scientific reports

OPEN

Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification

Pranav Singh[1], Raviteja Chukkapalli[2], Shravan Chaudhari[2], Luoyao Chen[1], Mei Chen[1], Jinqian Pan[1], Craig Smuda[3] & Jacopo Cirrone[1,4✉]

Advancements in clinical treatment are increasingly constrained by the limitations of supervised learning techniques, which depend heavily on large volumes of annotated data. The annotation process is not only costly but also demands substantial time from clinical specialists. Addressing this issue, we introduce the S4MI (Self-Supervision and Semi-Supervision for Medical Imaging) pipeline, a novel approach that leverages advancements in self-supervised and semi-supervised learning. These techniques engage in auxiliary tasks that do not require labeling, thus simplifying the scaling of machine supervision compared to fully-supervised methods. Our study benchmarks these techniques on three distinct medical imaging datasets to evaluate their effectiveness in classification and segmentation tasks. Notably, we observed that self-supervised learning significantly surpassed the performance of supervised methods in the classification of all evaluated datasets. Remarkably, the semi-supervised approach demonstrated superior outcomes in segmentation, outperforming fully-supervised methods while using 50% fewer labels across all datasets. In line with our commitment to contributing to the scientific community, we have made the S4MI code openly accessible, allowing for broader application and further development of these methods. The code can be accessed at https://github.com/pranavsinghps1/S4MI.

Keywords Medical image analysis, Self-supervision, Semi-supervised learning

Medical imaging analysis plays a pivotal role in clinical decision-making, aiding in diagnosis, treatment planning, and monitoring. The advent of deep learning has significantly enhanced the capability to analyze medical images both effectively and efficiently, promising to automate aspects of the diagnostic process and thereby augment clinical decision-making. However, the efficacy of these deep learning techniques is heavily contingent upon the availability of well-annotated medical image data. Unlike in natural image processing, obtaining annotations in the medical domain is fraught with challenges due to privacy concerns, high costs, and the extensive time required for expert clinicians to produce accurate annotations[1]. Moreover, the unique requirements for expert knowledge in medical imaging mean that traditional crowd-sourcing for annotations is not viable. This results in medical imaging datasets being significantly smaller than their natural imaging counterparts, leading to suboptimal performance when neural networks are trained from scratch on such limited data volumes. Consequently, transfer learning has become a key strategy, leveraging knowledge acquired from source domain tasks to enhance performance on target domain tasks in medical imaging. Transfer learning typically involves initializing the network with weights from a pre-trained model and subsequently fine-tuning it with target domain data. While transfer learning has shown promise, especially when source and target datasets, as well as their respective output classes, are similar[2], the distinct nature of medical imaging often limits its applicability.

[1]Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA. [2]Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA. [3]Division of Rheumatology, NYU Grossman School of Medicine, 550 1st Ave, New York, NY 10016, USA. [4]Colton Center for Autoimmunity, NYU Grossman School of Medicine, Science Building, 435 E 30th St, New York, NY 10016, USA. ✉email: cirrone@courant.nyu.edu

The core motivation of our study is to address these challenges by comparing machine supervision with traditional fully-supervised approaches in the realm of medical imaging. The acquisition of labels in medical imaging is notably costly and time-intensive, a hurdle that machine supervision approaches, with their scalable and label-free nature, aim to overcome. By enabling the development of priors that can outperform supervised methods, machine supervision tackles a critical bottleneck in advancing clinical treatments-the heavy reliance on supervised learning techniques that necessitate extensive annotated data. This study aims to elucidate the efficiencies that machine supervision can introduce, particularly in light of the challenges posed by the need for domain-specific annotation, the limited size of medical datasets due to stringent privacy regulations and high annotation costs, and the indispensable requirement for expert knowledge in annotating medical images.

In recent times, the field of computer vision has witnessed remarkable strides due to achievements in learning algorithms such as Deep Metric Learning[3], self-distillation[4], masked image modeling[5], etc. These innovative approaches have enabled the learning of meaningful visual representations from unannotated image data, with fine-tuning models from these learners showing competitive performance gains[4]. There is a growing interest in exploring and applying these learning algorithms to medical imaging modalities, where annotations are notably scarce. These algorithms, typically deployed in self-supervised, semi-supervised, or unsupervised settings, involve a two-stage process: (1) pre-training to acquire representations from unannotated data, and (2) fine-tuning using annotated data to refine the model for the specific target task.

In the realm of self-supervised learning, our work specifically focuses on DINO (Distillation with NO labels)[4] and CASS (Cross-Architectural Self-supervision)[6] techniques, which are at the forefront of current research. These methods leverage joint embedding[7] based architecture, a cutting-edge approach that facilitates learning meaningful visual representations by aligning two different views of an image to the same embedding space. This strategy is particularly adept at extracting robust features from unannotated data, making it highly relevant to medical imaging analysis. Additionally, for semi-supervised learning, we delve into the cross-architectural, cross-teaching method[8], which represents a significant advancement towards utilizing machine supervision over traditional human annotation. The adoption of such state-of-the-art self/semi-supervised techniques marks a pivotal shift in medical imaging, emphasizing machine supervision's scalable and label-free advantages. This is crucial in the context of medical imaging, where the procurement of labels is not only cost-prohibitive but also requires extensive time and expertise.

Building upon these considerations, this work introduces the S4MI (Self-Supervision and Semi-Supervision for Medical Imaging) pipeline. This novel semi-/self-supervised learning approach is designed to directly address the aforementioned challenges by minimizing the dependency on extensive labeling efforts. By integrating cutting-edge self-supervised learning algorithms with the latest semi-supervised learning techniques, S4MI aims to significantly improve performance beyond the capabilities of traditional supervised approaches. This initiative represents a significant step towards the development of scalable healthcare solutions, leveraging the untapped potential of machine supervision to achieve and potentially exceed the efficacy of human-supervised methods in medical imaging analysis.

