Self-Supervised Learning for Accurate Liver View Classification in Ultrasound Images with Minimal Labeled Data

Abder-Rahman Ali* Anthony E. Samir† Peng Guo‡
Harvard Medical School and Massachusetts General Hospital
Boston, MA 02114 USA

Abstract

Conventional B-mode "grey scale" medical ultrasound and shear wave elastography (SWE) are widely used for chronic liver disease diagnosis and risk stratification. Liver disease is very common and is clinically and socially important. As a result, multiple medical device manufacturers have proposed or developed AI systems for ultrasound image analysis. However, many abdominal ultrasound images do not include views of the liver, necessitating manual data curation for model development. To optimize the efficiency of real-time processing, a pre-processing liver view detection step is necessary before feeding the image to the AI system. Deep learning techniques have shown great promise for image classification, yet labeling large datasets for training classification models is time-consuming and expensive. In this paper, we present a self-supervised learning method for image classification that utilizes a large set of unlabeled abdominal ultrasound images to learn image representations. These representations are then applied on the downstream task of liver view classification, resulting in efficient classification and alleviation of the labeling burden. In comparison to two state-of-the-art (SOTA) models, ResNet-18 and MLP-Mixer, when trained for 100 epochs the proposed SimCLR+LR approach demonstrated outstanding performance when only labeling "one" image per class, achieving an accuracy similar to MLP-Mixer (86%) and outperforming the performance of ResNet-18 (70.2%), when trained on 854 (with liver: 495, without liver: 359) B-mode images. When trained on the whole dataset for 1000 epochs, SimCLR+LR and ResNet-18 achieved an accuracy of 98.7% and 79.3%, respectively. These findings highlight the potential of the SimCLR+LR approach as a superior alternative to traditional supervised learning methods for liver view classification. Our proposed method has the ability to reduce both the time and cost associated with data labeling, as it eliminates the need for human labor (i.e., SOTA performance achieved with only a small amount of labeled data). The approach could also be advantageous in scenarios where a subset of images with a particular organ needs to be extracted from a large dataset that includes images of various organs.

*aali25@mgh.hrvard.edu
†asamir@mgh.harvard.edu
‡pguo@mgh.harvard.edu

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a prevalent form of chronic liver disease characterized by the accumulation of excess fat in the liver, leading to damage and inflammation. The incidence of NAFLD is predicted to rise from 20.8% to 22.9% in Canada between 2019 and 2030 [11], while the overall prevalence of NAFLD in the U.S. is estimated to be 24% [1]. The upward trend in incidence is likely to result in an increase in the economic burden, which is already high, with annual NAFLD-related US medical costs estimated to be $103 billion [12]. The stage of liver fibrosis at the time of diagnosis has been shown to be the best indicator of negative outcomes for patients with excess liver fat [8]. However, liver biopsy, the current standard for detecting excess fat and fibrosis stage, is invasive, costly, and subject to sampling error and interpretative variability. Non-invasive alternatives, such as medical ultrasound, have been developed to address these limitations. Shear wave elastography (SWE) is a non-invasive ultrasound method that can measure changes in liver stiffness as liver fibrosis progresses, making it a useful biomarker for NAFLD diagnosis. It has been shown to be highly effective in diagnosing cirrhosis and moderately accurate for intermediate fibrosis stages in adults with NAFLD [13]. Proper placement of a region of interest (ROI) on the liver in the B-mode image is crucial for obtaining accurate and reliable results in SWE. The first step in developing an automated system for measuring liver fibrosis using SWE would thus be to segment the liver capsule. The use of SWE technology for liver fibrosis staging has been widely adopted by multiple medical ultrasound manufacturers, which has led to the development of AI systems designed to augment SWE and liver disease detection. However, not all abdominal ultrasound images

