CROSS-DOMAIN SELF-SUPERVISED DEEP LEARNING FOR ROBUST ALZHEIMER'S DISEASE PROGRESSION MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing successful artificial intelligence systems in practice depends both on robust deep learning models as well as large high quality data. Acquiring and labeling data can become prohibitively expensive and time-consuming in many real-world applications such as clinical disease models. Self-supervised learning has demonstrated great potential in increasing model accuracy and robustness in small data regimes. In addition, many clinical imaging and disease modeling applications rely heavily on regression of continuous quantities. However, the applicability of self-supervised learning for these medical-imaging regression tasks has not been extensively studied. In this study, we develop a cross-domain self-supervised learning approach for disease prognostic modeling as a regression problem using 3D images as input. We demonstrate that self-supervised pre-training can improve the prediction of Alzheimer's Disease progression from brain MRI. We also show that pre-training on extended (but not labeled) brain MRI data outperforms pre-training on natural images. We further observe that the highest performance is achieved when both natural images and extended brain-MRI data are used for pre-training.

1 INTORDUCTION

Developing reliable and robust artificial intelligence systems requires advanced and efficient deep learning techniques, as well as the creation and annotation of large volumes of data for training. However, the construction of a labeled dataset is often time-consuming and expensive, such as in the medical imaging domain, due to the complexity of annotation tasks and the high expertise required for the manual interpretation of medical images. To alleviate the lack of annotations in medical imaging, transfer learning from natural images is becoming popular ((Liu et al., 2020b), (McKinney et al., 2020), (Menegola et al., 2017), (Xie et al., 2019)). Although numerous experimental studies indicate the effectiveness of fine-tuning from either supervised or self-supervised ImageNet models ((Alzubaidi et al., 2020), (Graziani et al., 2019), (Heker & Greenspan, 2020), (Zhou et al., 2021),(Hosseinzadeh Taher et al., 2021),(Azizi et al., 2021)), it does not always improve the performance due to domain mismatch problem ((Raghu et al., 2019)).

In the meantime, self-supervised learning has demonstrated great success in many downstream applications in computer vision, where the labeling process is quite expensive and time-consuming((Doersch et al., 2015), (Gidaris et al., 2018), (Noroozi & Favaro, 2016), (Zhang et al., 2016),(Ye et al., 2019), (Bachman et al., 2019),(Tian et al., 2020a), (Henaff, 2020), (Oord et al., 2018)]. Medical research and healthcare are especially well-poised to benefit from self-supervised learning approaches, given the prevalence of unprecedented amounts of medical images generated by hospital and non-hospital settings. Despite demand, the use of self-supervised approaches in the medical image domain has received limited attention. Only few studies have investigated the impact of self-supervised learning in the medical image analysis domain for limited applications including classification ((Liu et al., 2019), (Sowrirajan et al., 2021),(He et al., 2020b),(Azizi et al., 2022),(Zhu et al., 2020),(Liu et al., 2020a)) and segmentation ((Ronneberger et al., 2015),(Bai et al., 2019),(Chaitanya et al., 2020),(Spitzer et al., 2018)).

In this study, we focus on developing a self-supervised deep learning model for predicting the progression of Alzheimer's disease (AD). AD is a slowly progressing disease caused by the degeneration of brain cells, with patients showing clinical symptoms years after the onset of the disease. Therefore, accurate prognosis and treatment of AD in its early stage is critical to prevent non-reversible and fatal brain damage. Prognostic prediction, modeled as a regression task rather than a classification, helps provide more granular estimates of disease progression. However, little progress has been made because of the difficult design requirements and, lack of large-scale and homogeneous datasets, and noisy endpoints that are potentially hard to predict.

We establish a cross-domain self-supervised transfer learning approach for medical image analysis that leverages self-supervised learning on both unlabeled large-scale natural images and domain-specific medical imaging to learn transferable and generalizable representations for medical images. Such representations can be further fine-tuned and deployed for downstream tasks such as disease progress prediction using limited labeled data from the clinical setting. Our transfer learning strategy benefits from both large-scale self-supervised representations and domain-specific self-supervised representations in a unified deep learning framework. We develop a detailed and extensive framework to evaluate the performance of our model fine-tuned from different supervised and self-supervised models pre-trained on either natural images or medical images, or both. Our extensive experimental results reveal that (1) self-supervised pre-training on natural images followed by intermediate self-supervised learning on unlabeled domains-specific medical data outperforms all other transfer learning techniques, which is very promising because replacing supervised approaches with self-supervised learning can significantly reduce the need for data annotation (2) self-supervised models pre-trained on medical images outperform same models pre-trained on natural images indicating that self-supervised learning on medical images yield discriminative feature representations for regression task.