In this work, we evaluate the performance of state-of-the-art machine supervision approaches against traditional transfer learning across two pivotal tasks in medical image analysis: segmentation and classification. Our comparison spans three challenging datasets from distinct medical modalities, including histopathology slide images and skin lesion images. Additionally, we explore the effectiveness of machine supervision by varying the amount of annotated data used for fine-tuning. The machine supervision approaches examined in this study include[4,6] for classification and[8,9] for segmentation.

Data

To compare transfer learning from ImageNet with machine supervision and show its general applicability, we selected three datasets representative of those clinicians encounter in real life, featuring varying levels of class imbalance and sample sizes.

- **Dermatomyositis**: This dataset comprises 198 RGB samples from 7 patients, each image measuring 352 by 469 pixels. With this dataset, we conduct multilabel classification, aiming to classify cells based on their protein staining into TFH-1, TFH-217, TFH-Like, B cells, and cells that do not conform to the aforementioned categories, labeled as 'others.' We utilize the F1 score as our evaluation metric on the test set.
- **Dermofit**: This dataset consists of 1,300 image samples captured with an SLR camera, spanning the following ten classes: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus/Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Intraepithelial Carcinoma (IEC), Pyogenic Granuloma (PYO), Haemangioma (VASC), Dermatofibroma (DF), and Melanoma (MEL). Each image in this dataset is uniquely sized, with dimensions ranging from 205×205 to 1020×1020 pixels.
- **ISIC-2017**: Within this dataset, 2,000 JPEG images are distributed among three classes: Melanoma, Seborrheic Keratosis, and Benign Nevi. The evaluation of the test set is based on the Recall score.

Methods
Classification

For classification, we compare existing self-supervised techniques with transfer learning, the de facto norm in medical imaging. As mentioned, due to a lack of data in medical imaging, classifiers are often trained using initialization from other large visual datasets like ImageNet[10]. The architectures are then fine-tuned on the target medical imaging dataset using these initializations. Alternatively, we can train using self-supervised approaches to provide better classification priors for the target medical imaging dataset. In self-supervised learning, an auxiliary task is performed without labels to learn fine-grained information about the image. This auxiliary task

can be performed in various ways, for example, by corrupting the image, followed by its reconstruction, creating copies of the same image (positive pairs), and minimizing the distance between them using two differently parameterized architectures or redundancy reduction. In our study, we focus on DINO[4] and CASS[6]. DINO learns by creating augmentation-asymmetric copies of the input image, whereas CASS creates architecturally asymmetric copies followed by similarity maximization between the copies. We pre-trained using these two self-supervised techniques for 100 epochs, followed by 50 epochs of fine-tuning with labels. We perform fine-tuning with 10% and 100% label fractions. For this, we use all of the available labels per image while using 10% or 100% of the total labels available.

Classification pipeline additional details
For the ImageNet supervised classification approach, we used images and their corresponding labels as our inputs. For training with x% labels, we only used x% of the entire training images and labels while keeping the number of labels per image the same. For the self-supervised approaches DINO and CASS, for pertaining, we start with only images. During the fine-tuning process, we initialize the networks with their pre-trained weights and use corresponding image-label pairs. Similar to the supervised approach, we also fine-tune these architectures for two label fractions, 10% and 100%. Further details can be inferred from our open-sourced code base.

Segmentation
This section will first introduce standard implementation designs mutually applied to all models, followed by additional implementation details for the semi-supervised model. Lastly, we will describe the procedure of testing different cross-entropy weight initialization.

Approach specific additional details
Semi-supervised approach The semi-supervised model uses a batch size of 16, and each batch consists of eight labeled images and eight unlabeled images. In the last column of Fig. 5, we compare the performances between semi- and fully-supervised models in 100% labeled-ratio scenario; for this, we re-adjusted the batch to be 15 labeled images and one unlabeled image to approximate the fully-supervised setting while still retaining the "cross teaching" component so that the loss remains consistent and hence comparable with other semi-supervised label ratios. Additionally, when comparing the performance between fully and semi-supervised models, we adopt the same practice from Luo et al.[8] and use Swin Transformer to compare DEDL[11] with Resnet34 backbone as they have a similar number of parameters

Unsupervised approach We implemented PiCIE for Dermofit, Dermatomyositis, and ISIC-2017 datasets similar to the supervised and self/semi-supervised methods. We made slight changes that suit the datasets and ensure smooth learning, such as using SGD optimizer instead of Adam optimizer, keeping the learning rate the same as the original implementation (1e−4), and adding the StepLR (Step Learning Rate) scheduler provided by PyTorch. This is because we observed that the model sometimes failed to learn and was stuck at predicting all the image pixels as either foreground (thing) or background (stuff) with the Adam optimizer. We used batch sizes of 64 and 128, depending on the availability of CUDA RAM and dataset sizes. We trained the PiCIE unsupervised pipeline for 50 epochs with the ResNet34 backbone as a feature extractor and retained the unsupervised clustering technique described in[9] to achieve optimal segmentation performance. The hardware used was Nvidia RTX8000, similar to the previous methods. For PiCIE, labels were only used for validation and testing, not during training.

Supervised approach We adopted Singh and Cirrone's[12] approach for the supervised part of our comparison. For this, we utilize image and mask pairs during training, validation, and testing with a ResNet-34-based U-Net[13].

Common implementations
We implemented all models in Pytorch[14] using a single NVIDIA RTX-8000 GPU with 64 GB RAM and 3 CPU cores. All models are trained with an Adam optimizer with an initial learning rate (lr) of 3.6e−4 and a weight decay of 1e−5. We also set a cosine annealing scheduler with a maximum of 50 iterations and a minimum learning rate of 3.4e−4 to adjust the learning rate based on each epoch. For splitting our dataset into training, validation, and testing sets, we use a random train-validation-test split (70–10–20%), except in the ISIC2017 dataset, where we adopt the train/val/test split according to[15] for match-up comparison (57–8.5–34%). The batch size is 16, and we use data augmentation to enrich the training set using random rotation, random flip, and a further resizing to 224×224 to fit in Swin Transformer's patch size requirement[16]. Note that for 3-channel datasets (Dermofit, ISIC), we add a pre-processing step that normalizes the red channel of the RGB color model as proposed by[17]. We repeat all experiments with different seed values five times and report the mean value in the 95% confidence interval in all tables. Similar to classification, we fine-tune with multiple label fractions, with semi-supervised and DEDL. When we mention that we fine-tune with x% labels, we use all labels per image but only x% of the total available image-label pairs.

