Bar-Ilan University

Multiple Sclerosis Segmentation Using Semi-Supervised Learning and Soft Labeling

Zeev Hananis

Submitted in partial fulfillment of the requirements for the Master's Degree in the
Multidisciplinary Brain Research Center,
Bar-Ilan University

Ramat Gan, Israel 2023

This work was carried out under the supervision of
Prof. Mina Teicher,
Multidisciplinary Brain Research Center,
Bar-Ilan University, Israel.

Contents

Abstract

Multiple Sclerosis (MS) is a chronic autoimmune disease of the central nervous system that affects more than 2.8 million people globally. MS lesions segmentation refers to differentiating lesion-infected tissues from healthy ones. Deep convolutional neural networks (CNNs) approach has out scaled other machine learning methods in various computer vision tasks, including MS lesion segmentation tasks. However, applying CNN approaches for MS lesion segmentation is challenging, as MS lesions are heterogeneous and characterized by unclear and irregular boundaries. The visual complexity of MS lesions makes the process of manual segmentation extensive in time and labor, and results in a lack of accuracy and agreement between different experts' annotations. In addition, Deep learning methods typically rely on large training datasets with high-quality manual annotations since the network training requires tuning of many parameters, which makes them limited in their adoption and application when trained on small datasets, such as those available for MS. To tackle these issues, we propose to incorporate semi-supervised learning with soft labeling and uncertainty estimation techniques for unlabeled data distillation into the state-of-the-art (SOTA) nnU-Net architecture. We hypothesize:

1) Soft labeling will improve performance compared to baseline.

2) Enlargement of our database in unlabeled data and using semi supervised learning will improve performance compared to baseline.

For training, we used the annotated public training database of the ISBI 2015 challenge and the Tel Aviv Sourasky Medical Center (TASMC) dataset. We evaluated the final models on the ISBI 2015 challenge test dataset using the challenge's metrics.

Our soft labeling method has outperformed the current SOTA ISBI leaderboard architecture, achieving the best dice score coefficient of 0.686 and improving the overall challenge score (93.01) compared to the vanilla nnU-Net (92.87).

1. Introduction

1.1 Multiple Sclerosis

Multiple Sclerosis (MS) is a chronic autoimmune disease of the central nervous system that affects more than 2.8 million people globally. Pathologically, MS is defined by the breakdown of the myelin sheaths that protect the nerve fibers (demyelination) (1). Mobility, hand function, eyesight, cognition, bowel and bladder function, sensory, spasticity, pain, depression, and coordination are all affected, and most individuals notice some level of impairment in most of these domains as early as the first year of the disease (2). Currently, MRI is the modality of choice for diagnosis and treatment response assessment in patients with MS. The standard MRI protocol for routine MS diagnosis and follow-up includes: T1 weighted imaging acquired before and after contrast agent injection (T1WI, T1WI+C), T2 weighted imaging (T2WI), and fluid-attenuated inversion recovery (FLAIR) (3). Typical MS lesions may appear hyperintense on T2WI and FLAIR, hypointense on T1W, and may show contrast enhancement on T1WI+C (Figure 1). In clinical routine, visual based radiological inspection of lesions size, number, type, shape, character and location are used for patient's assessment (4).

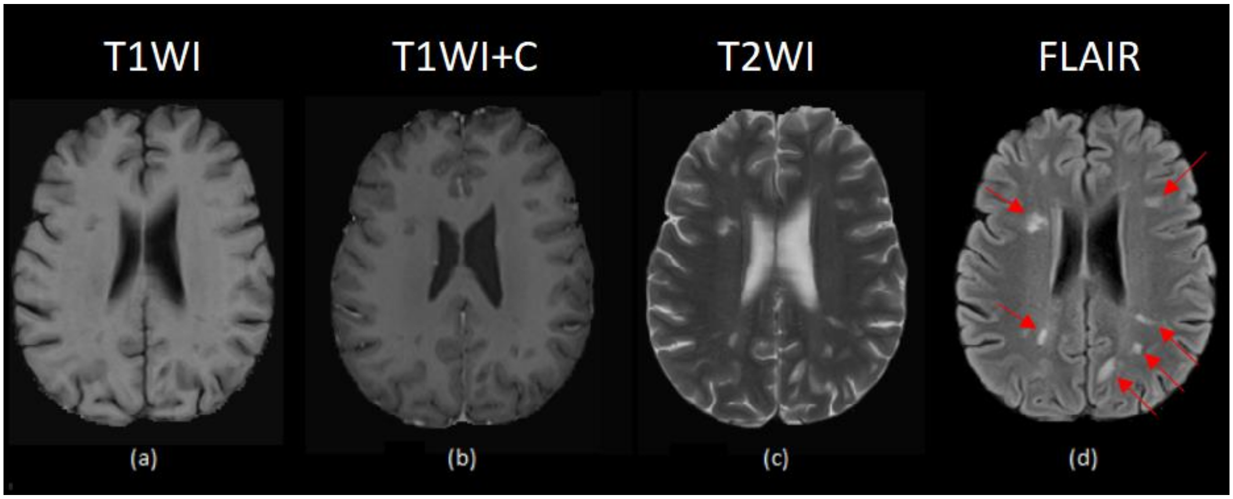

Figure 1. Multiple sclerosis case as featured on different MRI modalities. a. T1 weighted imaging (T1WI), b. T1WI + contrast agent (T1WI+C), c. T2 weighted imaging (T2WI), d. fluid-attenuated inversion (FLAIR). Red arrows indicate MS lesions.

Although such qualitative approaches are usually sufficient for a general diagnosis and may also give a good indication of the severity of the patients' condition, quantitative lesions segmentation may enable early therapy response assessment, and has the potential to improve the personalization of therapy for individual patients by elucidating the underlying relationship between image features and patient condition (1,5).

1.2 Automatic methods for MS lesions segmentation

In recent years, deep convolutional neural networks (CNNs) have been widely used in medical image analysis, including MS lesion segmentation, showing superior performance over conventional machine learning methods (6). The majority of currently used CNN techniques for MS lesions segmentation fall into one of the following two categories: patch-wise segmentation, semantic-wise segmentation. A Patch-wise CNN classifier is trained to classify a pixel using the data from the pixel-centered patch. Semantic segmentation trains a CNN model to directly classify each pixel of the whole input image in a single forward propagation (7). A summary of some of the state-of-the-art (SOTA) methods for automated MS lesion segmentation (6), published based on the public challenge ISBI 2015 (8) is given in Table 1.

Authors	Algorithms	Score	DSC	PPV	LFPR	LTPR	VPCC
Zhang et al., 2021 (6)	ALL-Net	93.32	0.639	0.914	0.122	0.533	0.86
Zhang et al., 2019 (9)	Tiramisu	93.11	0.641	0.902	0.155	0.54	0.867
Isensee et al., 2021 (10)	nnU-Net	92.87	0.679	0.847	0.159	0.523	0.865
Zhang et al., 2021 (11)	GEO Loss	92.73	0.643	0.887	0.132	0.48	0.854
Ma et al., 2021 (12)	Low-precision Ensemble	92.55	0.661	0.838	0.151	0.491	0.854
Hashemi et al., 2018 (13)	Asymmetric Loss	92.48	0.584	0.921	0.087	0.414	0.858
Aslani et al., 2019 (14)	Multi-Branch	92.12	0.611	0.899	0.139	0.41	0.867
Andermatt et al., 2017 (15)	Recurrent Gated Units	92.07	0.629	0.845	0.201	0.487	0.862
Valverde et al., 2017 (16)	Cascaded Network	91.33	0.63	0.787	0.153	0.367	0.866
Birenbaum et al., 2016 (17)	Multi-View	90.07	0.627	0.789	0.498	0.568	0.822

Table 1. Scores and metrics of the state-of-the-art MS lesion segmentation algorithms on the ISBI-2015 testing dataset (6) (see materials and methods). DSC=Dice Similarity Coefficient, PPV= Positive Predictive Value, LFPR=Lesion FPR, LTPR=Lesion TPR, VPCC=Volumes Pearson's Correlation Coefficient.