2 Related works

The recent development and success of self-supervised learning techniques, including contrastive learning ((Wu et al., 2018),(He et al., 2020a),(Chen et al., 2020c),(Chen et al., 2020b),(Chen et al., 2020a),(Grill et al., 2020),(Misra & Maaten, 2020)), mutual information reduction ((Tian et al., 2020b)), clustering ((Caron et al., 2020),(Li et al., 2020)), and redundancy-reduction methods ((Zbontar et al., 2021), (Bardes et al., 2021)) in computer vision indicate their effectiveness in improving the performance of AI systems. These methods train models on different pretext tasks to enable the network to learn high-quality representations without label information. SimCLR (Chen et al., 2020c) maximizes agreement between representations of different augmentations of the same image by using a contrastive loss in the latent space. Barlow Twins (Zbontar et al., 2021) measure the cross-correlation matrix between the embedding of two identical networks and its goal is to make this cross-correlation close to the identity matrix. SwAV (Caron et al., 2020) simultaneously clusters the images while enforcing consistency between cluster assignments produced for differently augmented views of the same image, instead of comparing features directly as in contrastive learning.

Subsequently, self-supervised learning has been employed for medical imaging applications including classification and segmentation to learn visual representations of medical images by incorporating unlabeled medical images. While some approaches have designed domain-specific pretext tasks (Bai et al., 2019), (Spitzer et al., 2018),(Zhuang et al., 2019),(Zhu et al., 2020)], others have adjusted well-known self-supervised learning methods to medical data (He et al., 2020b),(Li et al., 2021),(Zhou et al., 2020),(Sowrirajan et al., 2021). Very recently (Azizi et al., 2022) has applied SimCLR on a combination of unlabeled ImageNet dataset and task specific medical images for medical image classification; their experiments and improved performance suggest that pre-training on ImageNet is complementary to pre-training on unlabeled medical images.

Although aforementioned approaches demonstrate improvement of the performance on challenging medical datasets, all of them are limited to classification and segmentation tasks and their benefits and potential effects for the prognosis prediction tasks, as regression tasks, have not been studied. Formulating progress prediction as a regression rather than a traditional classification problem leads to a more fine-grained measurement scale which is crucial for real-world applications. Therefore, the development of self-supervised networks is in great demand for efficient data-utilization in medical

imaging for disease prognosis. To the best of our knowledge, this is the first study of developing a self-supervised deep convolution neural network on medical data images from various cross-domain datasets to predict a granular understanding of disease progression.

3 METHODOLOGY

3.1 TASK AND DATA

In Alzheimer's Disease clinical trials, the most important cognitive test that is performed to assess current patient function and the likelihood of AD progression is Clinical Dementia Rating Scale Sum of Boxes (CDRSB). Our goal is to predict the progression of Alzheimer's Disease 12 months after the first patient visit. Specifically, our model uses 2D slice stacks of an MRI volume (which are collected at the first visit) as an input and predicts the CDRSB value, which is computed at month 12 (i.e., after one year). All individuals included in our analysis are around $8k$ from seven external datasets, including ADNI (Petersen et al., 2010), BIOFINDER (Mattsson-Carlgren et al., 2020), FACEHBI (Moreno-Grau et al., 2018), AIBL (Ellis et al., 2009), HABs (Dagley et al., 2017), BIOCARD (Moghekar et al., 2013), and WRAP (Langhough Koscik et al., 2021), and four internal studies including A, B, C, and D (dataset names are anonymized in order to allow for blind review). We only include the axial view of MR images for this paper analysis. MR volumes are standardized using standard preprocessing steps (Appendix A.1) and are cropped or padded to (224×224) to make them compatible with the ResNet-50 input size. Furthermore, using the predefined SynthSeg model ((Billot et al., 2021)), we generate segmented images from raw images in order to evaluate the information of different brain parts. The segmentation results only contain the brain region and masked out other details such as skull, etc. The results of segmented images are provided in the Appendix A.2.

