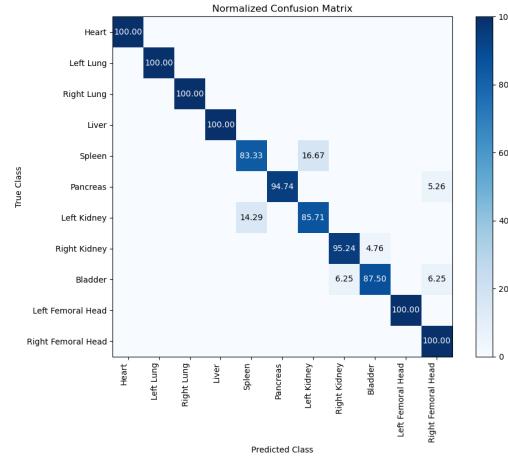


Figure 2: Normalized confusion matrix with Barlow twin pre-training and readout head fine-tuning of corona-
 l organ slices

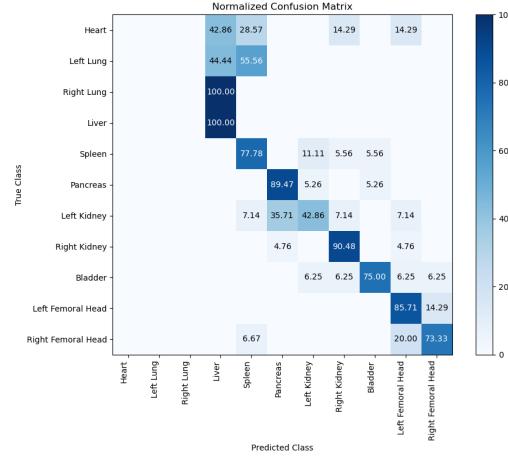


Figure 3: Normalized confusion matrix with ImageNet pre-training and readout head fine-tuning of corona-
 l organ slices

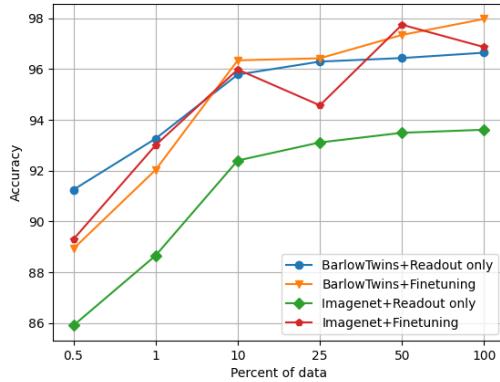
425 Conversely, the ImageNet pre-trained model
 426 records zero accuracy for heart, right or left lung.
 427 Accuracy for liver, pancreas, right kidney, and left
 428 femoral head surpasses 80%, while it exceeds 70%
 429 for spleen, bladder, and right kidney. Accuracy for
 430 left kidney is only 42.86%.

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

437 4.2 Colorectal Cancer Domain

438 4.2.1 Accuracy Analysis

439 The outcomes concerning the colorectal cancer
 440 dataset are visualized in Figure 4. The Barlow
 441 Twins pre-trained model demonstrates performance
 442 metrics of 91.25%, 93.26%, 95.78%, 96.29%,
 443 96.43%, and 96.64% accuracy for 0.5%, 1%, 10%,
 444 25%, 50%, and 100% training data, respectively,
 445 when only a readout head is trained. With an ad-
 446 dditional update of the CNN parameters, the model
 447 achieves accuracies of 88.93%, 92.03%, 96.34%,
 448 96.42%, 97.34%, and 97.97% on the same training
 449 subsets. For the ImageNet pre-trained model, accu-
 450 racy scores of 85.91%, 88.66%, 92.40%, 93.11%,
 451 93.49%, and 93.61% are achieved, with only a
 452 trained readout head. Whole-model fine-tuning
 453 leads to accuracies of 89.30%, 93.00%, 95.98%,
 454 94.57%, 97.74%, and 96.86% on the aforemen-
 tioned training subset sizes. In general, our ap-