1.3 3D Unet (nnUnet)

SOTA models for image segmentation are variants of the encoder-decoder architecture, also known as U-Net (19) (Figure 2). These U-net networks used for segmentation share a similar backbone: an encoder subnetwork which extracts features and learns an abstract representation of the input image, a decoder subnetwork which takes the abstract representation and generates a semantic segmentation mask, and skip connections, which connect parallel layers of both subnetworks, creating a shortcut connection and allowing an indirect gradient flow without degradation (20). Isensee et al. proposed the nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation, which is a robust and self-adapting framework based on 2D and 3D vanilla U-Nets (10). nnU-Net is an open-source U-Net based segmentation method that provides a fully automated pipeline, that can be used on any medical dataset for segmentation, including the ISBI dataset. The pipeline uses heuristic rules to automatically determine the data-dependent hyperparameters such as loss function, architecture, batch and patch size, without the need for user interference. The nnU-Net can produce an ensemble of for 2D, 3D and 3D-Cascade U-Net and choose the best configuration that will be used to produce the predictions at inference time. The nnU-Net has shown impressive adaptability and performance on dozens of public segmentation tasks, setting SOTA results or reaching the leaderboard in all of them, including on the MS ISBI 2015 competition, where it reached the SOTA DSC (10).

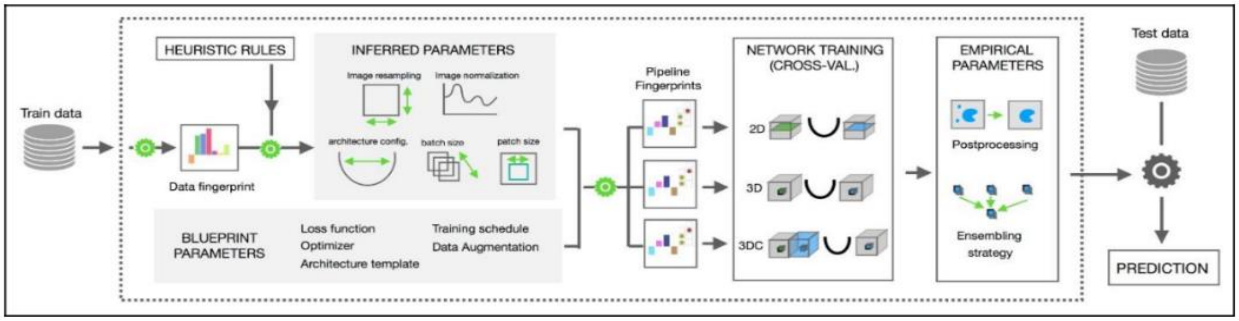

Figure 2: nnU-Net Architecture (10).

1.4 Challenges in automatic MS lesions segmentation

Despite the relative success, methods of automatic MS lesion segmentation do not reach a level of excellence. This may be due to a number of fundamental challenges unique to

MS lesion segmentation. MS lesions are heterogeneous and characterized by unclear and irregular boundaries. The visual complexity of MS lesions leads to inaccuracies and disagreement between experts' annotations (21). Due to the ambiguity of experts' annotations, automated lesion segmentation tools tend to perform poorly and are limited in their adoption and application in MS segmentation tasks (22). Figure 3 shows example of raters' disagreement on the ISBI 2015 training dataset

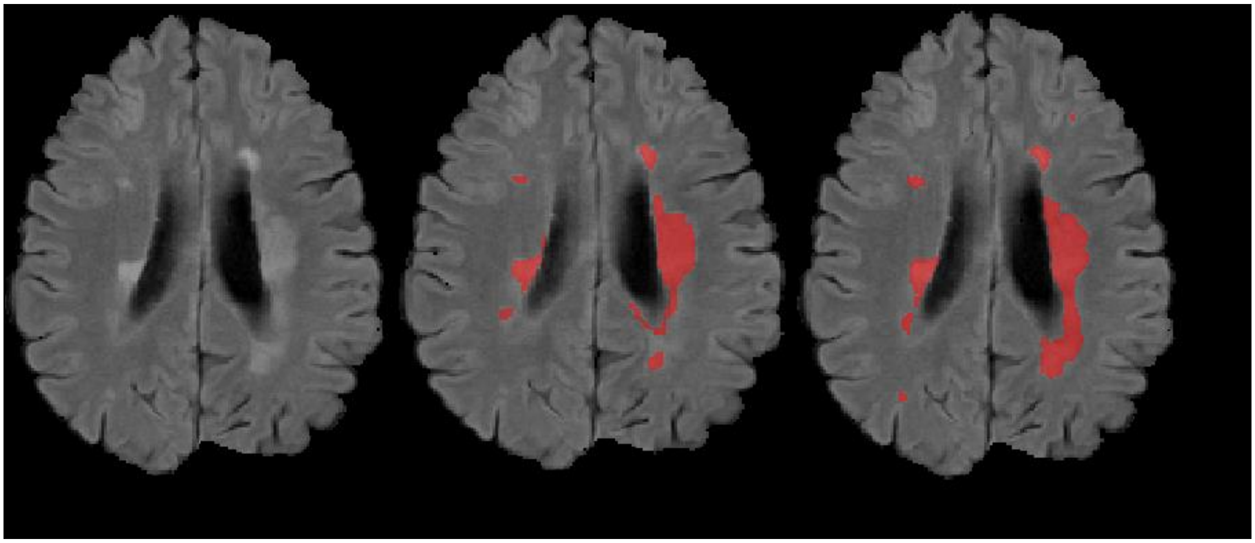

Figure 3: Example of raters' disagreement on the ISBI 2015 training dataset.

1.5 Soft labels

To tackle the issue of annotation inaccuracy, several studies proposed the use of Soft Labeling approaches. Instead of assigning a binary mask, a probability is assigned to each voxel, allowing the trained model some flexibility in areas where the annotations are less certain. The Use of soft labels was shown to provide a better precision-recall tradeoff and to achieve a higher average Dice similarity coefficient (DSC) (23-25). Figure 4 shows an example of soft labels usage in a case of experts' manual segmentation disagreement in MS patient (ISBI 2015 dataset) as shown by Kats E. et al. (24). They created soft labels from the manual annotations with a 3D morphological dilation protocol. Based on the anatomical knowledge, they expanded the labels in the boundaries of the lesions where the experts' annotations are uncertain (24).