We provide two datasets of medical images referred to as center-slice and 5-slice datasets. In the 5-slice dataset, for each subject, we prepare a 5-slice stack from its corresponding 3D MRI volume by taking the center slices of the brain sub-volumes and four adjacent slices at intervals equal to 5 (two to the right, two to the left). The center-slice imaging dataset only contains the center slice, while the 5-slice dataset includes the 5-slice stack. The number of unlabeled images in the center-slice dataset is $43,740$, and the number of unlabeled images in the 5-slice dataset equals $218,700$. Both of these datasets are used to train self-supervised models. During the self-supervised learning process, 92% and 8% of the data are assigned for the training and validation, respectively. Following the common practice in self-supervised learning, based on the minimum self-supervised validation loss, we select the best model and transfer its backbone weights to the supervised model. For the purpose of creating labeled datasets, we use MR images of participants' first visits. The participants are selected from five datasets, including ADNI, A, B, C, and D. For supervised training, we only consider the 2D center slice. The number of labeled data in this dataset is 935 for the train and validation sets, as well as 360 for the independent test set.

3.2 SELF-SUPERVISED LEARNING PLATFORM FOR REGRESSION PROBLEM

Following are the experiments that comprise our approach. 1) We assign our model with self-supervised pre-trained representations on unlabeled natural images in ImageNet. 2) We initialize the weights of model randomly then do self-supervised pre-training on medical images (Figure 1(a)). 3) We initialize weights with supervised pre-trained weights on labeled ImageNet and then self-supervised pre-training on medical images (Figure 1(b)). 4) We initialize the weights with self-supervised pre-trained weights on unlabeled ImageNet and then self-supervised pre-training on all layers on unlabeled medical images (Figure 1(c)). A number of data augmentations have been performed for all the self-supervised learning models discussed in this paper, including random gray scaling, random cropping with resizing, random Gaussian blurring, random rotation, and random flipping. We use a batch size of 50 and epoch of 1000 and implement an early-stop mechanism based on the validation data in order to prevent over-fitting.
After we have pre-trained the weights of self-supervised frameworks, we use their CNN backbone (in our experiments, this is ResNet-50) to fine-tune the regression task (Figure 1(d)) and then calculate the results on an independent set of data (Figure 1(e)). For supervised learning models, we perform data augmentation including random vertical and horizontal rotation, as well as random flipping. For supervised learning, a batch size of 25 and an epoch of 500 are used.

Self-supervised pre-training on natural images

There are a wide variety of self-supervised learning (SSL) models that are pre-trained on large-scale natural image datasets, such as ImageNet, that are commonly used for transfer learning. Such SSL pre-training techniques surpass supervised ImageNet models in several computer vision tasks. To exploit these benefits, in this experiment, we initialize the backbone encoder with weights from SSL models trained on ImageNet. While several types of self-supervised learning methods exist, in this work, we focus on three self-supervised learning methods, one contrastive approach including SimCLR, one redundancy reduction approach including BarLow Twins (BLT), and one clustering-based contrastive learning approach including SwAV which have yielded impressive results on natural image benchmarks. There are several other self-supervised learning strategies, but their performance on ImageNet yields comparable results.

Self-supervised pre-training on medical images

Although representations learned from natural images perform well on medical imaging tasks ((Mustafa et al., 2021)), they may not be optimal for the medical imaging domain due to the large distribution shift from natural images. Generally, medical images are monochromatic and have similar anatomical structures. This discrepancy could be minimized by further pre-training on medical data, either supervised or unsupervised. Due to the time and cost associated with the annotation of medical data, self-supervised learning can be considered a more practical and realistic alternative. Similar to self-supervised pre-training on natural images, we employ SimCLR, Barlow Twins, and SwAV to learn distinctive representations of unlabeled medical images effectively. These methods have all demonstrated good performance in the classification and segmentation of medical images.