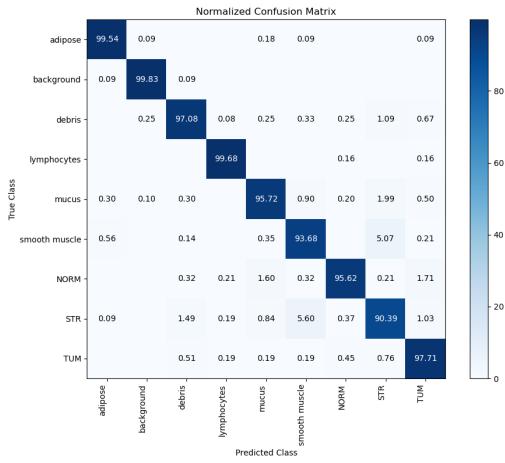


455 Figure 4: Accuracy on the colorectal cancer dataset with
 456 different percentages of data used with different fin-
 457 tunning strategies

458 proach yields the best results across different train-
 459 ing subsets. The only exception is observed for
 460 50% of the training data, where the fully fine-tuned
 461 ImageNet pre-trained model is slightly better. Im-
 462 portantly, the Barlow Twins pre-trained model sig-
 463 nificantly outperforms the ImageNet pre-trained
 464 model when only a readout head is trained, which
 465 can be attributed to superior feature quality fol-
 466 lowing pre-training. The performance differences
 467 decrease when training the entire models. Not-
 468 ably, fine-tuning the entire models only slightly
 469 surpasses our pre-training method’s performance,
 470 when training only the readout head, for training
 471 subset sizes of 10% and larger. Fine-tuning the
 472 entire model even reduces test set accuracy for
 473 smaller subset sizes, further emphasizing our fea-
 474 ture space’s quality. Increasing training samples
 475 leads to overall higher test set accuracy. Notable
 476 inconsistencies arise in the training runs of the Im-
 477 ageNet pre-trained model when fine-tuning the en-
 478 tire model, as training data increases from 10% to
 479 25%, and subsequently from 50% to 100%. This
 480 phenomenon stems from the vast difference in train-
 481 able parameters between a small readout head and
 482 the entire model. As a result, the model’s sensitiv-
 483 ity to hyperparameter choices increases. Although
 484 efforts are made to tune hyperparameters, the re-
 485 sulting models do not reach their anticipated peak
 486 performance levels.

487 4.2.2 Class-Specific Performance

488 Further insight into the quality of the Barlow Twins’
 489 pre-trained feature space is gained by examin-
 490 ing differences in confusion matrices, particularly
 491 when exclusively training a readout head. The con-
 492 fusion matrix for 100% of the data is presented in
 493 Figure 5. Classes such as adipose, background, and
 494 lymphocytes achieve near-perfect predictions with
 495 accuracies of 99.54%, 99.83%, and 99.68%, respec-
 496 tively. Debris and TUM classes attain accuracies of
 497 97.08% and 97.71%, while mucus and NORM
 498 classes achieve 95.72% and 95.62%, respectively.



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

497 Accuracy values slightly diminish for the STR
 498 and smooth muscle classes, reaching 90.39% and
 499 93.68%, respectively. The model shows a tendency
 500 to predict the STR class when the true class is
 501 smooth muscle, occurring in 5.07% of cases. Con-
 502 versely, when the true class is STR, the model
 503 leans towards predicting smooth muscle in 5.60%
 504 of cases. The high recall accuracies for adipose,
 505 background, and lymphocytes indicate effective lin-
 506 ear separation in the feature space for these classes.
 507 Although debris, mucus, NORM, and TUM classes
 508 exhibit slightly lower accuracy, they are not sig-
 509 nificantly afflicted by major confusions. Notably,
 510 distinguishing between STR and smooth muscle
 511 poses a significant challenge.