contain images of the liver, making it infeasible to automatically pass all images to an AI system for processing. A preprocessing step to check for the presence of the liver view in the image is necessary before feeding the image to the AI system. This step could improve the overall efficiency and effectiveness of the AI system in processing abdominal ultrasound images in a real-time setting. Deep learning has emerged as a powerful tool for image classification, and while various deep learning models have been proposed for this purpose, one of the main challenges in using such models for medical imaging is the significant expense (i.e., both in terms of time and cost) of labeling large datasets for training the classification model. In this paper, we introduce a contrastive self-supervised learning method for image classification that leverages a large dataset of unlabeled abdominal ultrasound images to learn image representations that are then applied to the downstream task of liver view classification, enabling effective classification, and reduction of the labeling burden. The remainder of the paper is presented as follows: Section 2 provides a survey of relevant literature. The proposed method is described in Section 3. The results of the proposed approach are analyzed and compared with SOTA methods in Section 4. Finally, the paper is concluded in Section 5.

2. Related work

Within the scope of the related work, it is noteworthy to mention that the literature presents a rather limited number of studies that have investigated and focused on the classification of abdominal organs using ultrasound images. Li et. al., [6] introduce an automatic abdominal organ recognition method in ultrasound images using deep neural networks (DNNs) and k-nearest-neighbor (k-NN) classification. The technique achieves high classification accuracy (96.67%) by combining fine-tuned ResNet-50 and DenseNet-121 feature extractors with k-NN and FC layer classifiers. The incorporation of principal component analysis (PCA) improves classification results by reducing overfitting and eliminating correlated features. The method effectively handles challenging cases and offers real-time performance. However, potential limitations include limited generalizability due to the small dataset, possible overfitting with fine-tuned networks, and dependence on hyperparameter choices. In [7], Reddy et. al. introduced a transfer learning-based framework for abdominal organ classification in ultrasound images, utilizing pretrained models such as AlexNet, VGGNet, GoogleNet, Inception, and ResNet. The approach demonstrated high performance, especially with ResNet-50, which achieved an average precision, recall, F1 score, and classification accuracy of 98.77%, 98.55%, 98.55%, and 98.77%, respectively. Despite its success, the study faced some challenges, such as the use of a relatively small dataset of 1906 images, which could impact generalizability. Furthermore, biases introduced by excluding images with unclear anatomical locations or superimposed annotations might affect the model's real-world applicability. Dadoun et al. [3] developed a framework for multi-label classification of abdominal organs in ultrasound images by adapting deep clustering with PICA [4] and semi-supervised learning with FixMatch [10]. The approach demonstrated better feature transferability than ImageNet initialization and achieved robust results with limited labeled examples. Key advantages include the use of a large database and the effective combination of self-supervised and semi-supervised learning. However, limitations encompass transforming multi-label classification into single-label classification, the introduction of an additional hyperparameter, and a limited evaluation of methods and datasets, necessitating further research.

3. Methodology

The proposed method (Fig.1) involves a two-stage training approach. In the first stage, a contrastive self-supervised learning technique (i.e., SimCLR [2]) is trained on a large dataset of unlabeled abdominal ultrasound images to obtain feature representations of the images (these feature representations will be fine-tuned on a downstream classification task utilizing logistic regression). In the second stage, a supervised learning approach (logistic regression) is used to train the downstream task, which involves a substantially smaller dataset (as compared to the first stage) of images with labeled liver view classes (i.e., with liver vs. without liver). We call our approach SimCLR+LR (LR: logistic regression).

3.1. SimCLR+LR

3.1.1 SimCLR framework

SimCLR (simple framework for contrastive learning of visual representations) is a framework for learning effective representations of data. The key idea behind SimCLR is to use a contrastive learning approach to learn representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. The framework comprises four major components, which we describe herein.