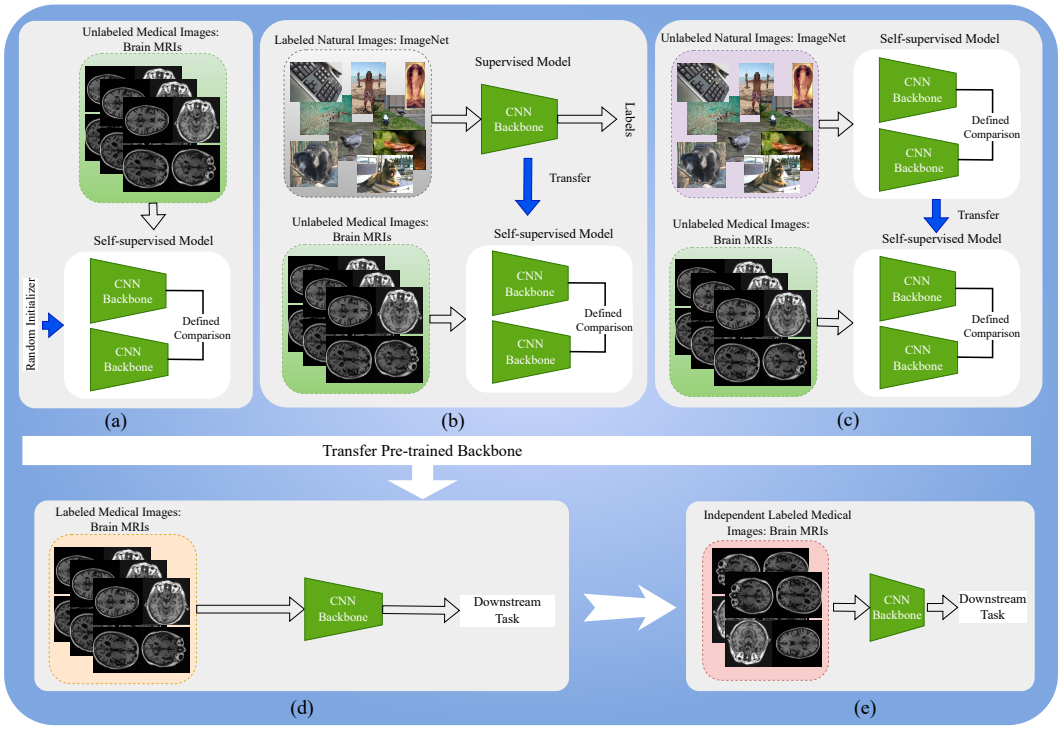

Figure 1: These figures demonstrate different strategies for preparing using in-domain images and how fine-tuning and independent testing are conducted. (a) Show a self-supervised pre-training process with random initialization for in-domain medical imaging. (b) Shows an in-domain pre-training procedure with supervised ImageNet initialization. (c) demonstrates in-domain self-supervised training with self-supervised ImageNet initialization. (d) illustrates in-domain fine-tuning following the transfer of backbone from one of the scenarios (a), (b) or (c). (e) demonstrates the application of the trained model on an independent test set within the domain.

Pre-training	Initialization	R^2
-	Random	0.06
Supervised	ImageNet	0.06
Self-supervised	SimCLR	0.10
	SwAV	0.08
	Barlow Twins	**0.14**

(a) pre-training on natural images

Pre-training	Initialization	R^2
-	Random	0.06
Self-supervised	SimCLR	0.17
	SwAV	0.10
	Barlow Twins	0.13

(b) pre-training on medical images

Table 1: Comparison of different pre-training schemes on downstream task.

Fine Tuning

During fine-tuning, we train the model using the weights of the pre-trained network as initialization for the downstream supervised task. For progress prediction, the standard ResNet-50 backbone ((He et al., 2016)) is followed by a linear layer head, where the backbone is initialized either randomly or with the pre-trained models. Training ResNet-50 involves freezing backbone parameters and only training the head. The mean square error (MSE) is used as a criterion for calculating model loss.

4 RESULTS

We propose the first benchmarking analysis to evaluate the efficacy of different pre-training models for disease progression prediction as a regression problem. (1) We examine the transferability of both supervised and self-supervised ImageNet models pre-trained on natural images for the purpose of assessing their performance on medical tasks; (2) we investigate the efficacy of pre-training models on domain specific medical images for transfer learning performance on medical tasks; and (3) we evaluate cross-domain self-supervised pre-training on both natural and medical datasets to explore the efficacy of the proposed platform for bridging the domain gap between natural and medical images. In order to evaluate the performance of our model, we used the coefficient of determination, which is indicated as R^2.

4.1 SELF-SUPERVISED MODELS ON NATURAL IMAGES PROVIDE MORE GENERALIZABLE REPRESENTATIONS

Experimental setup. This experiment evaluates the transferability of standard supervised ImageNet models with three popular SSL methods using officially released models including SimCLR, Barlow Twins, and SwAV. All SSL models are pre-trained on the ImageNet dataset and employ a ResNet-50 backbone. To establish a baseline, we train the target model using random initialization (i.e., without pre-training).