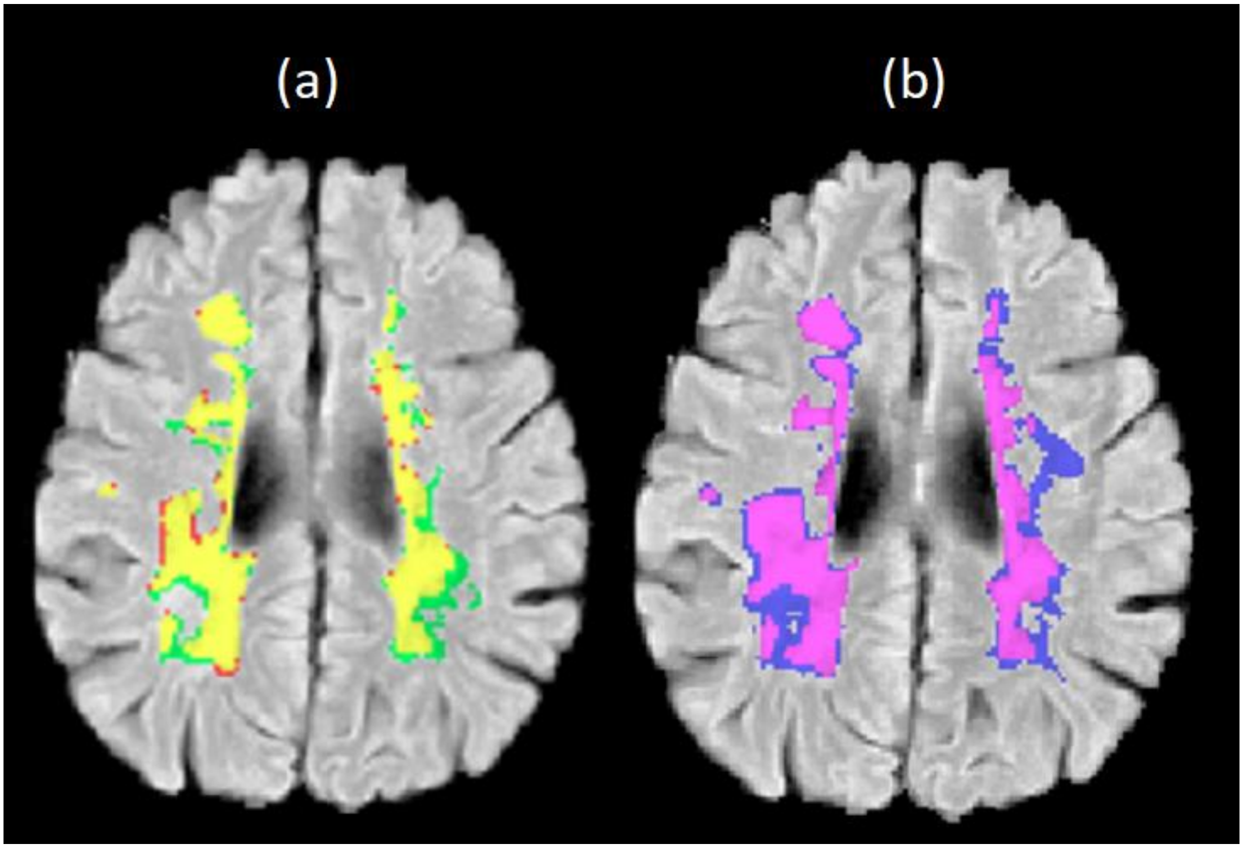

Figure 4. Soft labeling. a. experts' annotations. Yellow - agreement lesion voxels by both raters. Green/Red - voxels marked as lesion by rater 1 and rater 2 respectively. b. Purple - ground truth mask. Blue - soft mask dilation based on ground truth annotation (24).

1.6 Semi supervised learning

In addition, despite the success of CNN in medical image analysis, the lack of large-sized and well-annotated datasets still poses a major bottleneck in the progression and usability of automated methods for the clinical domain (26). In a search for more data-efficient CNN methods to overcome the need for large annotated datasets, there is a rising research interest in semi-supervised learning (SSL) and its applications to deep neural networks to reduce the amount of labeled data required (27). SSL can be introduced into lesion segmentation tasks when only a limited number of annotated training images are available, where a combination of labeled and unlabeled data can be exploited for training CNNs. Most of the semi-supervised methods use some variation of a teacher-student type framework, where a teacher model is trained on a limited annotated dataset in conjunction with a student model that is jointly trained on a combination of labeled and pseudo labeled datasets. Numerous studies have demonstrated that semi-supervised learning outperforms supervised methods and improves the overall performance of the system (28).

1.7 Test time augmentation

SSL has great potential for enlarging and enriching the training dataset, but without assessing the nature of the unlabeled data prediction, its contribution will be limited. Recently, test-time augmentation (TTA) approach was proposed for uncertainty estimation, data distillation in unannotated datasets (29). Uncertainty estimation has been implemented in the literature to measure how diverse the predictions for a given image are, thus increasing the reliability of medical imaging analysis systems (30).

1.8 Study aim

The aim of this study was to improve the SOTA nnU-Net based solution for MS lesion segmentation by incorporate Semi-supervised Learning and soft labeling with uncertainty estimation techniques. Segmentation models were trained and evaluated on the public ISBI 2015 dataset, and Tel Aviv Sourasky Medical Center (TASMC) MS datasets.

2. Methods

2.1 Database

2.1.1 Imaging data

ISBI data set: ISBI 2015 training dataset contains 21 annotated longitudinal cases taken from 5 patients. The test dataset contains 61 longitudinal cases taken from 14 patients. Each case includes fluid attenuation inversion recovery (FLAIR), T2 weighted image (T2WI) and T1 weighted image (T1WI). All images are realigned to the same space with fixed image size of 181×217×181 (8). The segmentation of the MS lesions was manually performed by two experts (Figure 3). The ISBI dataset was used for model training and evaluation.

TASMC data set: Included MRI data of 58 MRI unlabeled cases of 53 patients with MS scanned at the Tel Aviv Sourasky Medical Center (TASMC). MRI data was collected retrospectively from patients' routine clinical assessments performed at TASMC, with different MRI vendors and systems and various acquisition parameters. T1WI, T2WI, and FLAIR are available for all cases. The TASMC dataset was used for model training only.

2.1.2 Data splitting

ISBI training dataset was split at the in stratified manner – at the subject level into 80% training and 20% evaluation in a five-fold cross-validation manner. The unlabeled TASMC dataset was only allocated as part of the training dataset in each fold of training the student model, in the semi supervised teacher-student architecture.

2.2 Tools - algorithm and software

2.2.1 BPT Data preprocessing

Analysis was performed in a python environment and tested on a single graphical processing unit (GPU) CUDA device, NVIDIA Tesla V100-PCIE-32GB.

As part of our work, we have implemented and combined different available tools for data preprocessing into an all-in-one solution, written in the Python environment and referred here as BPT (brain preprocessing tool). Our implementation improved preprocessing performance by using novel tools and by simplifying the workflow. Data preprocessing pipeline is given in Figure 7. **The BPT pipeline includes:**

1. Images alignment: Images acquired from different imaging protocols differ in spatial relations and resolution. Image alignment is the process of finding the optimal transformation or mapping function which will align one image to another (31). The ISBI challenge images and labels were aligned based on the T1WI contrast. In order to incorporate the TASMC dataset into the training of the semi-supervised teacher-student architecture, in conjunction with the T1WI aligned ISBI training dataset, we aligned the FLAIR and T2WI contrasts in each case of the TASMC dataset to the T1WI image using the Elastix tool. Elastix is an intensity-based medical image registration framework that allows the configuration, testing, and comparison of different registration methods (32).
2. Bias field correction: When performing the MRI scans, there may be artifacts that cause distortion in the magnetic field. This distortion causes differences in brightness between areas that are identical in the tissue, and has been shown to affect automatic segmentation results (33). We use the N4 Bias Field Correction algorithm to correct unevenness in brightness caused by this distortion (34).
3. Brain extraction: In order to facilitate the model in the training phase we removed unnecessary areas like the skull and the eyes that are not relevant to the segmentation task. We are using a pre-trained model called HD-BET. It receives MRI scans of all relevant contrasts and outputs the brain image without the skull (35).

2.2.2 nnU-Net

The fully automated preprocessing pipeline begins by cropping the images to a unified size and removing as much background as possible to reduce computational costs. Afterwards, a resampling of all images was done, so all images have the same voxel spacing and was set to 1X1X1. Finally, a simple z-score normalization was applied to each image individually.

All networks were trained for 50 epochs with stochastic gradient descent, a batch size of 2 and patch size of 128×128×128 (with 32 feature maps at the highest resolution). The networks use Adam optimizer function and a custom learning rate scheduler with initial learning rate of 3×10−3.