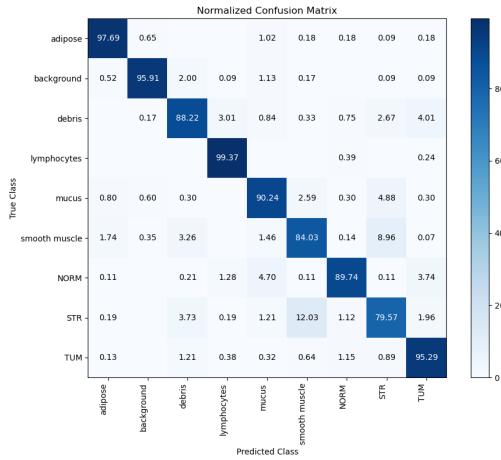


Figure 6: Normalized confusion matrix with Barlow Twins pre-training and readout head fine-tuning on 0.5 percent of colorectal cancer data

512 The confusion matrix for training on a 0.5% sub-
 513 set of the data is illustrated in Figure 6. Distinct
 514 class predictions are accurate: adipose (97.69%),
 515 background (95.91%), debris (88.22%), lympho-
 516 cytes (99.37%), mucus (90.24%), smooth muscle
 517 (84.03%), NORM (89.74%), STR (79.57%), and
 518 TUM (95.29%). It is noteworthy that the model
 519 tends to predict smooth muscle when the ground
 520 truth is STR (8.96% of cases), and vice versa
 521 (12.03% of cases). Adipose, background, lympho-
 522 cytes, and TUM classes only experience marginal
 523 accuracy decreases compared to training on the
 524 entire dataset, indicating robust linear separability.
 525 The choice of samples has a more pronounced im-
 526 pact on other classes, particularly in distinguishing
 527 between smooth muscle and STR, which heavily
 528 relies on a comprehensive sample representation
 529 along boundary regions. Importantly, the confusion
 530 between STR and smooth muscle is not exclusive

531 to the Barlow Twins-derived feature space; it per-
 532 sists across all settings, as detailed in Appendix
 533 [A.1.1](#). This phenomenon can be attributed to tis-
 534 sue morphology similarities reflected in analogous
 535 textures. Achieving linear differentiation between
 536 both tissue types evidently presents a more intricate
 537 challenge.

538 Contrasting our results with the usage of Barlow
 539 Twins on histopathological breast cancer data by
 540 [Zhang et al.](#) we explain the different outcomes
 541 given equal batch sizes with the short pre-training
 542 and fine-tuning time in their experimental set-up.
 543 Our 1000 epochs adopted from [He et al.](#) are crucial
 544 for attaining our feature capabilities compared to
 545 their 100 pre-training epochs, especially because
 546 their larger projection head of 4096 and ResNet-50
 547 should potentially have allowed for better perfor-
 548 mance of the Barlow Twins method.

4.3 Visual Feature Analysis

549 The application of T-SNE dimensionality reduction
 550 to the feature vectors derived from the colorectal
 551 cancer dataset reveals a noticeable differentiation
 552 between the backbone trained using Barlow Twins
 553 and the ImageNet pre-trained backbone from the
 554 pre-text task. This distinction is evident in Figures
 555 [7](#) and [8](#). Notably, the Barlow Twins’ features ex-
 556 hibit reduced dispersion and greater concentricity
 557 around cluster centers for a majority of classes. Im-
 558 portantly, an enhanced separation between clusters
 559 is observed across almost all cases. However, there
 560 are exceptions such as the background and adipose
 561 tissue classes, which are effectively distinguished
 562 by both backbones. Conversely, the smooth mus-
 563 cle and STR classes display overlap and reduced
 564 concentricity with both approaches, aligning with
 565 our earlier performance findings.

566 It is worth mentioning that the TUM class (col-
 567 orectal adenocarcinoma epithelium), which holds
 568 crucial significance in medical diagnosis, appears
 569 less scattered within other classes.

570 The application of additional supervised fine-
 571 tuning bridges the advantage of Barlow Twins’
 572 features over those produced by ImageNet, pro-
 573 vided labeled data is available. However, we ac-
 574 knowledge that the smaller liver tumor organ slices
 575 dataset prevents conclusive visual inferences. The
 576 T-SNE representations of the test set encompass
 577 only approximately 165 samples across 11 classes,
 578 resulting in a less coherent and interpretable depic-
 579 tion. For further details, refer to Appendix [A.2](#).