Results. Table 1a displays the results, from which we draw the following indications: (1) transfer learning from the supervised ImageNet model does not improve over random initialization. We attribute this performance to the significant domain shift between the pre-training and regression target task. In particular, supervised ImageNet models tend to capture domain-specific semantic features, which can be inefficient if the pre-training and target data distributions differ. This finding is consistent with recent studies on other medical tasks that demonstrate that transfer learning from supervised ImageNet pre-training does not always correlate with performance on either classification or segmentation of medical images((Dippel et al., 2021),(Vendrow & Schonfeld, 2022),(Hosseinzadeh Taher et al., 2021)). (2) Transfer learning from self-supervised ImageNet models provides superior performance compared with both random initialization and transfer learning from the supervised ImageNet model. The best self-supervised model (i.e., Barlow Twins) achieves performance improvement of 8% and 8% over random initialization and the supervised ImageNet model, respectively. As opposed to supervised pre-trained models, self-supervised pre-trained models, encode semantic features that are not biased toward any particular task, which can improve generalizbility cross domains.

4.2 SELF-SUPERVISED PRE-TRAINING ON MEDICAL IMAGES OUTPERFORMS IMAGENET MODELS

Experimental setup. We investigate the impact of pre-training medical datasets on self-supervised learning. We train three SSL methods, SimCLR, Barlow Twins, and SwAV on 11 cross-domain unlabeled medical imaging datasets (referred to as in-domain dataset). All of the SSL models are initialized randomly and later are fine tuned by using our labeled dataset.

Results. Table 1b shows the performance accuracy measured by the R^2 when SSL methods are pre-trained using the center slice dataset. We observe that (1) SimCLR pre-training on in-domain dataset achieves the highest performance and provides 7% and 4% performance boost compared to SwAV and Barlow Twins respectively. This observation may offer the superiority of contrastive learning for identifying significant MRI features for predicting progression of Alzheimer's disease in terms of CDRSB (2) SimCLR pre-training on in-domain dataset yields higher performance than either supervised or self-supervised pre-training on only the ImageNet dataset (as seen in Table 1a). Intuitively, pre-training on in-domain dataset encodes domain-specific features that reflect unique characteristics of medical images.

The Barlow Twins pre-training on in-domain data, however, does not improve the performance over the Barlow Twins pre-training on ImageNet. This performance suggests that features learned by Barlow Twins pre-trained on ImageNet are generalizable enough to medical images; And therefore pre-training on a limited number of unlabeled images compared with the ImageNet dataset (40k vs. 1.3 M) may offer limited performance gain for Barlow Twins method, which is a redundancy reductions based model.

4.3 IMPACT OF THE SIZE OF THE UNLABELED IN-DOMAIN DATASET

Experimental setup. We conduct further experiments to evaluate the advantages of increasing the number of unlabeled images for self-supervised pre-training. For each one of SSL methods, we train two separate models by pre-training on either center-slice or 5-slice datasets.

Results. As observed in Table 2, all three SSL models pre-training on 5-slice imaging dataset results in an improvement over pre-training on center-slice dataset, indicating that the SSL models can benefit from larger pre-training in-domain unlabeled datasets. Specifically, Barlow Twins pre-training on the 5-slice dataset provides 1% performance improvement over pre-training on center-slice dataset which may confirm our previous assumption that Barlow Twins, in contrast to SimCLR, needs more unlabeled in-domain images to constrain the ImageNet-based features for medical tasks.

Self-supervised Model	Dataset	R^2
SimCLR	center-slice	0.17
	5-slice	**0.19**
SwAV	center-slice	0.10
	5-slice	0.12
Barlow Twins	center-slice	0.13
	5-slice	0.14

Table 2: A comparison of the performance of self-supervised pre-training models in the fine-tuning step with datasets of different sizes.

4.4 CROSS-DOMAIN SELF-SUPERVISED LEARNING BRIDGES THE DOMAIN GAP BETWEEN NATURAL AND MEDICAL IMAGES

Experimental setup. We investigate the impact of self-supervised pre-training on both natural images and domain-specific medical images. To do so, we pre-train SimCLR on our 5-slice in-domain dataset with two different initialization schemes including (1) Supervised ImageNet (referred to as Labeled_ImageNet→In-domain), (2) Barlow Twins on the ImageNet dataset (referred to as Unla-

beled_ImageNet→In-domain). We select SimCLR and Barlow Twins because as indicated in Tables 1a and 1b, they provide the best results on natural and medical images.

We fine-tune all pre-trained models for the progress prediction task using the labeled 1k images in the training data. Figure 1 shows the cross-domain self-supervised learning employed for the regression problem.