Stochastic data augmentation. SimCLR employs a stochastic data augmentation module that transforms any given data example randomly, resulting in two correlated views of the same example, denoted $\widetilde{x}_i$ and $\widetilde{x}_j$, which we consider as a positive pair. In their work, the authors sequentially apply three simple augmentations: random cropping followed by resize back to the original size, random color distortions, and random Gaussian blur. The authors emphasize that the combination of random crop and color distortion is crucial to achieving good performance. Fig.2

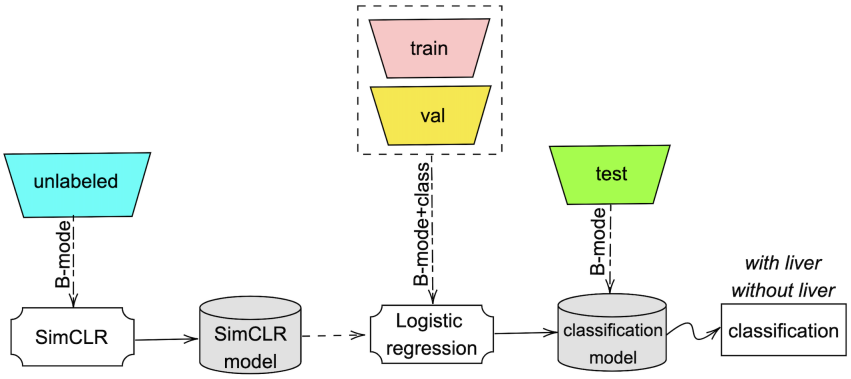

Figure 1. Proposed liver view classification approach

shows some examples on image pairs sampled with SimCLR augmentations applied on the images in the unlabeled dataset.

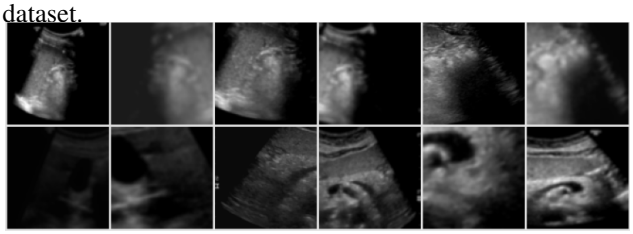

Figure 2. Augmented image pair samples of the abdominal ultrasound unlabeled dataset

Base encoder. The framework uses a neural network base encoder $f(\cdot)$ to extract representation vectors from the augmented data examples. The framework allows various choices of the network architecture without any constraints. The authors opt for simplicity and adopt the commonly used ResNet architecture to obtain $h_i = f(\widetilde{x}_i) = \text{ResNet}(\widetilde{x}_i)$, where $h_i \in \mathbb{R}^d$ is the output after the average pooling layer.

Projection head. A small neural network projection head $g(\cdot)$ maps representations to the space where contrastive loss is applied. The authors use a multi-layer perceptron (MLP) with one hidden layer to obtain $z_i = g(h_i) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}h_i)$, where σ is a rectified linear unit (ReLU) nonlinearity. The authors found that it beneficial to define the contrastive loss on the z_i's rather than the h_i's.

Contrastive loss function. The contrastive loss function is an essential component of the SimCLR framework, defined for a contrastive prediction task.. Given a set $\{\widetilde{x}_k\}$ including a positive pair of examples $\widetilde{x}_i$ and $\widetilde{x}_j$, the contrastive prediction task aims to identify $\widetilde{x}_j$ in $\{\widetilde{x}_k\}_{k\neq i}$ for a given $\widetilde{x}_i$. The contrastive loss function is defined as a softmax over the cosine similarities between the z_i's for the positive pair and the z_i's for the other examples in the batch:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(sim(z_i, z_k)/\tau)} \quad (1)$$

where z_i and z_j are the projections of the augmented views $\widetilde{x}_i$ and $\widetilde{x}_j$ of a given data example, obtained by passing them through the base encoder $f(\cdot)$ and the projection head $g(\cdot)$, and τ is a temperature parameter that controls the concentration of the distribution. The function $sim(z_i, z_j)$ measures the cosine similarity between the projections z_i and z_j and is defined as:

$$sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \cdot \|z_j\|} \quad (2)$$

where $\|z_i\|$ and $\|z_j\|$ represent the Euclidean norm of the vectors z_i and z_j, respectively.