Data-preprocessing
Following[11], we chose the processed image sample size to be 480×480 to avoid empty tile issues. To ensure uniformity, the image input size for the model remains consistent across all three datasets, as illustrated below. For the Dermatomyositis dataset, since it contains images of uniform sizes 1408×1876 each, we tiled them to a size of 480×480 and then used blank padding at the edges to make them fit in 480×480 sizes. This results in 12 tiled sub-images per sample, which are then resized to 224×224. In contrast, the Dermofit and the ISIC2017 datasets contain images of different sizes. Since the two datasets are about skin lesions, they have significantly denser and larger mask labels than the Dermatomyositis dataset. Thus, a different image preprocessing step is applied to the latter two datasets: bilinear interpolation to 480×480 followed by a resize to 224×224.

Results

We start by studying the classification pipeline in Sect. "Classification", followed by understanding the interpretability aspect of classification in Sect. "Classification". Since interpretability is ingrained in the segmentation task, we only study segmentation quantitatively in Sect. "Segmentation".

Classification

For classification, we advocate using self-supervision to learn useful discriminative representations from unlabelled data. Self-supervised techniques rely on an auxiliary pretext task for pre-training to learn useful representations. These representations are further improved and aligned with the downstream task through labeled fine-tuning. We benchmark and study the performance of transfer learning as well as two self-supervised techniques: (1) Distillation with No Labels (DINO)[4], a self-supervised approach designed for natural images, and (2) Cross cross-architectural self-supervision (CASS)[6] a self-supervision approach for medical imaging. We further expand on our methodology in Sect. "Classification".

We present the results of supervised and self-supervised trained classifiers in Table 1. We use the F1 score as the comparison metric, similar to its implementation in previous works[6]. The F1 score is defined as $F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$, where TP represents True Positive, FN is False Negative and FP is False Positive. We observe that CASS (Fig. 1) outperforms DINO and transfer learning consistently across all the datasets and backbones, with the, exception for the ISIC 2017 challenge dataset, where DINO outperforms CASS using the ResNet50 backbone. Interestingly, we also observe that ViT trained with CASS using just 10% labels, which saves significant annotation time, performs significantly better than the transfer learning-based supervised approach with 100% labels for the Dermatomyositis dataset. For the same case, DINO performs on par with the supervised approach while using 90% fewer labels. This shows the impact that machine supervision can have on improving access to classified medical images. DINO and CASS take an image to create asymmetry through augmentation (DINO) and architecture (CASS). These are then passed through different parameterized feature extractors (DINO) and architectural feature extractors (CASS) to create embeddings. Since the two embeddings are generated from the same image, they are expected to be similar; this is then used as the supervisory signal to maximize the similarity between the two embedding and update encoders. In CASS, architectural invariance is
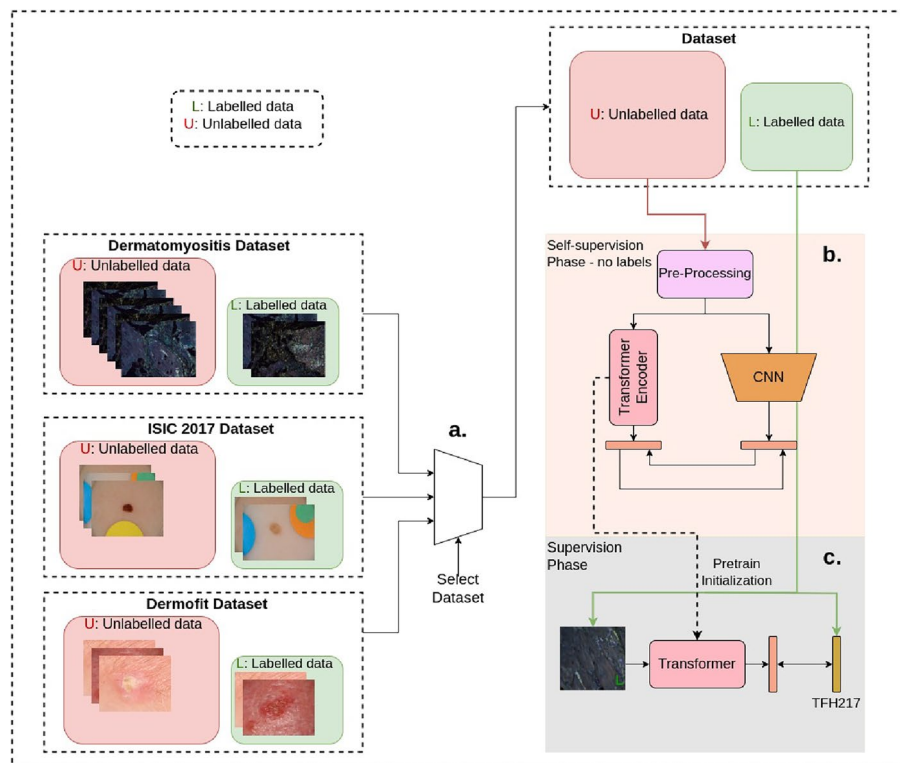

Figure 1. Applying CASS: In this figure we detail the steps involved in applying CASS, the best-performing machine supervision approach for classification[6]. We conducted experiments on three datasets listed on the left side: the Dermatomyositis, ISIC-2017, and Dermofit datasets. For training on a dataset, we initialize both the networks with their ImageNet weights and select one dataset at a time. To train CASS, we start by label-free pretraining as illustrated in Part (b) of Fig. 1. During pre-training, a CNN and a Transformer are trained simultaneously. In the case of the Dermatomyositis dataset, the finetuning is multi-label, while in the case of the ISIC-2017 and the Dermofit dataset, it is multi-class. This pre-training in (b) is followed by labeled fine-tuning as shown in Part (**c**), where image are fine-tuned one at a time.