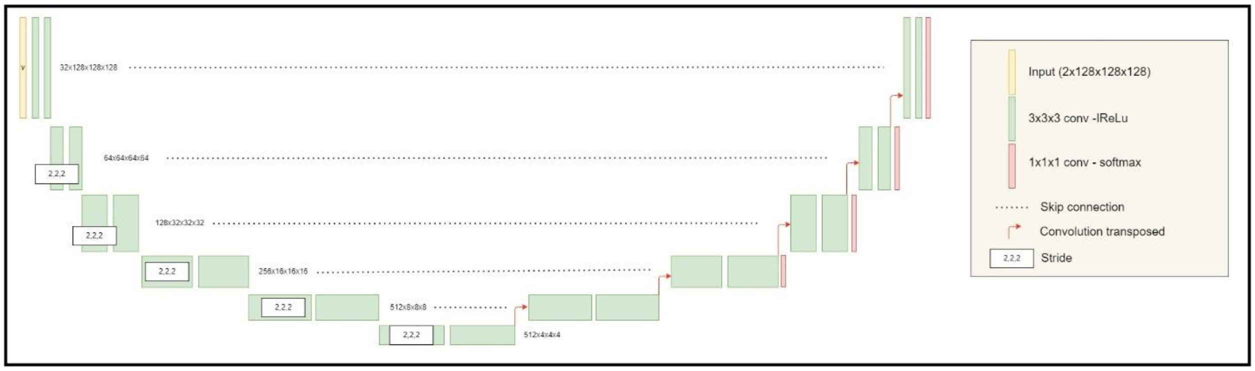

Figure 5: nnU-Net architecture.

2.3 Training procedure

We examined the different combinations of the methods suggested below. All models were evaluated on the ISBI test dataset.

2.3.1 Vanilla nnU-Net

Running a fully supervised training session with the vanilla nnU-Net configuration, to get a baseline score of the original architecture. Other sessions' results were compared to this score. Because ISBI data contains two raters' masks per each case, we composed and compared a network trained on the intersection of the masks and a network trained on the union of the masks (results in Table 2).

2.3.2 nnU-Net + Semi-supervised learning (SSL)

As proposed by Sedai et al. (25), we trained a semi-supervised teacher-student model where the teacher model was trained on the ISBI 2015 labeled dataset in a Bayesian manner to output segmentation uncertainties of the unlabeled TASMC dataset for the student model. After initializing the teacher model, we trained a student model based on the combination of labeled samples and unlabeled samples along with their pseudo labels (Figure 6).

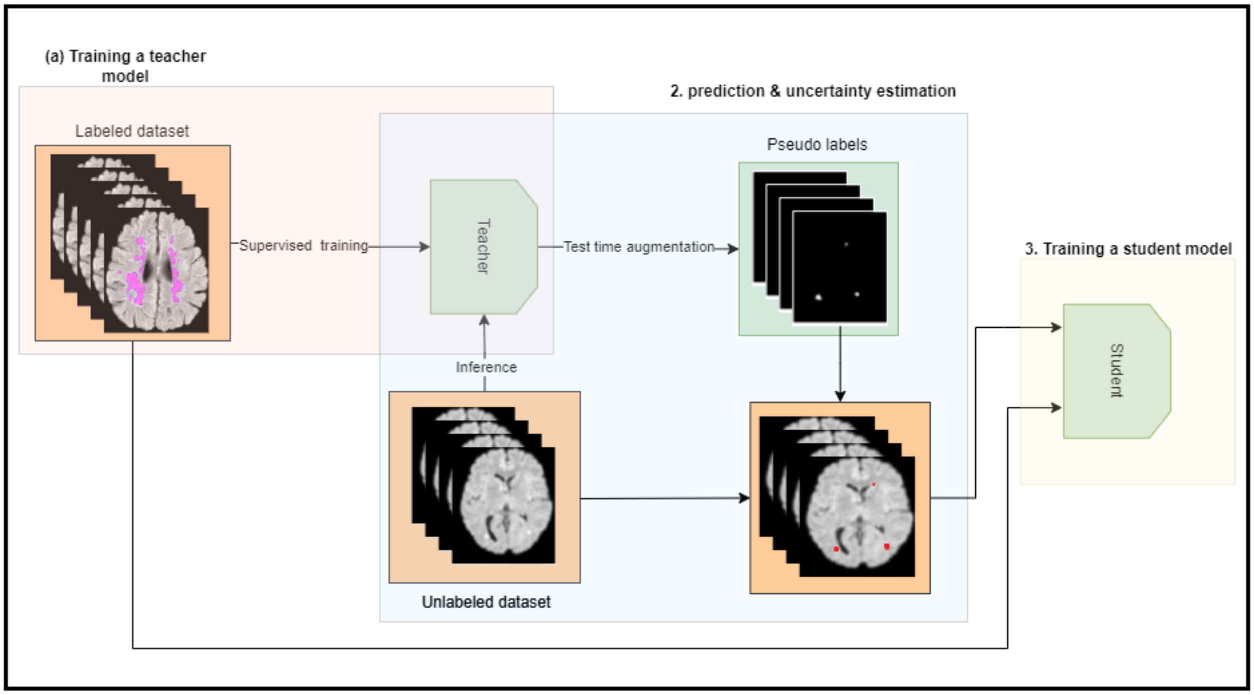

Figure 6. Illustration of the semi supervised architecture.

2.3.3 nnU-Net + Soft labeling (SL)

We trained the teacher models with soft labels as proposed by Kats et al. (24). To create soft labels from the original binary masks, we expanded the annotations of the lesions with a 3D morphological dilation protocol (denoted as γ). Specifically, we calculated the mean intensity of each lesion region separately, as they appear in FLAIR scans. Voxels located on the boundaries of the lesions with intensity values higher than the lesion mean value were assigned a soft label $0 < \gamma < 1$, which represents the probability of the voxel being a part of the lesion. To determine the optimal hyperparameters, we evaluated different combinations of the dilation size and the soft labels values (results shown in table 3-4). As in the baseline model, we compared the different merging methods of the two raters' annotations (Table 2).

To account for the probabilistic values of the soft labels, we used the modified dice loss suggested by Kats et al. (24). Its equation can be written it as follows:

$$(1)\ soft\ dice\ loss = -\frac{\sum_i (T_i+\gamma D_i)*P_i}{0.5\sum_i P_i+0.5\sum_i (T_i+\gamma D_i)}$$

Soft labeled mask is represented as T+γD by denoting the ground truth mask as matrix T, the dilated region as matrix D and the soft label assigned to voxels of the dilated region as γ. P is the predicted probabilistic mask.

2.3.4 nnU-Net + SSL + Test time augmentation (TTA) & Uncertainty estimation

The uncertainty estimation technique we used is based on the work of Sedai et al. (25). Test time augmentation (TTA) was previously used to improve prediction performance by combining predictions of multiple transformed versions of a test image (e.g. flipping, cropping, rotating, and scaling) (29). In this work, we used TTA as a Monte Carlo simulation process. We used a set of eight flip augmentations and calculated the aggregated entropy for each voxel in the predicted set. The average probability per class for each voxel is denoted by $\bar{P}$ is represented by:

$$(2)\ \bar{P}_c = \frac{1}{8}\sum_{k=1}^{K=8} p_k$$

The entropy of the average probability for each voxel is denoted by H is calculated as follows:

$$(3)\ \ H = -\sum_{c=1}^{C=2} \bar{P}_c * log\ \bar{P}_c$$

The entropy represents the uncertainty level of each voxel, higher entropy values represent voxel that the model was less confident about between the different augmentations. After obtaining the entropies for each voxel, we converted the matrix to normalized confidence map ω using equation 4:

$$(4)\ \omega = \ exp^{-\alpha H}$$

The normalized confidence map ω ∊ [0, 1], and α is a positive scalar hyperparameter. We then binarized the uncertainty map, zeroing values less than a threshold VPT (voxel probability threshold) and completely removing cases where the sum of the normalized confidence map was less than a threshold TST (total sum threshold). We empirically set VPT to 0.3 and TST to 100.

The student model was jointly trained on both the labeled samples along with their ground truth, and unlabeled samples along with their pseudo labels.

2.3.5 Integrated model

Finally, we tested the different hyperparameters combinations of all methods above to produce the optimal integrated model of nnU-Net + soft labeling + SSL + TTA.