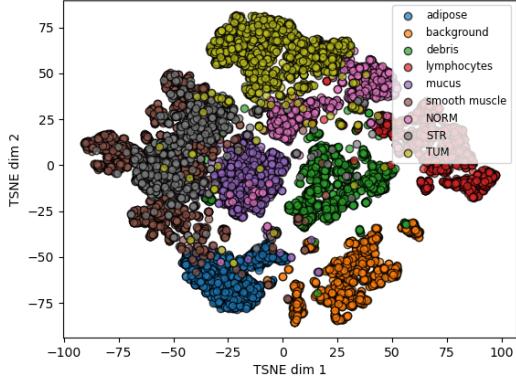


Figure 7: T-SNE projection of colorectal cancer feature vectors by ResNet-18 pre-trained with Barlow Twins.

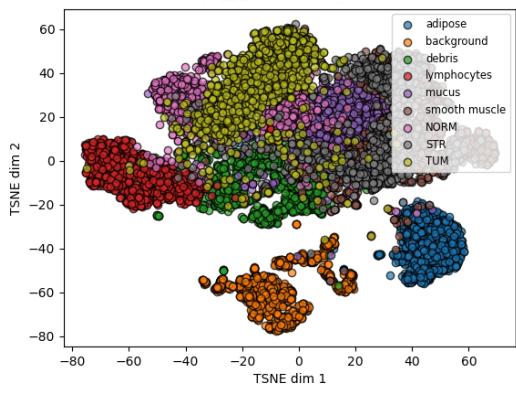


Figure 8: T-SNE projection of colorectal cancer feature vectors by ResNet-18 pre-trained on ImageNet.

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 (Shurab and Duwairi), and often focused on the center of the image (Haghghi 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, Zhaoan 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.121-challenge.org/data.html>. Accessed on August 15, 2023.
- Fatemeh Haghghi, 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,

758	Karsten U Kortuem, et al. 2020. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. <i>Nature Machine Intelligence</i> , 2(11):719–726.	813
759		814
760		815
761		816
762		
763	Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In <i>Proceedings of the IEEE conference on computer vision and pattern recognition</i> , pages 4700–4708.	817
764		818
765		819
766		820
767		821
768	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.	822
769		823
770		
771		
772		
773	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.	824
774		825
775		826
776		827
777	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. <i>arXiv preprint arXiv:1609.04836</i> .	828
778		829
779		
780		
781		
782	Muhammad Osama Khan and Yi Fang. 2023. Revisiting fine-tuning strategies for self-supervised medical imaging analysis.	830
783		831
784		832
785	Faeze Kiani. 2022. Texture features in medical image analysis: a survey.	833
786		
787	Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. 2022. Self-supervised learning in medicine and healthcare. <i>Nature Biomedical Engineering</i> , 6(12):1346–1352.	833
788		834
789		835
790		
791	Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.	836
792		837
793		838
794		839
795	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 <i>NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications</i> .	840
796		841
797		842
798		843
799		
800		
801	Haibo Li, Guorong Wu, and Qian Wang. 2019. Multi-modal medical image fusion: A survey. In <i>Fusion in Computer Vision</i> , pages 53–78. Springer.	844
802		845
803		846
804	Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. <i>IEEE transactions on knowledge and data engineering</i> , 35(1):857–876.	847
805		
806		
807		
808		
809	Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In <i>Proceedings of the European Conference on Computer Vision (ECCV)</i> .	850
810		851
811		
812		
813	Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak N. Patel. 2021. Reliable and trustworthy machine learning for health using dataset shift detection. <i>CoRR</i> , abs/2110.14019.	852
814		853
815		854
816		
817	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, <i>Resource-Efficient Medical Image Analysis</i> , volume 13543, pages 85–94. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.	855
818		856
819		857
820		858
821		
822		
823		
824	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. <i>2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 1–9.	859
825		860
826		861
827		862
828		863
829		864
830	Saeed Shurrah and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. 8:e1045.	865
831		866
832		867
833	Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. <i>arXiv preprint arXiv:1409.1556</i> .	868
834		869
835		870
836	Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In <i>2017 IEEE Winter Conference on Applications of Computer Vision (WACV)</i> , pages 464–472. IEEE.	871
837		872
838		873
839		874
840	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.	875
841		876
842		877
843		878
844	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.	879
845		880
846		881
847		882
848	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.	883
849		884
850		885
851		886
852	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.	887
853		888
854		889
855	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.	890
856		891
857		892
858		893
859	Jing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In <i>International Conference on Machine Learning (ICML)</i> .	894
860		895
861		896
862		897
863		898
864		899
865	Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. <i>Journal of Machine Learning Research</i> , 9:2579–2605.	900
866		901
867		902
868		903
869		904
870		905
871	Jiashu Xu. 2021. A review of self-supervised learning methods in the field of medical image analysis. <i>International Journal of Image, Graphics and Signal Processing</i> .	906
872		907
873		908
874		909
875		910
876		911
877		912
878		913
879		914
880		915
881		916
882		917
883		918
884		919
885		920
886		921
887		922
888		923
889		924
890		925
891		926
892		927
893		928
894		929
895		930
896		931
897		932
898		933
899		934
900		935
901		936
902		937
903		938
904		939
905		940
906		941
907		942
908		943
909		944
910		945
911		946
912		947
913		948
914		949
915		950
916		951
917		952
918		953
919		954
920		955
921		956
922		957
923		958
924		959
925		960
926		961
927		962
928		963
929		964
930		965
931		966
932		967
933		968
934		969
935		970
936		971
937		972
938		973
939		974
940		975
941		976
942		977
943		978
944		979
945		980
946		981
947		982
948		983
949		984
950		985
951		986
952		987
953		988
954		989
955		990
956		991
957		992
958		993
959		994
960		995
961		996
962		997
963		998
964		999
965		1000