Backbone	Technique	Autoimmune Dataset F1 Score		Dermofit Dataset F1 Score		ISIC 2017 Dataset Recall Score	
		10%	100%	10%	100%	10%	100%
ResNet-50	DINO	**0.8237 ± 0.001**	0.84252 ± 0.008	0.3749 ± 0.0011	0.6775 ± 0.0005	0.322 ± 0.009	**0.6407 ± 0.0363**
	CASS	0.8158 ± 0.0055	**0.8650 ± 0.0001**	**0.4367 ± 0.0002**	**0.7132 ± 0.0003**	**0.343 ± 0.002**	0.599 ± 0.0304
	ImageNet Transfer	0.819 ± 0.0216	0.83895 ± 0.007	0.33 ± 0.0001	0.6341 ± 0.0077	0.288 ± 0.091	0.5774 ± 0.0004
ViT B16	DINO	0.8445 ± 0.0008	0.8639 ± 0.002	0.332 ± 0.0002	**0.4810 ± 0.0012**	0.291 ± 0.005	0.5571 ± 0.076
	CASS	**0.8717 ± 0.005**	**0.8894 ± 0.005**	**0.3896 ± 0.0013**	0.4667 ± 0.0002	**0.321 ± 0.007**	**0.588 ± 0.0094**
	ImageNet Transfer	0.8356 ± 0.007	0.8420 ± 0.009	0.299 ± 0.002	0.456 ± 0.0077	0.277 ± 0.007	0.5322 ± 0.022

Table 1. Results for DINO, CASS, and Supervised (ImageNet-initialized transfer learning) methods on the Dermatomyositis, the Dermofit, and ISIC 2017 datasets. In this table, we compare the F1 score on the test sets. Bold values indicate the highest F1 score achieved by any method (DINO, CASS, or ImageNet transfer learning) for each combination of backbone, dataset, and label percentage. In all cases, self-supervised pre-training (CASS or DINO) improves performance over the ImageNet Transfer learning benchmark. Additionally, among CASS and DINO, we observe that CASS, with its focus on architectural invariance, outperforms DINO for almost all label fractions across three datasets for ResNet as well as for ViT.

used to create a positive pair instead of creating augmentation invariance, as in the case of DINO. This is because CNN and Transformers learn different representations from the same image[6].

Saliency maps
In segmentation, interpretability is ingrained in the task itself, as we can easily identify whether model predictions are aligned with the ground truth. To understand the decision-making process of trained neural classifiers, we study pixel attribution or saliency maps[18]. To accomplish this, we first train the neural network and compute the gradient of a class score with respect to the input pixels. Backpropagating the gradients that lead to a particular classification onto the input image helps us understand the most filtered/rewarding extracted features related to classification. We present the results using DINO, CASS-trained ResNet-50, and ViT Base/16 on the ISIC-2017 dataset in Fig. 2. Although most saliency maps fail to align with human-interested pathology, CASS-based saliency maps align more than those from DINO. This could be attributed to the joint training of CASS, where it
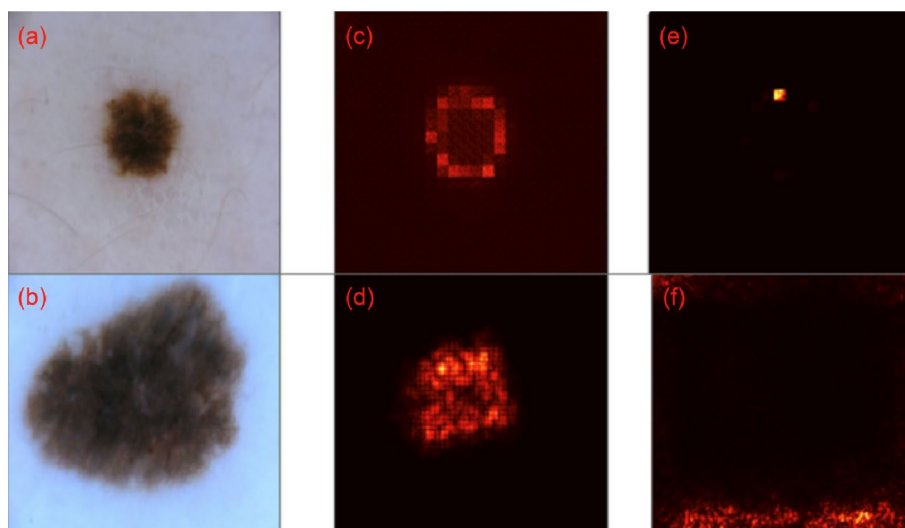

Figure 2. Visualization of saliency maps on a random sample from the ISIC-2017 dataset, left (**a**, **b**): data (input image), middle (**c**, **d**): saliency map from CASS, and right (**e**, **f**): saliency map from DINO-trained ViTB/16 at the top and ResNet-50 at the bottom (**f**). DINO's saliency map exhibits notable stochasticity, displaying a lack of strong correlation with the specific pathology under consideration. Conversely, in the case of CASS, the saliency maps demonstrate a significantly more aligned with the pathology of interest both for CNN as well as the Transformer.

trains a CNN that focuses on local information while a Transformer focuses on global or image-level features. Although both techniques fail to recognize and focus on the relevant area in all cases, CASS-trained architectures are consistently better focused on the relevant pathology than DINO-trained architectures.