The authors investigate the impact of model size and the use of a nonlinear projection head on the quality of learned representations. They find that, while increasing depth and width improves performance, unsupervised learning benefits more from bigger models than its supervised counterpart. They also show that a nonlinear projection head is better than no projection or a linear projection, and that the layer before the projection head is a better representation than the layer after. The authors conjecture that this is because the contrastive loss can induce a loss of information, and leveraging the nonlinear transformation of the projection head can help maintain more information in the layer before it. They conduct experiments to verify this hypothesis and find that the hidden layer before the projection head contains much more information about the transformation applied during pretraining than the layer after. The authors compared the NT-Xent loss, logistic loss, and margin loss, and found that the NT-Xent loss with adjustable temperature works the best. They also tested the importance of ℓ_2 normalization and temperature scaling in the NT-Xent loss and found that they are essential for achieving good performance. In terms of batch size, the authors found that larger batch sizes are beneficial for contrastive learning, particularly with shorter training periods. This is because larger batch sizes provide more negative examples, which helps facilitate convergence. Additionally, longer training periods

also provide more negative examples, improving results. The paper compares its approach to state-of-the-art methods using ResNet-50 in 3 different hidden layer widths. In linear evaluation, the results show that the paper's method outperforms previous approaches that require specifically designed architectures. In semi-supervised learning, the paper's approach significantly improves over state-of-the-art with both 1% and 10% of the labels. The transfer learning results show that when fine-tuned, the paper's self-supervised model significantly outperforms the supervised baseline on 5 datasets.

In our work, we use ResNet-18 as the base encoder network for SimCLR due to several factors, including computational efficiency, faster training, reduced overfitting, transferability, and suitability for self-supervised learning. By employing ResNet-18 (a less complex architecture than ResNet-50) we can reduce computational demands and accelerate training, making our model more practical for real-world clinical applications. Additionally, a shallower network like ResNet-18 mitigates the risk of overfitting on limited labeled data and enhances generalization, which is crucial in medical imaging domains where obtaining large amounts of labeled data is challenging.

3.1.2 Logistic regression for downstram task

Logistic regression [5] is a popular statistical technique used in supervised learning for binary classification tasks such as image classification. In the SimCLR+LR methodology, logistic regression is employed as the supervised learning approach for the downstream task, which is the classification of liver view classes (with liver vs. without liver). This task leverages the feature representations obtained from the unsupervised learning stage using the SimCLR framework. By utilizing logistic regression, the fine-tuning process adjusts the learned feature representations, enabling the model to predict the presence or absence of the liver in abdominal ultrasound images. The combination of SimCLR for unsupervised feature learning and logistic regression for supervised fine-tuning allows the method to benefit from the large-scale unlabeled dataset in the first stage and achieve accurate classification results with a smaller labeled dataset in the second stage (i.e., logistic regression).

4. Results and discussion

Liver ultrasound DICOM exams were collected from the electronic medical record (EMR) of the Massachusetts General Hospital (MGH). In particular, training, validation, and testing data were sampled from 903 abdominal ultrasound exams (each exam contained a sequence of DICOM files); the sampled images were labeled by MGH radiologists as either belonging to the "with liver" class (i.e., liver capsule can be viewed in the B-mode image) or "without liver" class (i.e., liver capsule is not viewable) to be used in the supervised stage (i.e., logistic regression) of the proposed approach. The self-supervised learning part (i.e., SimCLR) was trained on 17,441 unlabeled abdominal B-mode images, and the trained model was subsequently fine-tuned on a downstream task of liver view classification in ultrasound images involving logistic regression. The labeled training and validation B-mode images consisted of 854 (with liver: 495, without liver: 359) and 90 (with liver: 51, without liver: 39) images, respectively. Experiments were performed on a test set of 150 (with liver: 101, without liver: 49) images to evaluate the proposed approach. Fig.3 displays examples of the training data and corresponding classes.