2.3.6 Evaluation and inference

The optimal hyperparameters of the different experiments were determined based on the scores on the 5-fold validation dataset. The metrics that were used for evaluation are: Dice score coefficient (DSC), precision (positive predictive value) and recall (sensitivity). DSC is the ratio between twice the intersection to the total prediction and ground-truth. TP, FP, FN denote the number of voxel-wise true positives, false positives, and false negatives, respectively.

$$(5)\ DSC = \frac{2TP}{2TP+FP+FN}$$

Precision (PPV) is the ratio of TP to all the predicted voxels.

$$(6)\ PPV = \frac{TP}{TP+FP}$$

Recall is the ratio of TP to all the ground-truth voxels.

$$(7)\ \text{Recall} = \frac{TP}{TP+FN}$$

In addition, the final models in each experiment were evaluated on the ISBI test dataset using the evaluation score originally proposed in the ISBI 2015 challenge:

$$(8)\ ISBI\ score = \frac{1}{2\grave{N}}\sum^{2n}\left(\frac{DSC}{8}+\frac{PPV}{8}+\frac{LTPR}{4}+\frac{1-LFPR}{4}+\frac{VPCC}{4}\right)$$

$\grave{N}$ is a normalization factor considering the inter-rater variation and the number of samples, n is the total number of samples.

VPCC is the Pearson’s correlation coefficient of the lesion volumes between the ground-truth and the prediction.

Lesion false positive rate (LTPR) is the lesion-wise ratio of false positives to the sum of false positives and true negatives.

$$(9)\ LFPR = \frac{FP}{FP+TN}$$

Lesion true positive rate (LTPR) is the lesion-wise ratio of true positives to the sum of true positives and false negatives.

$$(10)\ \ LTPR = \frac{TP}{TP+FN}$$

3. Results

3.1 BPT pipeline

3.2 Soft labeling generation

3.2.1 Dilation methods compression

To obtain the best soft labeling configuration, we needed to examine what is the optimal dilation method. We compared between three conditions (shown in figure 9):

Based on the intersection between ISBI rater 1 and 2 segmentations + soft labels:

1. Dilation based on the intersection of rates (Fig. 9c).

2. soft labels Union – dilation based on intersection with addition of disagreement regions as soft labels (Fig. 9d)

3. Dilation based on the union of rates (Fig. 9e).

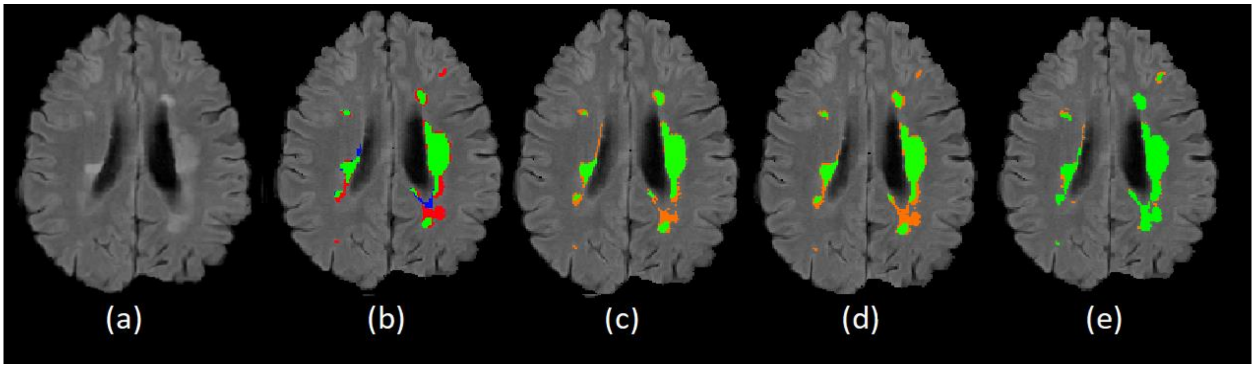

Figure 9: Dilation methods. a. FLAIR image b. original masks of both raters. Green – voxels of agreement between the two raters. Blue/Red – voxels marked as lesion by rater 1 and rater 2 respectively. c. soft labeling based on raters' intersection: Green - voxels of agreement. Orange – soft label dilation. d. soft union labeling: Green - voxels of agreement. Orange – soft label dilation and voxel where raters disagreed. e. soft labeling based on raters' union: Green – voxels marked as lesion by one of the raters. Orange – soft label dilation.

Hyperparameters were fixed with a soft labels value of 0.3 and a dilation size of 1mm 3D ball structuring element. We also included a comparison of baseline models trained on the intersection of the raters' masks and the union of the raters' masks. All models were trained on the ISBI training dataset with the default configurations of nnU-Net. The results were conducted per rater on the validation dataset in a 5-fold cross-validation manner (Table 2). The training time of a single model was about 4-6 hours.

The Best average DSC was achieved with the Intersection + soft union method. Figure 10 shows an example of predictions of the soft labels methods we examined on one of the validation dataset's cases.

Superior merging method – intersection + soft union labels

	Rater 1			Rater 2		
Model	DSC	Precision	Recall	DSC	Precision	Recall
Kats E. et al. (14)	0.699	0.805	0.631	0.620	0.815	0.506
Intersection (baseline)	0.746	0.898	0.648	0.690	0.918	0.561
Union (baseline)	0.766	0.729	0.827	0.750	0.773	0.751
Inter. + Soft labels	0.764	0.857	0.703	0.722	0.790	0.619
Inter. + Soft union labels	**0.770**	0.851	0.718	0.746	0.892	0.643
Union + Soft labels	0.761	0.705	0.844	**0.753**	0.761	0.769

Table 2: Soft labeling methods evaluation on validation dataset.

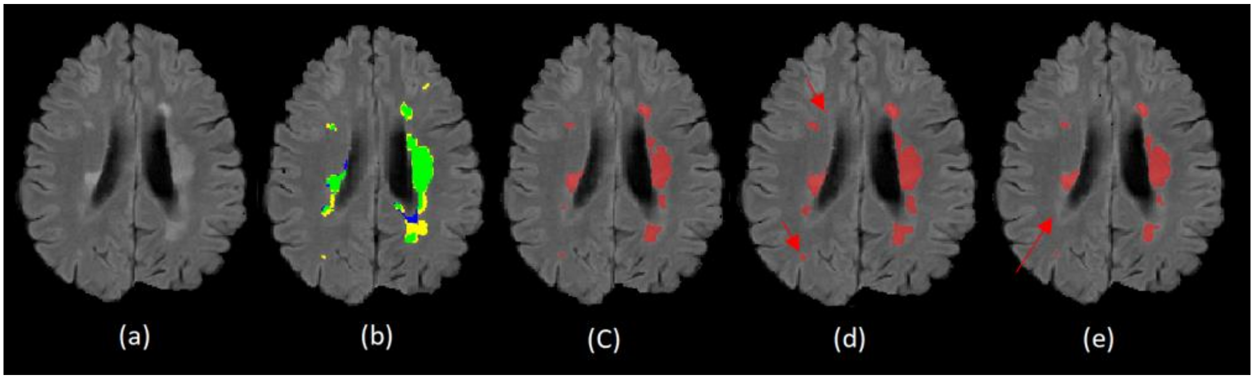

Figure 10: Comparison of predictions of soft labels methods on validation dataset. b. Ground truth masks. Green – regions of agreement between raters. Blue\Yellow - annotations of rater 1 and 2 respectively. c. Prediction of intersection + soft union labels model. d. Prediction of raters' union dilation model. e. Prediction of raters' intersection dilation model. Red arrows indicate region of false positive/negative predictions.

3.2.2 <u>Soft labels hyperparameters comparison</u>

In the same manner, we examined the optimal hyperparameters for soft label generation. First, we evaluated our method for different soft label values. The dilation size was fixed to a value of 1mm 3D ball structuring element. The results were conducted per rater on the validation dataset in a 5-fold cross-validation manner (Table 3). The networks were trained with intersection + soft union masks as labels. The training time of a single model was about 4-6 hours. We found that a soft label value of 0.5 had produced the best DSC average.