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

868 Zhirong Zhang, Yoshua Bengio, Moritz Hardt, Ben-
869 jamin Recht, and Oriol Vinyals. 2020. A compre-
870 hensive overview of self-supervised learning. *arXiv*
871 *preprint arXiv:2006.10029.*

872

A Appendix

A.1 Performance results

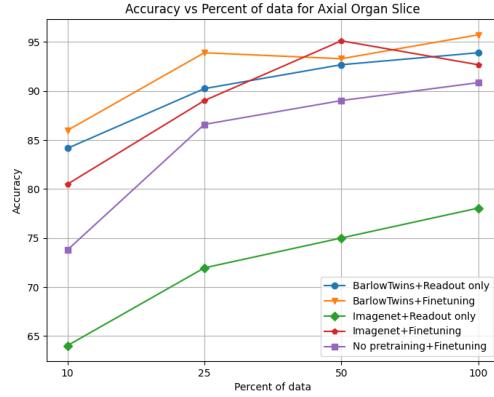


Figure 9: Accuracy on the axial organ slice dataset with different percentages of data used with different fine-tuning strategies

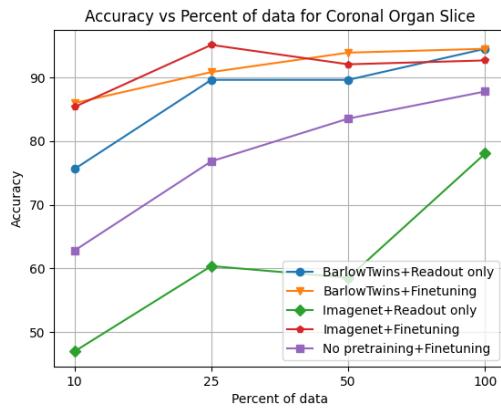


Figure 10: Accuracy on the coronal organ slice dataset with different percentages of data used with different fine-tuning strategies

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.