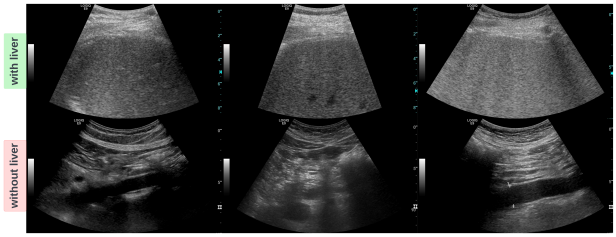

Figure 3. Samples of training images and their associated class: "with liver" or "without liver"

This section commences with an analysis of the dataset utilized in our research, followed by a comparative evaluation of our proposed approach against ResNet-18 and MLP-Mixer.

4.1. Data analysis

The goal of our study is to classify liver images into two classes: "with liver" for images where the liver capsule can be viewed, and "without liver" for images where the liver capsule is not viewable. To form the class mean (prototype image) of our training dataset, we computed the arithmetic mean of the pixel values of all the images in the class. This eventually results in an image that emphasizes the distinctive features that are common to all images in the class. The prototype images for the "with liver" and "without liver" classes are shown in Fig.4.a. and Fig.4.b.

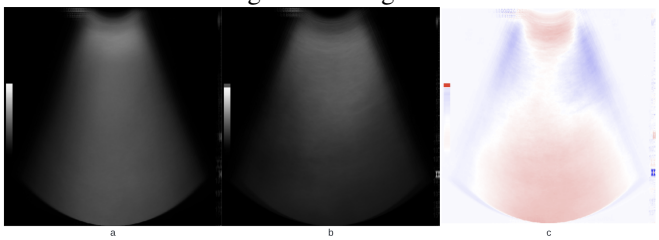

Figure 4. Prototype image: (a) "with liver" class (b) "without liver" class (c) difference between (a) and (b)

The prototype image for the "with liver" class, which appears homogeneous, indicates that the majority of the im-

ages in this class show a clear view of the liver with few artifacts or noise. Conversely, the prototype image for the "without liver" class, which appears less homogeneous with dark areas, suggests that the majority of the images in this class do not show the liver, and instead contain other organs or tissue. The difference image (Fig.4.c.) represents the differences between the prototype images of the two classes, "with liver" and "without liver". The blue-white-red (BWR) colormap was used to create the difference image, where positive values are displayed in red, negative values in blue, and the zero crossing in white. The difference image provides insight into the features that are most distinctive between the "with liver" and "without liver" classes in the training dataset. In particular, regions that appear in red correspond to features that are more prevalent in the "with liver" class, while regions that appear in blue correspond to features that are more prevalent in the "without liver" class. Regions that appear in white may correspond to features that are common to both classes or to areas where the difference between the two classes is minimal (i.e., non-distinctive features of either class).

4.2. Evaluation

Our experiments were performed on a NVIDIA GeForce RTX 2080-Ti GPU. The initial set of experiments aimed to evaluate the impact of the number of training images per class and the number of epochs per experiment on the liver view classification performance. Fig.5 depicts the impact of training duration and number of images per class (i.e., randomly selected) on liver view classification accuracy. The figure shows that increasing the number of training epochs results in higher classification accuracy. Moreover, it can be observed that even with a limited number of labeled images, remarkable classification performance can still be accomplished. Notably, the optimal classification accuracy was obtained when 50 images were used per label. Specifically, employing this configuration resulted in liver view classification accuracies of 95.3%, 96%, and 99.3% when trained for 100, 500, and 1000 epochs, respectively (summarized in Table 1). The outstanding performance achieved with 50 images per label can be attributed to two main factors: (i) the absence of similar features between the two classes (i.e., Fig.4) in the images used in this experiment: when training a model, the features present in the training data are crucial in determining how well the model can distinguish between different classes. If the features in the training images used in the experiment are highly distinctive and not similar between the two classes, it can make it easier for the logistic regression model to accurately classify new images (i.e., test images); (ii) the capacity of the images utilized for training to optimize the accuracy of logistic regression: in this case, having as few as 50 images per class can allow the model to learn and generalize effectively since the training data set is composed of representative examples (i.e., this is a topic we intend to investigate in future work related to active learning [9]).