Results. The results are shown in Table 3a. We observe that the best performance is achieved when both unlabeled ImageNet and in-domain datasets are utilized for pre-training. Specifically, Unlabeled_ImageNet→In-domain pre-training surpasses both in-domain and ImageNet pre-trained best performing models. It achieve 15%, 7%, and 2% performance boosts compared to Random initialization, pre-training only on ImageNet, and pre-training only on in-domain dataset, respectively. These results, in line with previous studies ((Hosseinzadeh Taher et al., 2021) ,(Azizi et al., 2021)), imply that pre-training on ImageNet is complementary to pre-training on in-domain datasets, resulting in more powerful representations for medical applications. We also note that Labeled_ImageNet→In-domain pre-training is inferior to Unlabeled_ImageNet→In-domain. These observations restate the efficacy of self-supervised models in delivering more generic representations that can be used for target tasks with limited data, resulting in reduced annotation costs. Appendix A.3 also includes the analysis of these cross-domain experiments but for center-slice dataset.

Pre-training Method	Pre-training Dataset	R^2
Random	-	0.06
Barlow Twins	ImageNet	0.14
SimCLR	In-domain	0.19
Barlow Twins → SimCLR	Unlabeled_ImageNet → In-domain	**0.21**
Supervised ImageNet → SimCLR	Labeled_ImageNet → In-domain	0.18

(a) This table shows best performance model in each domain and their combination in cross-domain self-supervised learning setting

Pre-training Method	Pre-training Dataset	R^2 on Independent test set
Random	-	0.04
Barlow Twins	ImageNet	0.06
SimCLR	In-domain	0.10
Barlow Twins → SimCLR	Unlabeled_ImageNet → In-domain	**0.16**
Supervised ImageNet → SimCLR	Labeled_ImageNet → In-domain	0.12

(b) Results on independent dataset

Table 3: Results of different pre-training schemes on both (a) validation R^2 and (b) independent test R^2

4.5 PERFORMANCE ON TEST SET

Experimental setup. The robustness of our best model, i.e. Barlow Twins→SimCLR on the 5-slice dataset is evaluated and compared with best performing models initialized by either ImageNet or in-domain dataset. Our test set includes 360 unseen patient images.

Results. As presented in Table 3a, the best performance is achieved when both unlabeled ImageNet and in-domain dataset are utilized for pre-training. In particular Unlabeled_ImageNet→In-domain pre-training yields a significant improvement of 10% and 6% over in-domain and ImageNet pre-trained models, respectively. These results confirm our previous observation that supervised pre-trained models on natural images encode semantic features that are generalizable across domains. Moreover, Labeled_ImageNet→In-domain pre-training is inferior to Unlabeled_ImageNet→In-domain indicating that supervised ImageNet models encode domain-specific

semantic features, which may be inefficient when the pre-training and target data distributions are far apart.

5 DISCUSSION

We present a cross-domain self-supervised learning framework for predicting the progression of Alzheimer's disease, which is formulated as a regression problem. This formulation helps to provide more granular estimations of disease progression but can be very noisy and subjective across studies – making it challenging to establish treatment efficacy in small patient populations. The goal of this problem is to predict CDRSB score as an endpoint for defining the stage of Alzheimer's disease. This framework is enforced to use pre-training on natural images followed by pre-training on domain-specific medical images to alleviate domain shift and improve generalization of pre-trained features. Our extensive experiments demonstrates the effectiveness of our approach to combat the lack of large-scale annotated data for training deep models for progression prediction. Our results show an improvement in prediction accuracy and more robust prediction performance for patients with Alzheimer's disease. The proposed framework has potentials to identify patients at higher risk of progressing to AD and help develop better therapies at lower cost to society. As future works, we will expand our dataset to include more than 5-slice from each image, as well as other views such as coronals and/or sagittals, allowing us to introduce more in-domain information to our pre-training model. We also hope to deploy this pre-training strategy on other medical-domain regression tasks.

REFERENCES

Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, J Santamaría, Ye Duan, and Sameer R. Oleiwi. Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10(13):4523, 2020.

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

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 541–549. Springer, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, and Juan Eugenio Iglesias. Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution. *arXiv preprint arXiv:2107.09559*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Alexander Dagley, Molly LaPoint, Willem Huijbers, Trey Hedden, Donald G McLaren, Jasmeer P Chatwal, Kathryn V Papp, Rebecca E Amariglio, Deborah Blacker, Dorene M Rentz, et al. Harvard aging brain study: dataset and accessibility. *Neuroimage*, 144:255–258, 2017.