Segmentation

Image Segmentation is another important task in medical image analysis. The area of interest—usually containing the pathology—is segmented from the rest of the slide image. We study the performance of transfer learning, semi-supervised, and unsupervised approaches for segmentation. In computer vision, it is widely acknowledged that image segmentation presents a more intricate challenge than image classification. This distinction arises from the fact that image segmentation necessitates the meticulous classification of individual pixels, whereas image classification solely requires making predictions at the image level. Moreover, due to the aforementioned rationale, annotating images for segmentation exhibits a significantly higher complexity level. Therefore, the present study aims to investigate the effectiveness of unsupervised, semi-supervised, and fully supervised techniques across four distinct label fractions instead of the conventional two in classification. In our case, x% label fractions indicate that we only use x% labels from the dataset. We include the unsupervised approach[9] to study the performance for 0% labels scenario. We compare the performance of DEDL (Data-Efficient Deep Learning framework)[11], a transfer learning-based approach, against[8], a semi-supervised approach using the Swin transformer-based U-Net model (Fig. 3).

We choose[11] due to its significant promise in histopathology imaging, and[8] for its state-of-the-art approach to leveraging unannotated data in MR segmentation. The semi-supervised approach[8] trains two segmentation architectures—a Swin transformer-based U-Net and a CNN U-Net (as shown in Fig. 3), while DEDL[11] and unsupervised approach[9] only train one architecture.

To ensure a fair comparison, we make sure the number of overall parameters trained is the same in each case over three datasets and four-label fractions.
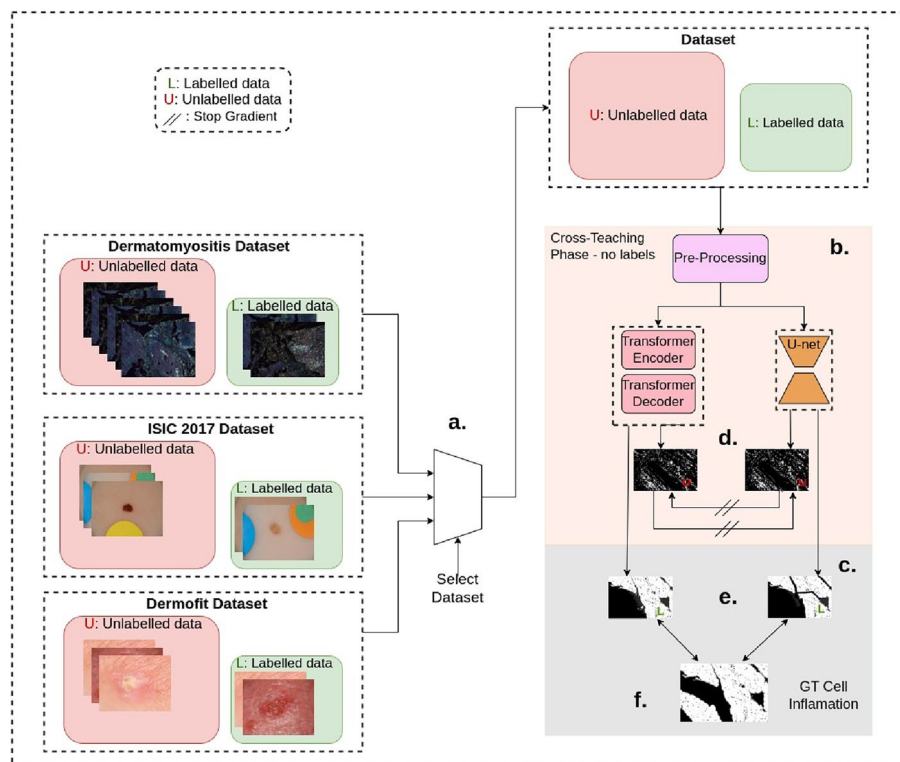

Figure 3. In this figure we present the semi-supervised pipeline as described in Sect. "Segmentation". Similar to the classification experiments, we evaluate the segmentation pipeline on three challenging medical image segmentation datasets—the Dermatomyositis, ISIC-2017, and the Dermofit dataset. We use one dataset at a time to train the semi-supervised architecture. Unlike the classification pipeline, semi-supervised learning involves simultaneous learning from labeled and unlabeled data. In Part (b) of Fig. 3, we start by training the data in an unlabeled fashion and during the same iteration labeled data is also passed to the architecture as shown in part (c) of the figure. Predictions from passing inputs of the labeled images yield learned predictions as shown in Part (e). Unsupervised loss is then calculated by comparing the outputs of the CNN and the Transformer (as shown in Part (d) of this figure) using the $\mathcal{L}_{\text{Unsupervised}}$ in Sect. "Segmentation". This unsupervised loss is then added to the supervised loss denoted by $\mathcal{L}_{\text{Supervised}}$ in Sect. "Segmentation". The supervised loss is calculated against the ground truth as shown in part (e) and (f) in this figure.

In their study[8], compared the performance of Swin-U-Net trained in semi- and fully-supervised fashion for segmentation. Observing that Swin-U-Net exhibits superior performance in a semi-supervised setting, we extend this comparison to include a semi-supervised Swin-U-Net and a similarly parameterized ResNet-34 based U-Net, trained under both unsupervised and fully-supervised approaches. Furthermore, from[11], we use their best-reported ResNet-34-based U-Net model.

The evaluations are based on three datasets, as detailed in Sect. "Data". To ensure our results are statistically significant, we conduct all experiments with five different seed values and report the mean values in the 95% confidence interval (C.I.) over the five runs. For comparison, we use IoU (Intersection over Union) in line with previous works in the field[12]. IoU or Jaccard index for two images U and V is defined as $IoU(U, V) = \frac{|U \cap V|}{|U \cup V|}$. We present these results in Fig. 5.