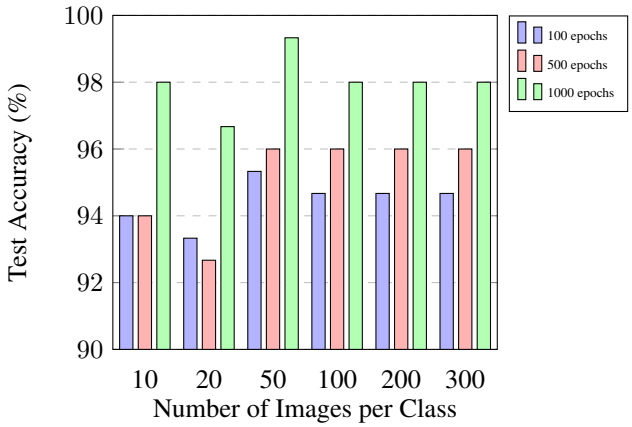

Figure 5. Liver view classification performance of SimCLR+LR using different training dataset sizes per class and varying number of epochs

Epochs	Liver view classification accuracy
100	95.3%
500	96.0%
1000	99.3%

Table 1. Classification accuracies of liver views using SimCLR+LR with varying training epochs (based on 50 labeled images per class)

In the process of evaluating the proposed approach, we compared it to two state-of-the-art (SOTA) models, namely ResNet-18 and MLP-Mixer. This evaluation was carried on the original dataset where training, validation, and testing datasets consisted of 801 (with liver: 495, without liver: 306), 90 (with liver: 51, without liver: 39), and 150 (with liver: 101, without liver: 49) B-mode images, respectively. Based on the findings illustrated in Fig.6 (summarized in Table 2), it is apparent that the SimCLR+LR approach outperforms ResNet-18 by a significant margin across different epochs. Specifically, upon training the models for 100 epochs, the liver view classification accuracy achieved by the SimCLR+LR method was 94.5%, which is substantially higher than the accuracy of 70.2% obtained by ResNet-18. Furthermore, after training the models for 500 epochs, the SimCLR+LR method exhibited an accuracy of 97.3%, while ResNet-18 achieved an accuracy of 78.7%. Finally, after training for 1000 epochs, the SimCLR+LR approach demonstrated an impressive accuracy of 98.7%, which significantly surpassed the 79.3% accuracy obtained by ResNet-18.

After training the models (SimCLR+LR, ResNet-18, and MLP-Mixer) for 100 epochs, the liver view classification accuracy of images belonging to the "with liver" and "with-

Model	100 epochs	500 epochs	1000 epochs
SimCLR+LR	94.5%	97.3%	98.7%
ResNet-18	70.2%	78.7%	79.3%

Table 2. Comparison of liver view classification accuracy between SimCLR+LR and ResNet-18 across different epochs (when trained on the whole labeled dataset)