Jonas Dippel, Steffen Vogler, and Johannes Höhne. Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*, 2021.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

JR Ellis, Pradeep Jonathan Nathan, Victor L Villemagne, RS Mulligan, Timothy Saunder, K Young, Catherine L Smith, J Welch, Michael Woodward, Keith A Wesnes, et al. Galantamine-induced improvements in cognitive function are not related to alterations in $\alpha 4\beta 2$ nicotinic receptors in early alzheimer's disease as measured in vivo by 2-[18f] fluoro-a-85380 pet. *Psychopharmacology*, 202(1):79–91, 2009.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Mara Graziani, Vincent Andrearczyk, and Henning Müller. Visualizing and interpreting feature reuse of pretrained cnns for histopathology. In *Irish Machine Vision and Image Processing Conference (IMVIP 2019), Dublin, Ireland*, 2019.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020a.

Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, and Pengtao Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*, 2020b.

Michal Heker and Hayit Greenspan. Joint liver lesion segmentation and classification via transfer learning. *arXiv preprint arXiv:2004.12352*, 2020.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 3–13. Springer, 2021.

Rebecca Langhough Koscik, Bruce P Hermann, Samantha Allison, Lindsay R Clark, Erin M Jonaitis, Kimberly D Mueller, Tobey J Betthauser, Bradley T Christian, Lianlian Du, Ozioma Okonkwo, et al. Validity evidence for the research category,"cognitively unimpaired–declining," as a risk marker for mild cognitive impairment and alzheimer's disease. *Frontiers in Aging Neuroscience*, 13:688478, 2021.

Hongwei Li, Fei-Fei Xue, Krishna Chaitanya, Shengda Luo, Ivan Ezhov, Benedikt Wiestler, Jianguo Zhang, and Bjoern Menze. Imbalance-aware self-supervised learning for 3d radiomic representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46. Springer, 2021.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10632–10641, 2019.

Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11):3429–3440, 2020a.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020b.

Niklas Mattsson-Carlgren, Antoine Leuzy, Shorena Janelidze, Sebastian Palmqvist, Erik Stomrud, Olof Strandberg, Ruben Smith, and Oskar Hansson. The implications of different approaches to define at (n) in alzheimer disease. *Neurology*, 94(21):e2233–e2244, 2020.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

Afonso Menegola, Michel Fornaciali, Ramon Pires, Flávia Vasques Bittencourt, Sandra Avila, and Eduardo Valle. Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pp. 297–300. IEEE, 2017.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Abhay Moghekar, Shanshan Li, Yi Lu, Ming Li, Mei-Cheng Wang, Marilyn Albert, Richard O'Brien, BIOCARD Research Team, et al. Csf biomarker changes precede symptom onset of mild cognitive impairment. *Neurology*, 81(20):1753–1758, 2013.

Sonia Moreno-Grau, Octavio Rodríguez-Gómez, Angela Sanabria, Alba Pérez-Cordón, Domingo Sánchez-Ruiz, Carla Abdelnour, Sergi Valero, Isabel Hernandez, Maitée Rosende-Roca, Ana Mauleon, et al. Exploring apoe genotype effects on alzheimer's disease risk and amyloid β burden in individuals with subjective cognitive decline: The fundacioace healthy brain initiative (facehbi) study baseline results. *Alzheimer's & Dementia*, 14(5):634–643, 2018.

Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pp. 728–744. PMLR, 2021.

Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 663–671. Springer, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020b.

Edward Vendrow and Ethan Schonfeld. Understanding transfer learning for chest radiograph clinical report generation with modified transformer architectures. *arXiv preprint arXiv:2205.02841*, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Huidong Xie, Hongming Shan, Wenxiang Cong, Xiaohua Zhang, Shaohua Liu, Ruola Ning, and Ge Wang. Dual network architecture for few-view ct-trained on imagenet data and transferred for medical imaging. In *Developments in X-ray Tomography XII*, volume 11113, pp. 184–194. SPIE, 2019.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6210–6219, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 398–407. Springer, 2020.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical image analysis*, 64:101746, 2020.

Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 420–428. Springer, 2019.

A APPENDIX

A.1 PREPROCESSING

MR volumes are standardized using the following preprocessing steps. First, a brain mask is inferred for each volume using SynthSeg (https://github.com/BBillot/SynthSeg), a deep learning segmentation package. During training, the volumes and segmentations are resampled isotropically to 1mm voxel size, standardized to canonical (RAS+) orientation, intensity rescaled to 0,1 and Z-score normalized.