Superior soft label value – 0.5

	Rater 1			Rater 2		
Soft label value	DSC	Precision	Recall	DSC	Precision	Recall
0.1	0.695	0.910	0.570	0.744	0.885	0.614
0.3	0.770	0.851	0.718	0.746	0.892	0.643
0.5	**0.771**	0.777	0.783	**0.75**	0.825	0.704
0.7	0.769	0.748	0.811	0.75	0.801	0.717

Table 3: Soft label values evaluation on validation dataset.

Similarly, we determined the best dilation size, with a fixed soft label value of 0.5. As shown in table 3, a 1mm 3D ball structuring element dilation has resulted in best DSC in for raters.

Superior dilation shape – 1mm 3D ball element

	Rater 1			Rater 2		
Dilation size (mm)	DSC	Precision	Recall	DSC	Precision	Recall
1	**0.771**	0.777	0.783	**0.750**	0.825	0.704
2	0.769	0.843	0.720	0.726	0.876	0.635
3	0.762	0.854	0.703	0.717	0.885	0.616

Table 4: Dilation size evaluation on validation dataset.

3.3 SSL with uncertainty estimation

Our semi supervised method is a teacher-student based architecture. We split the SSL optimization procedure into two parts:

1) Teachers output optimization – find the optimal pseudo labels
2) Student training optimization – training the student model with and without soft labeling the pseudo labels.

3.3.1 Teacher's output optimization

After obtaining the optimal soft labeling configuration, the chosen teacher model was used for generating pseudo labels from the TASMC unlabeled dataset to train the student model with. As described in the methods section, TTA was done on each predicted image by augmenting the image with a flip augmentation in eight different axial combinations, and an entropy mask was created based on equations 1 and 2. For the next step of generating the normalized confidence map ω, we needed to find the optimal α parameter. Similarly, to the soft labeling hyperparameters optimization process, all other parameters were fixed to examine different values of α.

The pseudo labels were generated with a teacher model that was trained on the ISBI training dataset, using Intersection + Soft labels masks with a soft labels value of 0.5 that were produced with 1mm 3D ball structuring element dilation.

After calculating the probability map (Figure 11a), we zeroed voxels with values smaller than 0.3 and removed cases where the sum of the map was less than 100 (Figure 11b). The probability masks were binarized, and the student models were trained on the soft labeled ISBI training dataset in conjunction with the binary pseudo labeled TASMC training dataset (Figure 11c).

Superior α value – 0.3

	Rater 1			Rater 2		
α	DSC	Precision	Recall	DSC	Precision	Recall
0.3	0.793	0.820	0.777	0.757	0.855	0.691
0.5	0.796	0.816	0.784	0.751	0.844	0.692
0.7	0.772	0.859	0.708	0.737	0.895	0.626

Table 5: Dilation size evaluation on validation dataset.

3.3.2 Student training optimization

Next, we wanted to examine if dilating the binary pseudo labels (pseudo labels + soft labels) would improve the student's performance over the binary masks. The pseudo labels for both conditions were generated with a teacher model that was trained on Intersection + Soft labels masks with a soft labels value of 0.5 that were produced with a 1mm 3D ball structuring element dilation. We set α to 0.3, and after binarizing the normalized prediction masks (equations 2-4), we dilated the binary pseudo labels using soft labels with a value of 0.5 that were produced with 1mm 3D ball structuring element dilation. Figure 11 shows an example case of each stage of the process.

integration of pseudo labels + soft labels resulted in best dice score average.

	Rater 1			Rater 2		
method	DSC	Precision	Recall	DSC	Precision	Recall
Pseudo-labels	0.793	0.820	0.777	0.757	0.855	0.691
Pseudo-labels + soft labels	**0.801**	0.825	0.783	**0.757**	0.854	0.691

Table 6: pseudo labels dilation evaluation on validation dataset.

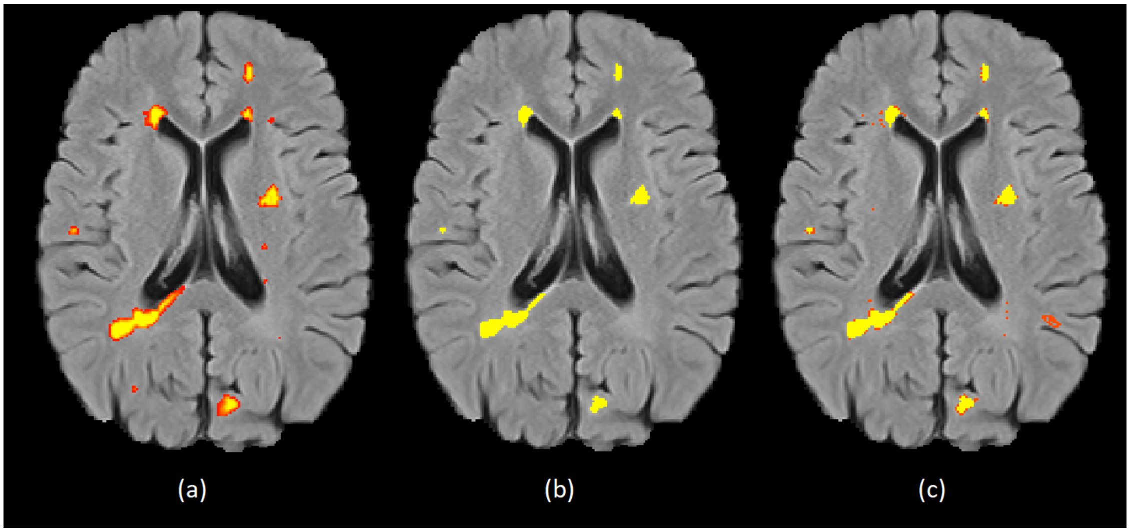

Figure 11: Teacher's output optimization. a. the teacher's output normalized probability mask ∈ [0, 1]. b. binarizing the mask, values less than the 0.5 threshold were zeroed out. c. dilation of the binary mask with a 1mm 3D ball structuring element and soft label value of 0.5.

3.4 Segmentation results

As a final evaluation, we compared the optimal models we produced at each training step with the SOTA algorithms from the ISBI challenge leaderboard.

The specific configurations we used are as follows:

Soft labels model: A nnU-Net trained on the ISBI training dataset using the soft union dilation of both raters' annotations created with a 1mm 3D ball structuring element dilation with soft labels value of 0.5.

SSL + soft labels: A nnU-Net student model. The pseudo labels used for training, were generated from the soft labels model above, converted to normalized probability masks using α value of 0.3 and binarized with a probability threshold of 0.5. The model was trained on the ISBI annotated training dataset in conjunction with the TASMC pseudo labeled dataset.

All models were trained for 50 epochs, with the default nnU-Net configuration.

We evaluated the models on the ISBI test dataset, using the ISBI challenge metrics and score (equations 5-10).

The results are shown in Table 7, representative segmentation results are given in Figure 12.

Algorithm	ISBI Score	DSC	PPV	LFPR	LTPR	VPCC	Epochs
Soft labels	93.01	0.686	0.85	0.129	0.503	0.871	50
SSL + Soft labels	92.45	0.653	0.85	0.111	0.43	0.87	50
ALL-Net (6)	93.32	0.639	0.914	0.122	0.533	0.86	140
Tiramisu (9)	93.11	0.641	0.902	0.155	0.54	0.867	100+
nnU-Net (10)	92.87	0.679	0.847	0.159	0.523	0.865	1000

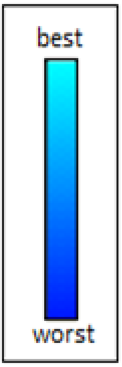

Table 7: Segmentation results on ISBI test dataset. ISBI score = Challenge computed score. DSC = Dice Similarity Coefficient, PPV= Positive Predictive Value, LFPR=Lesion false positive rate, LTPR=Lesion True positive rate, VPCC=Volumes Pearson's Correlation Coefficient. Colors bar shows the results' quality, lighter blue indicates better results.