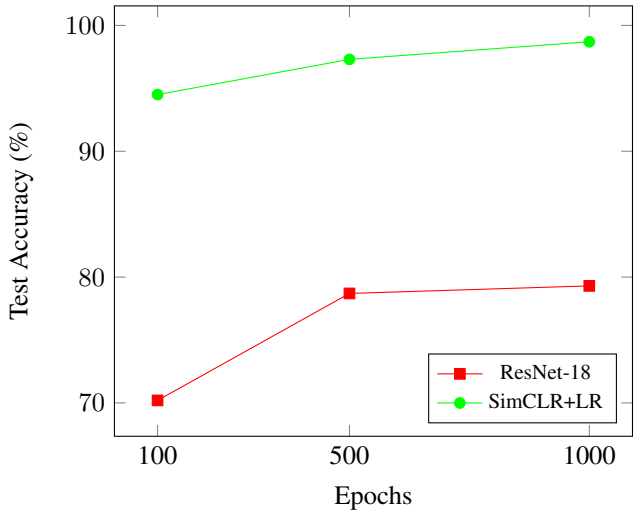

Figure 6. Overall accuracy of SimCLR+LR and ResNet-18 across different epochs

out liver" classes was evaluated in a per-class comparison (Fig.7 and Fig.8). The results show that the SimCLR+LR approach outperforms both ResNet-18 and MLP-Mixer for both classes. Specifically, the SimCLR+LR approach achieved an accuracy of 95.1% for images categorized as "with liver", while ResNet-18 and MLP-Mixer achieved 79.2% and 82.2%, respectively. Similarly, for the "without liver" class, the SimCLR+LR approach demonstrated an accuracy of 93.9%, which is higher than the accuracy of ResNet-18 and MLP-Mixer at 61.2% and 89.8%, respectively. Overall, the accuracy for the three models is 94.5%, 70.2%, and 86% for SimCLR+LR, ResNet-18, and MLP-Mixer, respectively. Table 3 summarizes these findings. We would like to emphasize that training SimCLR+LR with only "one" labeled image for each class (i.e., 2 images in total) achieved a performance similar to MLP-Mixer (86%), which has been trained on the entire dataset comprising 801 B-mode images (with liver: 495, without liver: 306). These results highlight the potential of the SimCLR+LR approach as a viable alternative to traditional supervised learning methods for liver view classification.

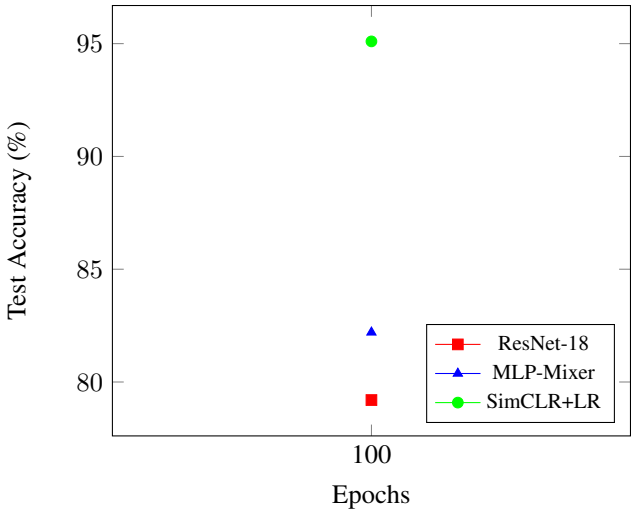

Figure 7. "with liver" class accuracy of SimCLR+LR, ResNet-18, and MLP-Mixer when trained for 100 epochs

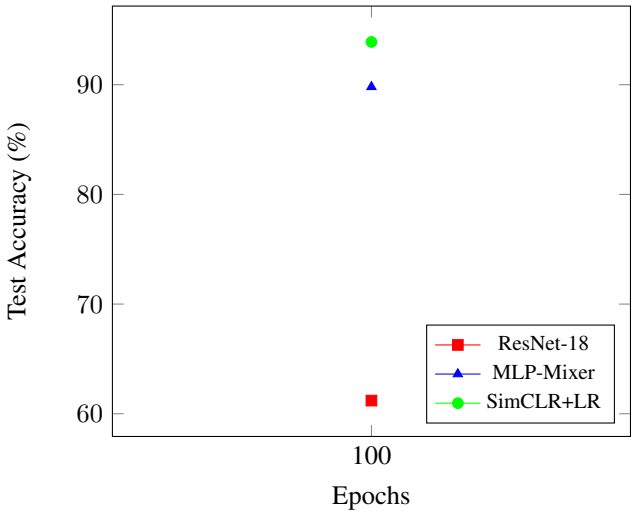

Figure 8. "without liver" class accuracy of SimCLR+LR, ResNet-18, and MLP-Mixer when trained for 100 epochs

Model	WithLiver	WithoutLiver	Overall accuracy
SimCLR+LR	95.1%	93.9%	94.5%
ResNet-18	79.2%	61.2%	70.2%
MLP-Mixer	82.2%	89.8%	86.0%