Semi-supervised approach
Semi-supervision approaches utilize the dataset's labeled and unlabeled parts for learning. Techniques generally fall under (1) Adversarial training (e.g. DAN[19]) (2) Self-Training (e.g. MR image segmentation[20]) (3) Co-Training (e.g. DCT[21]) and (4) Knowledge Distillation (e.g. UAMT[22]). We study the co-training-based semi-supervision segmentation technique[8], depicted in Fig. 4. The model consists of two segmentation architectures, a U-Net (trained from scratch) and a pre-trained Swin Transformer, adapted from the Swin Transformer proposed by Liu et al.[16]. Both architectures are updated simultaneously using a combined loss over the labeled and unlabeled images.

The U-Net model is trained as follows:

(1) The **labeled data** (L) is evaluated using the average of dice and cross-entropy loss

$$\mathcal{L}_{\text{Supervised}} = \frac{1}{2}[\text{DICE}(\text{pred}_{\text{U-Net}}, L) + \text{CrossEntropy}(\text{pred}_{\text{U-Net}}, L)]$$

(2) For **unlabeled data** (U) cross-teaching strategy is used to cross-supervised between CNN and Transformer: to update U-Net, output from Transformer, $\text{pred}_{\text{Transformer}}$ (vice versa) is used as "sudo ground truth":

$$U_{\text{sudo}} = \text{argmax}(\text{pred}_{\text{Transformer}})$$

(3) The unsupervised loss for U-Net is conducted between prediction and pseudo-ground truth:

$$\mathcal{L}_{\text{Unsupervised}} = \text{DICE}(\text{pred}_{\text{U-Net}}, U_{\text{sudo}})$$

(4) The final loss is the summation of unsupervised and supervised loss.

Unsupervised approach
Unsupervised approaches for learning from unlabelled images are mostly limited to curated and object-centric images; however, some recently proposed methods like[9] achieve semantic segmentation with the help of unsupervised clustering methods. The machine aims to discover and identify the different foreground objects and separate them from the background. In this paper, the datasets under study only contain one foreground object, which eases learning; however, unsupervised segmentation continues to be a challenging task. Cho et al.[9] proposes a technique PiCIE (Pixel-level feature Clustering using Invariance and Equivariance) to incorporate geometric consistency to learn invariance to photometric transitions and equivariance to geometric transitions. The two main constraints here are (1) to cluster the pixels having similar appearance together with the same label
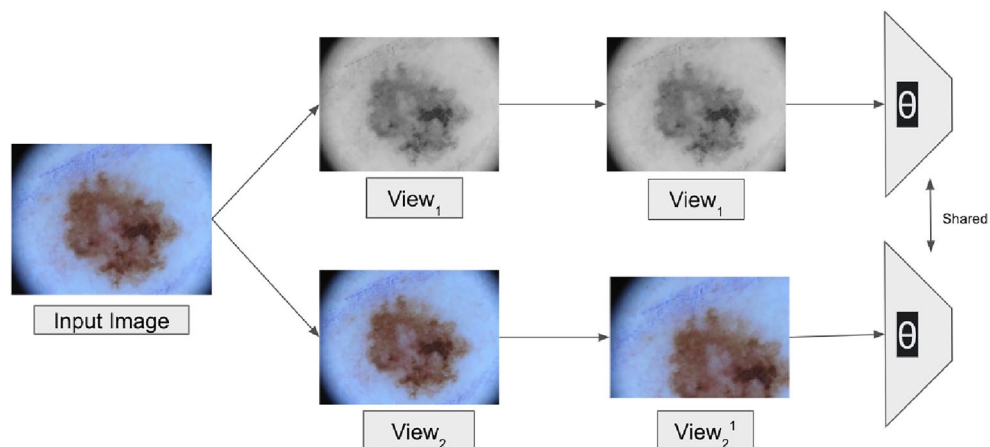