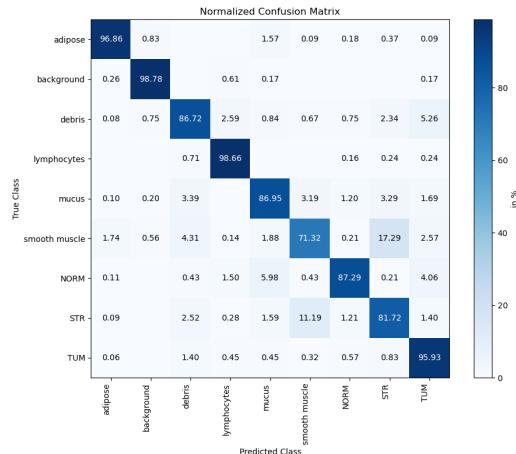


Figure 11: Normalized confusion matrix with ImageNet pre-training and fine-tuning the whole model of 0.5 percent of colorectal cancer data

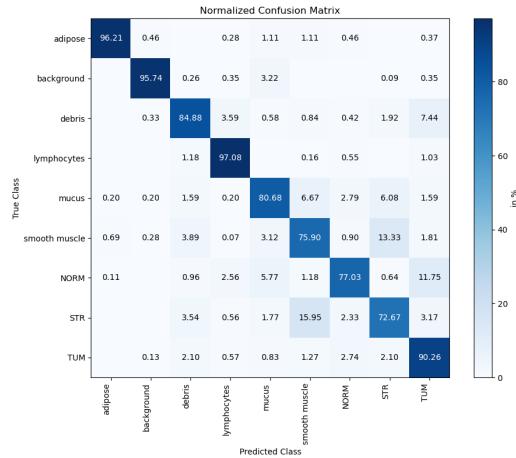


Figure 12: Normalized confusion matrix with ImageNet pre-training and only readout head fine-tuning of 0.5 percent of colorectal cancer data

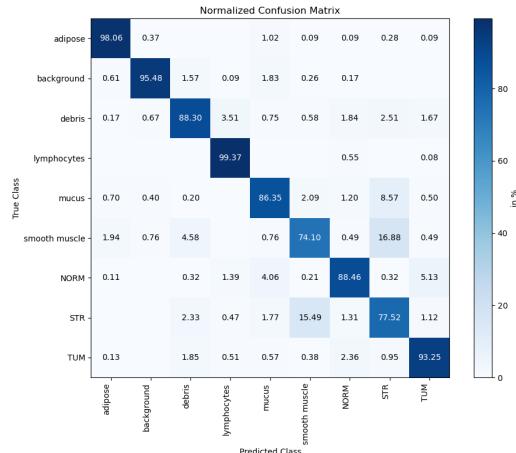


Figure 13: Normalized confusion matrix with Barlow Twins pre-training and fine-tuning the whole model of 0.5 percent of colorectal cancer data

880

A.2 Visual feature analysis

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

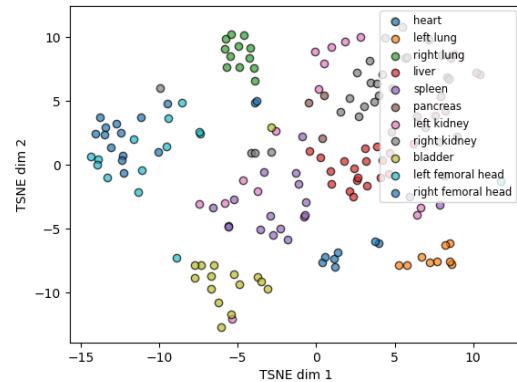


Figure 14: TSNE projection of axial organ slice feature vectors obtained from Barlow Twins pre-trained ResNet-18 with fine-tuned readout head.

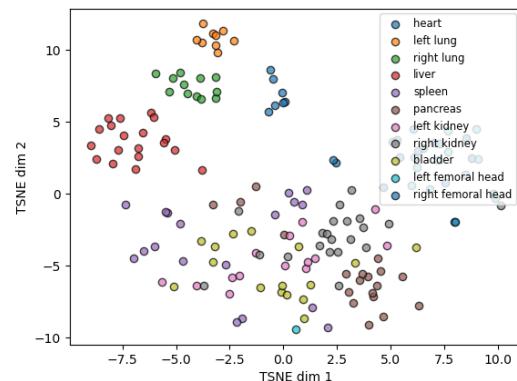


Figure 15: TSNE projection of axial organ slice feature vectors obtained from ImageNet based pre-trained ResNet-18 with fine-tuned readout head.

883