A.1.1 SELF-SUPERVISED MODELS

SimCLR
In this method, the learning rate is $1e-4$, Adam is used as an optimizer and NT-Xent loss is used as a loss function. We use an implementation of SimCLR in Pytorch-lightning repository for creating our framework. `https://github.com/Lightning-AI/lightning-bolts.git`
Barlow Twins
This method utilizes LARS as an optimizer and $1e-4$ as a learning rate scheduler, with CosineWarmup serving as the learning rate scheduler. This repository is used as a basis `https://github.com/SeanNaren/lightning-barlowtwins.git`].
SwAV
For this method, we use Adam as the optimizer, with a learning rate of $1e-4$. Like SimCLR, we use the SwAV base model implementation from the Pytorch-lightning repository `https://github.com/Lightning-AI/lightning-bolts.git`.

A.2 ABLATION STUDIES ON BRAIN SEGMENTATION

In order to find how much information we can use from all parts of brain. We also run some of our early experiments using segmented slices instead of raw slices. Table 4 and Table 5These results shows that although in some experiments using segmentation achieve slightly higher performance, their results in components of cross-domain setting is lower than raw images. Also, results indicate that in 5-slice dataset, results on raw images are higher than segmented images.

Pre-training	Initialization	R^2
-	Random	0.05
Supervised	ImageNet	0.05
Self-supervised	SimCLR	0.08
	SwAV	0.08
	Barlow Twins	0.07

(a) pre-training on natural images

Pre-training	Initialization	R^2
-	Random	0.05
Self-supervised	SimCLR	0.18
	SwAV	0.11
	Barlow Twins	0.15

(b) pre-training on medical images

Table 4: Comparison of different pre-training schemes on downstream task using segmented images.

A.2.1 EXAMPLE OF SEGMENTATION OF ONE BRAIN SLICE

Figure 2 shows an example of brain segmentation which mask out all parts of head including skull and etc. and only keep the brain tissue.

Self-supervised Model	Dataset	R^2
SimCLR	center-slice	0.18
	5-slice	0.16
SwAV	center-slice	0.10
	5-slice	0.09
Barlow Twins	center-slice	0.15
	5-slice	0.15

Table 5: Comparison between self-supervised models performance in different dataset size

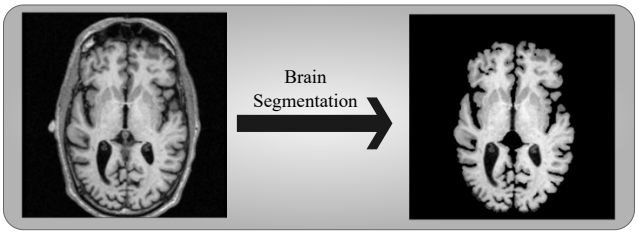

Figure 2: Example of extracting brain tissue from raw 2D MRI slices. In this figure left-side image shows raw(original) slice and right-side image shows segmented image

A.3 Ablation Studies on affect of dataset size on SimCLR pre-training settings

The purpose of conducting this ablation study is to additionally analyse the effect of in-domain database size on 3 experiments of SimCLR. We partially shows Table 6 in section 4.4 with the exception of adding center-slice results as well. We observe when there is a higher number of available in-domain unlabeled images, SSL pre-training on natural images is likely complementary to SSL pre-training on in-domain dataset and it can reduce the domain shift between the pre-training and target tasks. However by using less unlabeled in-domain images i.e. center-slice dataset, Unlabeled_ImageNet→In-domain, yields worse performance, restating that features learned by natural images are not always transferable to down-stream tasks in different domains. Moreover by looking only at center-slice results Labeled_ImageNet is able to boost performance higher compare to Unlabeled_ImageNet. But this observation is vice versa in 5-slice dataset which has more data.

Pre-training Method	Pre-training Dataset		R^2	R^2 on Independent test set
SimCLR	In-domain	center-slice	0.17	0.11
		5-slice	0.19	0.10
Barlow Twins → SimCLR	Unlabeled_ImageNet → In-domain	center-slice	0.16	0.06
		5-slice	**0.21**	**0.16**
Supervised ImageNet → SimCLR	Labeled_ImageNet → In-domain	center-slice	0.18	0.13
		5-slice	0.18	0.12

Table 6: Comparison between R^2 results of different SimCLR pre-training settings using different size of dataset