Figure 4. In this figure, we depict the Unsupervised Segmentation Approach: PiCIE pipeline[9]. *View*$_1$ and *View*$_2$ represent two photometrically transformed views of the input image, whereas *View*$_2^1$ represents the geometric transformation of *View*$_2$. Cross-view training is then used to train the architecture shared between the two views (parameterized by θ in the figure); we have expanded further on this in Sect. "Segmentation".
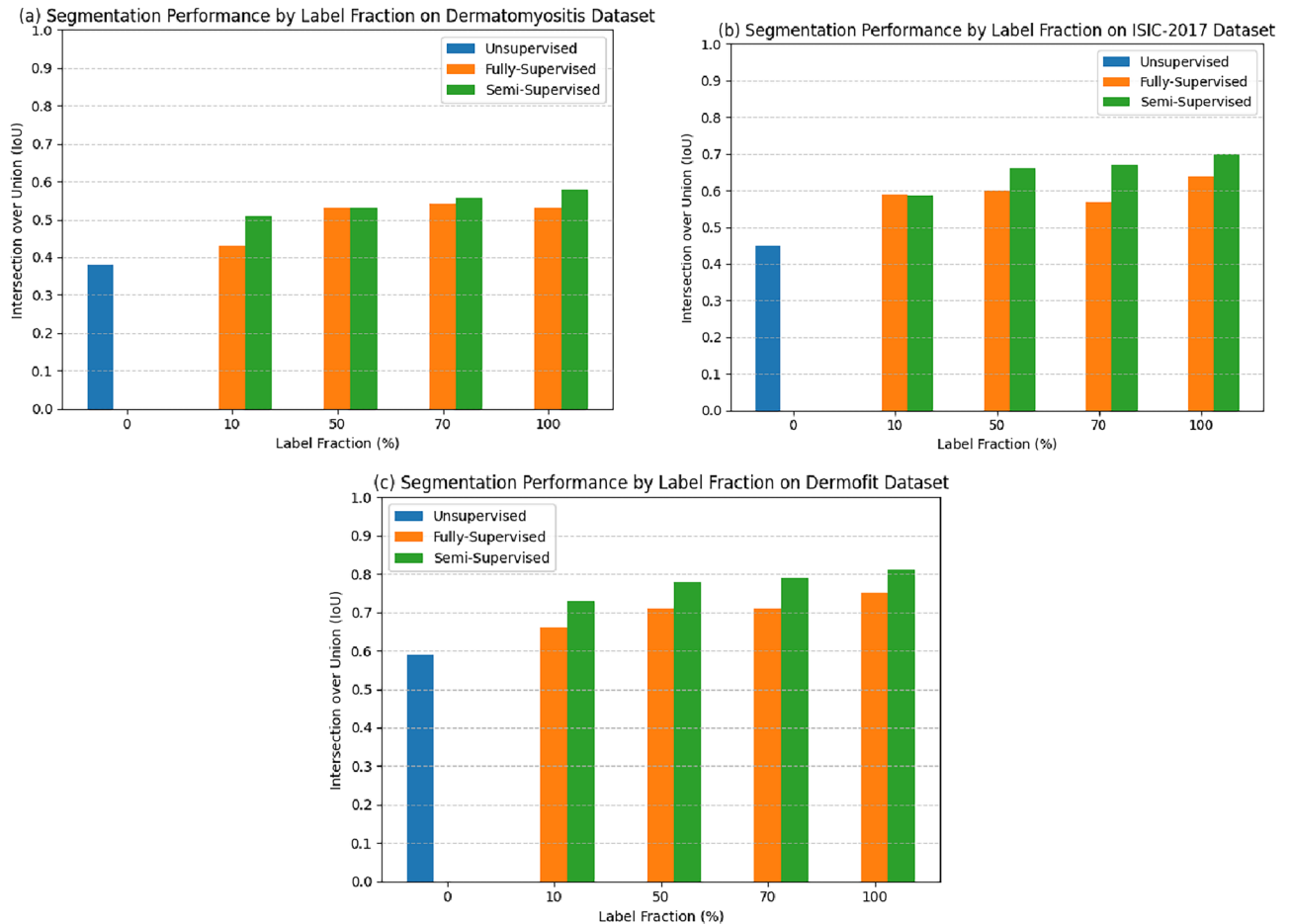
Figure 5. In this figure, we present the results for the Dermatomyositis dataset, ISIC-2017, and the Dermofit dataset in panels (**a**), (**b**), and (**c**), respectively. We compare the segmentation performance of Full, Semi, and Unsupervised architectures across these datasets, considering different percentages of label fraction (x-axis). Performance is evaluated using the Intersection Over Union (IoU), depicted on the y-axis, to compare results across all three datasets. IoU values range from 0 to 1. The blue bar represents the performance of the unsupervised approach, PiCIE[9], which, by definition, does not require any labels for fine-tuning. Consequently, we present results for PiCIE using 0% label fractions. Remarkably, we observe that the semi-supervised approach surpasses the fully-supervised approach by requiring 50% fewer labels per image across all three datasets.

for all the pixels belonging to a cluster and (2) to ensure the pixel label predictions follow the invariance and equivariance constraints mentioned earlier. With these objectives, Cho et al.[9] trains a ConvNet-based semantic segmentation model in a purely unsupervised manner (without using any pixel labels). PiCIE uses the alternating strategy between clustering the feature representations and using the cluster labels as pseudo labels to train the feature representation proposed by DeepCluster[23].

Figure 4b shows a brief overview and an end-to-end pipeline of the PiCIE approach. For each input image, two photometric transformations, like jittering the pixel colors, etc., are randomly sampled to create $View_1$ and $View_2^1$. One of these views ($View_1$) is directly fed to the network, while geometric transformation is applied to the other view before feeding it to the same network ($View_2^1$). The network then produces the two feature representations for every pixel corresponding to the two views. Since we expect equivariance for geometric transformations, they are applied to feature representations of the original image before clustering the pixels of the feature representations. The inductive biases of invariance and equivariance to photometric and geometric transformations are introduced by aggregating the clustering losses within the same view and across the two views. Using this cross-view training strategy, the training is guided to prevent predicted pixel labels from changing with jittering of pixel colors. Still, at the same time, if the image is warped geometrically, then the warping would also be reflected in the labeling similarly[9].

Comparision

Table 5 shows the evaluation results of these methods for the three datasets. The evaluation metric applied is intersection over union (IOU). The unsupervised approach[9] sets the lower bound on the segmentation performance and illustrates the difficulty of unsupervised segmentation and its ability to outperform even the state-of-the-art unsupervised approaches using a small number of annotations (10%). Semi-supervision continuously

outperformed the transfer learning across all label fractions across all datasets except the 10% case for ISIC, where the performance is close. For all the datasets, semi-supervision with 50% labels outperforms transfer learning from Imagenet with 100% labels. These results show a significant performance gain by moving to machine supervision approaches instead of transfer learning from Imagenet. Figures 1 and 3 show the best-performing approaches for learning from limited annotation datasets for classifications and segmentation, respectively.

Discussion

We observed a significant improvement in the performance of trained classifiers with the incorporation of pre-training self-supervised objectives, surpassing methods reliant on fully annotated data. When comparing the two training regimes, CASS consistently outperformed DINO in almost all cases under study. Further analysis into the pixel attribution of the trained methods, to understand their decision-making processes, revealed that CASS typically focuses on relevant pixels more accurately than DINO. Therefore, employing self-supervision-additional training with reduced human supervision-enhanced performance.

We explored three training regimes for segmentation: supervised, semi-supervised, and unsupervised. Although implementing the unsupervised approach necessitates some adaptation and hyperparameter tuning, the Dermofit dataset consistently demonstrates that the semi-supervised approach outperforms the fully-supervised approach, irrespective of the labeling percentage. Notably, we achieved consistent performance improvements using a semi-supervised approach-wherein both labeled and unlabeled data per batch were utilized to enable the model to learn better representations. On the Dermatomyositis dataset, the performance using no labels is closely comparable to that of the supervised and semi-supervised approaches with 10% labels. Performance between semi-supervised and fully-supervised approaches converges when labeling ranges between 50% and 70%. However, on the ISIC-2017 dataset with 10% labels, the fully-supervised approach initially shows superior performance compared to the semi-supervised approach; yet, as the percentage of labels increases beyond 10%, semi-supervised begins to outperform fully-supervised.