Figure 12, demonstrates a test case where the intersection + soft union labels model outpreformed the official ISBI challange results of the nnU-Net, specificly in the areas marked with the red arrows, where the vanilla nnU-Net didn't fully maksed the lesions regions.

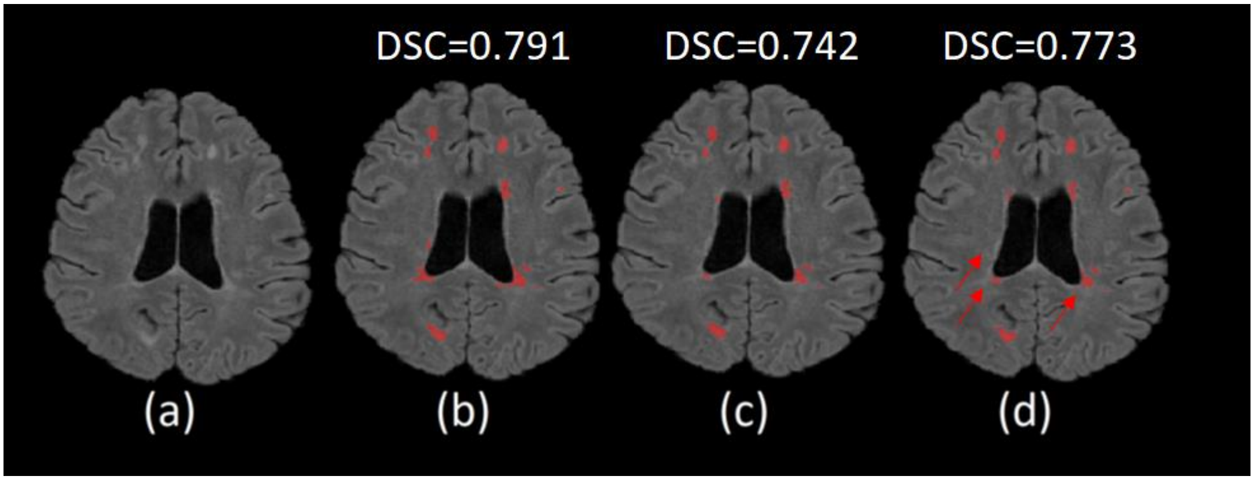

Figure 12: Comparison of predictions with DSC on ISBI test dataset. a. Example test case b. Prediction of the soft labels model. c. Prediction of SSL + soft labels model. d. prediction of the vanilla nnU-Net model. Red arrows indicate region of false negative predictions.

4. Discussion

In this study, we proposed an integration of soft labeling and a simple semi-supervised teacher-student method for improving the nnU-Net architecture for multiple sclerosis lesion segmentation tasks where the data is unbalanced and lesions are ambiguous. In this section, we will go over the experiment's steps and discuss its results.

Our intersection + soft union labels method, with 1mm 3D ball geometry dilation and 0.5 soft labels value, achieved an ISBI score of 93.01 and is ranked 3rd on the ISBI challenge leaderboard (Table 1). It outperformed the vanilla nnU-Net surpassing its results, with a reduction of epochs by a factor of 100 (table 7). The highest DSC 0.686 and the second lowest LFPR we achieved, indicate that training a model with soft labels results in a model that is more sensitive in the lesions' boundaries compared to models trained with binary mask.

Study innovation

BPT improves the preprocessing pipeline

We utilized an automated MRI brain image preprocessing tool (BPT). This pipeline integrates SOTA solutions for MRI images realignment, brain extraction and intensity normalization (32-35). The tool The integrated tool receive raw dicom files, and provide end-to-end solution for the data preparation as input data for the deep learning model. The BPT tool has been proven to be robust and to work on images produced by different vendors with different image configurations, contrasts, and orientations (Figure 8).

Soft labeling improves model generaliation

In this work, we defined soft labels as the dilated probability masks of the ground truth annotations in the regions of the lesions' borders, where voxels have a similar characteristics to those inside the lesions. These regions are usually represent areas of annotators uncertainty and disagreement. Kats E. et al. (14) have constructed a modified dice loss function that takes into account the probabilistic soft label mask. We began this experiment by determining which dilation method will result in the optimal performance. Using the ISBI training dataset, we compared three type of raters' masks merge methods: intersection + soft labels, intersection + soft union, and union + soft labels. Overall, among all the model we tested, The intersection + soft union method gave the best DSC average and the optimal tradeoff between precision and recall on the validation dataset, and was superior to the results gained by Kats et al. (Table 2). We also showed that the soft labeling methods, intersection/union + soft labels surpass the parallel baseline methods on the validation dataset. Our intersection + soft union labels method, with 1mm 3D ball geometry dilation and 0.5 soft labels value, has also resulted in a better score 93.01 and IS ranked 3rd on the ISBI challenge leaderboard (Table 1), outperforming the official nnU-Net benchmark results. The highest DSC (0.686) and the second lowest LFPR we achieved, demonstrate the benefit of soft labels, making the model more sensitive to topological changes in the boundaries of the lesions compared to the parallel model trained on binary masks.

A future implementation that may improve our results is to make the soft labeling more voxel-specific, taking into account the relative intensity and location of each boundary voxel, and assigning it a specific soft label value, giving voxels with high intensity and proximity to the center of the lesion a higher probability to be regarded as such.

Semi supervised learning (SSL)

We have implemented a teacher-student architecture. To mitigate the effect of the pseudo labels created by the teacher model and preventing the student model to be biased, we produced the pseudo labels using uncertainty estimation, where voxels with low certainty were removed from the predicted mask.

Based on the validation dataset, the teacher-student architecture that yield the best results was a as follows:

Teacher: the optimal soft label model we discussed earlier, that was trained with intersection + soft union labels method, with 1mm 3D ball geometry dialtion and 0.5 soft labels value.

Student: trained on the same soft labels as the teacher model in conjunction with the TASMC dataset.

The pseudo labels were produced by the teacher with uncertainty estimation, using the optimal α hyperparameter (Table 5) to create the normalized probability mask (equation 4). The masks were binarized with a threshold of 0.5, and the binary masks were dilated using soft labels (pseudo labels + soft labels) (Figure 11).

The architecture has achieved the best DSC on the validation dataset, compared to all other method we evaluated. However, we were not able to improve the test results we achieved with the soft labels model, receiving a score of 92.45 (Table 7) ranking seventh on the leaderboard (table 1). Nevertheless, this method achieved the second best FTPR score, meaning that the model is not overestimating and have a lower false positives rate compared to the vanilla nnU-Net benchmark.

Study limitations

One of the technical issue we faced during while working with 3D nnU-Net architecture, is the need for extremely high computational power. Although we had high performance computational infrastructure, with 2 available GPUs (CUDA device, NVIDIA Tesla V100-PCIE-32GB), the nnU-Net architecture we used, with three contrast channels (FLAIR, T1W, T2W) with high resolution 3D images and more than 31 million parameters to fine tune, required a significant capacity and computational resources. Due to this fact, We had to shorten the models' training time so that we could conduct all the trails we planned. The default nnU-Net model is usually trained for 1000 epochs. To balanced between the running time and number of models, we limited the training to 50 epochs. Nevertheless, we were still able to outperformed the original ISBI challenge, and achieve a better DSC and an overall score with our soft label model (intersection + soft union labels). Perhaps, if we had trained the model for a longer period of time, we could have achieved even better results.

Secondly, the unlabeled training dataset used by the semi supervised models was rather small, containing 58 samples. SSL enables to reduce the model's reliance of annotated dataset, but on the other hand, it requires a considerable amount of unlabeled dataset to be able to generalize well.

Future directions

This work is still ongoing. We are currently implementing an automated MS images preprocessing and lesions segmentation pipeline, to be integrated in the Ichilov environment and that will be used for research and clinical purposes.