Table 3. Liver view classification accuracy evaluated in a per-class comparison for SimCLR+LR, ResNet-18, and MLP-Mixer when trained for 100 epochs

5. Conclusion

We proposed a novel two-stage approach for liver view classification in abdominal ultrasound images called SimCLR+LR. Our approach achieved high accuracy even with a limited number of labeled images and outperformed the

state-of-the-art models, ResNet-18 and MLP-Mixer. The proposed approach involves a contrastive self-supervised learning technique in the first stage, which was trained on a large dataset of unlabeled abdominal ultrasound images to obtain feature representations of the images. In the second stage, the learned feature representations were fine-tuned on a supervised learning approach (i.e., logistic regression) which was trained on a smaller labeled dataset of liver view classes. Our proposed approach has the potential to save time and cost when labeling data as it does not require human labor (i.e., few labeled data required to reach SOTA performance). It could also be beneficial in situations where a set of images with a specific organ need to be selected from a large data set that contains images of different organs. This could save time and resources by allowing researchers or medical professionals to quickly and efficiently identify and access the images they need, rather than having to manually sift through a vast collection of images. In general, self-supervised learning presents a promising substitute for transfer learning strategies that rely on ImageNet pre-trained models. By pre-training a model on a domain-specific dataset, this can potentially yield improved performance and applicability to the target task. Future work includes investigating active learning as a data-centric AI method, and expanding the proposed approach to multi-class organ view classification.

References

[1] Tamoore Arshad, Pegah Golabi, Linda Henry, and Zobair M Younossi. Epidemiology of non-alcoholic fatty liver disease in north america. *Current Pharmaceutical Design*, 26(10):993–997, 2020. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[3] Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler, and Nicholas Ayache. Deep clustering for abdominal organ classification in us imaging. 2022. 2

[4] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8849–8858, 2020. 2

[5] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002. 4

[6] Keyu Li, Yangxin Xu, Ziqi Zhao, and Max Q-H Meng. Automatic recognition of abdominal organs in ultrasound images based on deep neural networks and k-nearest-neighbor classification. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1980–1985. IEEE, 2021. 2

[7] D Santhosh Reddy, P Rajalakshmi, and MA Mateen. A deep learning based approach for classification of abdominal organs using ultrasound images. *Biocybernetics and Biomedical Engineering*, 41(2):779–791, 2021. 2

[8] Arun J Sanyal, Mark L Van Natta, Jeanne Clark, Brent A Neuschwander-Tetri, AnnaMae Diehl, Srinivasan Dasarathy, Rohit Loomba, Naga Chalasani, Kris Kowdley, Bilal Hameed, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. *New England Journal of Medicine*, 385(17):1559–1569, 2021. 1

[9] Burr Settles. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 10:S00429ED1V01Y201207AIM018, 2012. 5

[10] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[11] Mark G Swain, Alnoor Ramji, Keyur Patel, Giada Sebastiani, Abdel Aziz Shaheen, Edward Tam, Paul Marotta, Magdy Elkhashab, Harpreet S Bajaj, Chris Estes, et al. Burden of nonalcoholic fatty liver disease in canada, 2019–2030: a modelling study. *Canadian Medical Association Open Access Journal*, 8(2):E429–E436, 2020. 1

[12] Michal Witkowski, Søren Ilsøe Moreno, João Fernandes, Pierre Johansen, Margarida Augusto, and Sunita Nair. The economic burden of non-alcoholic steatohepatitis: a systematic review. *PharmacoEconomics*, pages 1–26, 2022. 1

[13] Guangqin Xiao, Sixian Zhu, Xiao Xiao, Lunan Yan, Jiayin Yang, and Gang Wu. Comparison of laboratory tests, ultrasound, or magnetic resonance elastography to detect fibrosis in patients with nonalcoholic fatty liver disease: a meta-analysis. *Hepatology*, 66(5):1486–1501, 2017. 1