Additionally, unlike the unsupervised approach, the supervised and the semi-supervised approaches don't treat this problem as pixel clustering and are less prone to overfitting to the dominant pixel distribution. Since classification requires image-level identification of classes instead of pixel-level identification, segmentation is a more difficult objective. Yet, introducing reduced human supervision with pseudo-labels in the semi-supervised approach improved performance beyond both the supervised and unsupervised paradigms, to the point where the semi-supervised approach for segmentation outperformed fully-supervised methods while requiring 50% fewer labels across all evaluated datasets.

Conclusion

The primary focus of our experimental investigations revolved around two fundamental medical imaging tasks: classification and segmentation. Significantly, our findings from employing the S4MI (Self-Supervision and Semi-Supervision for Medical Imaging) framework indicate a promising shift away from traditional, purely transfer-learning-based supervised methodologies. Specifically, in the realm of classification, our empirical evidence consistently shows that self-supervised training within the S4MI framework outperforms conventional supervised approaches across both Convolutional Neural Network (CNN) and Transformer models.

Moreover, in the comparative analysis of the two self-supervised techniques, CASS demonstrates superiority by exhibiting enhanced performance in nearly all scenarios compared to DINO. Utilizing reduced supervision in segmentation has yielded favorable outcomes. Specifically, applying the semi-supervised approach, as outlined in[8], has resulted in notable enhancements in performance across all label fractions for the three datasets under consideration.

A notable inference emerged from our investigation, wherein the complete elimination of supervisory signals through the utilization of unsupervised algorithms resulted in comparable performance solely in the context of the dermatomyositis dataset when compared to architectures with 10% supervision, an observation that underscores the inherent capacity of unsupervised methodologies in distinct contextual situations.

Based on a comprehensive analysis of our experimental findings, adopting machine-level supervision through the S4MI framework reduces dependence on human supervision and yields substantial advantages in both time efficiency and accuracy of medical image analysis. The findings of our study make significant empirical contributions to the fields of medical imaging and limited-supervision techniques, thereby stimulating future research in this area. The distribution of our S4MI pipeline as open-sourced code will aid other researchers by saving labeling time and improving image analysis quality. This, in turn, will facilitate the advancement of healthcare solutions, leading to improved patient care outcomes. We are optimistic that our study will act as a catalyst for meaningful dialogue and collaborative endeavors within the medical imaging community, propelling progress in this critical domain.

Data availability

We utilized three datasets for our experiments: Dermatomyositis, Dermofit, and ISIC-2017. The Dermatomyositis dataset, while private and available from NYU Langone, is subject to restrictions and was used under license for this study; therefore, it is not publicly available. Data are however available upon reasonable request and with permission of NYU Langone to Jacopo Cirrone at cirrone@courant.nyu.edu. To facilitate open-source reproduction, we conducted our experiments on the Dermofit and ISIC-2017 datasets, both of which are publicly available at ISIC Data repository and DERMOFIT Project Datasets.

References

1. Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M. & Smith, K. What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9225–9234 (2022).
2. Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S. & Xie, P. Transfer learning or self-supervised learning? A tale of two pretraining paradigms. arXiv:2007.04234 (2020).
3. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021).
5. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q. & Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9653–9663 (2022).
6. Singh, P., Sizikova, E. & Cirrone, J. Cass: Cross architectural self-supervision for medical image analysis. arXiv:2206.04170 (2022).
7. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y. *et al.* A cookbook of self-supervised learning. arXiv:2304.12210 (2023).
8. Luo, X., Hu, M., Song, T., Wang, G. & Zhang, S. Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In *Medical Imaging with Deep Learning.* https://openreview.net/forum?id=KUmlnqHrAbE (2022).
9. Cho, J. H., Mall, U., Bala, K. & Hariharan, B. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. arXiv:2103.17070 (2021).
10. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
11. Singh, P., & Cirrone, J. A data-efficient deep learning framework for segmentation and classification of histopathology images. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* 385–405 (2023). Springer
12. Singh, P. & Cirrone, J. A data-efficient deep learning framework for segmentation and classification of histopathology images (2022).
13. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, Vol. 32, 8024–8035 (Curran Associates, Inc., 2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
15. Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* 168–172 (IEEE, 2018).
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
17. Thanh, D. N. H., Erkan, U., Prasath, V. B. S., Kumar, V. & Hien, N. N. A skin lesion segmentation method for dermoscopic images based on adaptive thresholding with normalization of color models. In *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)* 116–120 (2019). https://doi.org/10.1109/ICEEE2019.2019.00030
18. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv:1610.02391 (2022).
19. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D. P. & Chen, D. Z. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part III* 408–416 (Springer, 2017). https://doi.org/10.1007/978-3-319-66179-7_47.
20. Bai, W. *et al.* Semi-supervised learning for network-based cardiac MR image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2017* (eds Descoteaux, M. *et al.*) 253–260 (Springer, 2017).
21. Peng, J., Estrada, G., Pedersoli, M. & Desrosiers, C. Deep co-training for semi-supervised image segmentation. *Pattern Recognit.* **107**, 107269 (2020).
22. Yu, L., Wang, S., Li, X., Fu, C.-W. & Heng, P.-A. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 605–613 (Springer, 2019).
23. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. arXiv:1807.05520 (2018).

Author contributions

J.C. conceptualized the S4MI pipeline and led the study design. P.S., R.C., L.C., and S.C. were instrumental in implementing and optimizing the self-supervised and semi-supervised learning techniques. P.S., L.C., M.C., and J.P. made significant contributions to the analysis of the medical imaging datasets. P.S. was responsible for the statistical analysis and validation of the results. C.S., as the clinician collaborator, provided critical insights into the clinical applicability of the research and assisted in interpreting the results from a clinical perspective. J.C. oversaw the entire project, ensuring adherence to scientific rigor, and contributed to the writing and editing of the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information