In addition, this work is currently under preperation for publicationUnder this framework, we are collecting additional cases to try and improve the results of the SSL models. .

5. References

1. Lladó X, Oliver A, Cabezas M, et al. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Information Sciences 2012;186(1):164-185.
2. Kister I, Bacon TE, Chamot E, et al. Natural history of multiple sclerosis symptoms. International journal of MS care 2013;15(3):146-156.
3. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 2005;58(6):840-846.
4. Traboulsee A, Simon J, Stone L, et al. Revised recommendations of the consortium of MS centers task force for a standardized MRI protocol and clinical guidelines for the diagnosis and follow-up of multiple sclerosis. American Journal of Neuroradiology 2016;37(3):394-401.
5. Pagnozzi AM, Gal Y, Boyd RN, et al. The need for improved brain lesion segmentation techniques for children with cerebral palsy: A review. International Journal of Developmental Neuroscience 2015;47:229-246.
6. Zhang H, Zhang J, Li C, et al. ALL-Net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. NeuroImage: Clinical 2021;32:102854.
7. Zeng C, Gu L, Liu Z, Zhao S. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. Frontiers in Neuroinformatics 2020;14:610967.
8. Carass A, Roy S, Jog A, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 2017;148:77-102.
9. Zhang H, Valcarcel AM, Bakshi R, et al. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer; 2019. p. 338-346.
10. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 2021;18(2):203-211.
11. Zhang H, Zhang J, Wang R, et al. Geometric loss for deep multiple sclerosis lesion segmentation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI): IEEE; 2021. p. 24-28.
12. Ma T, Zhang H, Ong H, et al. Ensembling low precision models for binary biomedical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. p. 325-334.
13. Hashemi SR, Salehi SSM, Erdogmus D, Prabhu SP, Warfield SK, Gholipour A. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 2018;7:1721-1735.
14. Aslani S, Dayan M, Storelli L, et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage 2019;196:1-15.
15. Andermatt S, Pezold S, Cattin PC. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. International MICCAI Brainlesion Workshop: Springer; 2017. p. 31-42.

16. Valverde S, Cabezas M, Roura E, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 2017;155:159-168.
17. Birenbaum A, Greenspan H. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. Deep learning and data labeling for medical applications: Springer; 2016. p. 58-67.
18. Ghafoorian M, Karssemeijer N, Heskes T, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. NeuroImage: Clinical 2017;14:391-399.
19. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support: Springer; 2018. p. 3-11.
20. Kazerouni IA, Dooly G, Toal D. Ghost-UNet: An asymmetric encoder-decoder architecture for semantic segmentation from scratch. IEEE Access 2021;9:97457-97465.
21. Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Computerized Medical Imaging and Graphics 2018;70:83-100.
22. Gabr RE, Coronado I, Robinson M, et al. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: a large-scale study. Multiple Sclerosis Journal 2020;26(10):1217-1226.
23. Chang Q, Yan Z, Lou Y, Axel L, Metaxas DN. Soft-Label guided semi-supervised learning for Bi-ventricle segmentation in cardiac cine MRI. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI): IEEE; 2020. p. 1752-1755.
24. Kats E, Goldberger J, Greenspan H. Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019): IEEE; 2019. p. 1563-1566.
25. Sedai S, Antony B, Rai R, et al. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer; 2019. p. 282-290.
26. Yuan M, Liu Z, Wang F, Jin F. Rethinking labelling in road segmentation. International Journal of Remote Sensing 2019;40(22):8359-8378.
27. Ouali Y, Hudelot C, Tami M. An overview of deep semi-supervised learning. arXiv preprint arXiv:200605278 2020.
28. Chebli A, Djebbar A, Marouani HF. Semi-supervised learning for medical application: A survey. 2018 International Conference on Applied Smart Systems (ICASS): IEEE; 2018. p. 1-9.
29. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 2019;338:34-45.
30. Sun L, Wu J, Ding X, Huang Y, Wang G, Yu Y. A Teacher-Student Framework for Semi-supervised Medical Image Segmentation From Mixed Supervision. arXiv preprint arXiv:201012219 2020.
31. El-Gamal FE-ZA, Elmogy M, Atwan A. Current trends in medical image registration and fusion. Egyptian Informatics Journal 2016;17(1):99-124.
32. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 2009;29(1):196-205.
33. Li C, Gore JC, Davatzikos C. Multiplicative intrinsic component optimization (MICO) for MRI bias field estimation and tissue segmentation. Magnetic resonance imaging 2014;32(7):913-923.

34. Raffelt D, Dhollander T, Tournier J-D, et al. Bias field correction and intensity normalisation for quantitative analysis of apparent fibre density. Proc Intl Soc Mag Reson Med. Volume 25; 2017. p. 3541.

35. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. Human brain mapping 2019;40(17):4952-4964.

תקציר

טרשת נפוצה היא מחלה אוטואימונית כרונית של מערכת העצבים המרכזית הפוגעת ביותר מ-2.8 מיליון אנשים ברחבי העולם. פילוח נגעי טרשת נפוצה מתייחס להבחנה בין רקמות נגועות בנגעים לרקמות בריאות. שימוש ברשתות עצביות מתקפלות העפילה על שיטות אחרות של למידת מכונה במשימות ראייה ממוחשבת שונות, כולל משימות פילוח נגעים בטרשת נפוצה. עם זאת, יישום גישות רשתות עצביות מתקפלות לפילוח נגעי טרשת נפוצה הוא מאתגר, שכן נגעי טרשת נפוצה הינם הטרוגניים ומאופיינים בגבולות לא ברורים ולא סדירים. המורכבות החזותית של נגעי טרשת נפוצה הופכת את תהליך הפילוח הידני לתובעני בזמן ובמאמץ, ומביאה לחוסר דיוק והסכמה בין תיוגי מומחים שונים. בנוסף, שיטות למידה עמוקה מסתמכות בדרך כלל על מערכי אימון גדולים עם תיוגים ידניים באיכות גבוהה, שכן אימון הרשת דורש כוונון של פרמטרים רבים, מה שהופך אותם למוגבלים באימוץ וביישום שלהם כאשר הם מאומנים על מערכי נתונים קטנים, כמו אלה הזמינים עבור תקשת נפוצה. כדי להתמודד עם בעיות אלו, אנו מציעים לשלב למידה מפוקחת-למחצה עם טכניקות תיוג רך ואומדן אי ודאות עבור זיקוק נתונים ללא תיוגים בארכיטקטורת nnU-Net המתקדמת. אנו משערים:

1) תיוג רך ישפר את הביצועים בהשוואה לקו הבסיס.

2) הגדלה של מסד הנתונים שלנו בנתונים לא מתויגים ושימוש בלמידה מפוקחת למחצה ישפרו את הביצועים בהשוואה לקו הבסיס.

לצורך הערכה, השתמשנו במאגר הציבורי של אתגר ISBI 2015 ובמערך הנתונים של המרכז הרפואי תל אביב סוראסקי. הערכנו את המודלים הסופיים בעזרת מערך הנתונים של האתגר ISBI 2015 תוך שימוש במדדים של האתגר.

שיטת התיוג הרכה שלנו עלתה על ביצועי ארכיטקטורת המובילות, והשיגה את dice score הטוב ביותר של 0.686. כמו, כן השיטה שיפרה את ציון האתגר הכולל (93.01) בהשוואה לnnU-Net המקורי (92.87).

עבודה זו נעשתה בהדרכתה של פרופ' מינה טייכר

המרכז הרב תחומי לחקר המוח ע״ש גונדה של אוניברסיטת בר-אילן.

אוניברסיטת בר-אילן

פילוח נגעים של טרשת נפוצה בעזרת למידה מפוקחת-למחצה ותיוג רך

זאב חנניס

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך ביחידה ללימודים בין תחומיים, המרכז הרב תחומי לחקר המוח של אוניברסיטת בר-אילן

רמת גן תשפ"ב