2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Semi-Supervised Learning Framework to Leverage Proxy Information for Stroke MRI Analysis

Jennifer Polson*, Haoyue Zhang*, Kambiz Nael, Noriko Salamon, Bryan Yoo, Namkug Kim, Dong-Wha Kang, William Speier, and Corey W. Arnold.

Abstract— **Treating acute ischemic stroke (AIS) patients is a time-sensitive endeavor, as therapies target areas experiencing ischemia to prevent irreversible damage to brain tissue. Depending on how an AIS is progressing, thrombolytics such as tissue-plasminogen activator (tPA) may be administered within a short therapeutic window. The underlying conditions for optimal treatment are varied. While previous clinical guidelines only permitted tPA to be administered to patients with a known onset within 4.5 hours, clinical trials demonstrated that patients with signal intensity differences between diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) sequences in an MRI study can benefit from thrombolytic therapy. This intensity difference, known as DWI-FLAIR mismatch, is prone to high inter-reader variability. Thus, a paradigm exists where onset time serves as a weak proxy for DWI-FLAIR mismatch. In this study, we sought to detect DWI-FLAIR mismatch in an automated fashion, and we compared this to assessments done by three expert neuroradiologists. Our approach involved training a deep learning model on MRI to classify tissue clock and leveraging time clock as a weak proxy label to supplement training in a semi-supervised learning (SSL) framework. We evaluate our deep learning model by testing it on an unseen dataset from an external institution. In total, our proposed framework was able to improve detection of DWI-FLAIR mismatch, achieving a top ROC-AUC of 74.30%. Our study illustrated that incorporating clinical proxy information into SSL can improve model optimization by increasing the fidelity of unlabeled samples included in the training process.**

Stroke, MRI, Semi-supervised Learning, Deep Learning

I. Introduction

Stroke remains a leading cause of long-term disability; acute ischemic stroke (AIS) accounts for 87% of the 795,000 strokes that are diagnosed each year. [1] For ischemic stroke patients, treatments such as thrombolysis and thrombectomy aim to restore blood flow to areas experiencing ischemia. Successful intervention for both treatments is contingent upon many clinical factors. Thrombolytics reperfuse tissue that is experiencing ischemia but that is not yet infarcted. Until recently, thrombolysis was only recommended for patients meeting certain clinical criteria and with a known symptom onset within 4.5 hours. [2] As many as 35% of patients were deemed ineligible for this treatment due to unknown time since stroke (TSS). [3] However, the recent WAKEUP clinical trial provided a new avenue for patients to receive thrombolytic treatment. [4] Using MR imaging, a neuroradiologist may assess differences in signal between diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) series when TSS is unknown. [5] Several clinical trials have illustrated that ischemic tissue is visible almost immediately after stroke onset, in contrast to ischemic tissue appearing several hours after onset on FLAIR imaging. [6] The most recent version of the American Heart Association guidelines for treating acute stroke assert that, for patients with unclear time of symptom onset, MRI can be performed to identify areas with DWI-FLAIR mismatch that could benefit from thrombolytic treatment. [7] This can offer another therapeutic avenue for these patients; however, this assessment is prone to a large amount of reader variability. [6] Some semi-automated quantitative methods have been developed to determine DWI-FLAIR mismatch, but threshold-based methods may not be reliable on multi-institutional datasets due to differences in MR acquisition protocols. [8] An automated method to detect this observation could help clinicians in the decision to offer thrombolysis. These techniques can parallel previous approaches to automatically determine lesion age by classifying the time since stroke onset. [9]–[11]

Deep learning can leverage medical images to automatically classify signal differences; one primary challenge with this approach is that acquiring high-quality annotations is costly to implement at a large scale, as it requires assessments from multiple domain experts. Approaches to tackle classification tasks with manually intensive labels aim to address small sample size. [12] Semi-supervised learning (SSL), for example, has been explored widely across multiple medical domains by assigning pseudolabels to unlabeled datasets, and incorporating them into the training dataset. [13] This carries the risk of confirmation bias, that is, incorrectly assigning pseudolabels and therefore influencing the loss function to optimize to the incorrect minimum. [14] Another commonly-utilized approach is weakly-supervised learning, whereby samples with unverified labels are evaluated for their proximity to the fully-annotated label. [15] In this work, we propose an automatic, semi-supervised framework that jointly learns from fully and weakly labeled samples. Our framework performs label inference on a large

*Haoyue Zhang and Jennifer S Polson contributed equally.

This work was supported by the following grants: NIH T32EB016640, NIH R01NS100806, and an NVIDIA Academic Hardware Grant.

J. Polson, H. Zhang, and W. Speier are with the Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA, USA.

K. Nael, N. Salamon, and B. Yoo are with the Department of Radiology, University of California, Los Angeles, Los Angeles, CA, USA.

N. Kim is with the Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea.

D. Wang is with the Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea.

C.W. Arnold is with the Departments of Radiology, Pathology, and Bioengineering, and Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, CA, USA (Email: cwarnold@ucla.edu).

dataset by leveraging weak proxy information. In addition to the implementation of established regularization methods, we evaluate the utility of proxy labels to inform sample selection from a set of weakly labeled data. We evaluate this framework using our previously-published deep learning backbone for stroke images, and we do so by testing classification performance on a separate validation dataset from an external institution.

II. Methods

A. Dataset

The data used for this study comprised patient MR imaging studies from two institutions, all of which were diffusion-weighted. Our preprocessing protocol for image series involved skull stripping, N4-bias field correction, registration to an anatomical atlas, intensity normalization, and histogram matching. [16] Each neuroradiologist performed assessments independently and in the same order; the DWI-FLAIR mismatch label was determined by a majority vote. The development dataset was drawn from the existing UCLA Health stroke registry. This cohort contains 417 patients from 2011-2019 who underwent diffusion-weighted MRI with a known TSS. For external validation, we used a cohort from Asan Medical Center (AMC) totaling 355 patients. [11] All of these patients have a known TSS; in addition, three neuroradiologists performed DWI-FLAIR mismatch assessments for 72 and 55 patients from the UCLA and AMC datasets, respectively. Patient records were collected in accordance with respective IRB approval and HIPAA compliance standards. Informed consent was waived via the exemption for retrospective data.

B. Problem Formulation

Our dataset contained two labels: time clock (TSS) and tissue clock (DWI-FLAIR mismatch). One the one hand, TSS is available for all patients in both datasets, but it serves as an imperfect proxy for the the underlying tissue changes that create ischemic tissue that has not yet experienced infarction. Conversely, DWI-FLAIR mismatch serves as a better approximation of salvageable tissue targeted by thrombolytic treatments, but it is manually intensive to generate these labels across the entire dataset, and it is prone to inter-reader variability. Thus, we will consider TSS as a weak proxy label, and DWI-FLAIR mismatch as a full label. Each fully-labeled patient can be categorized as one of the following:

$$Weak\ Proxy\ 1,\ Target\ 1\ (Clean)$$
$$Weak\ Proxy\ 1,\ Target\ 0\ (Noisy)$$
$$Weak\ Proxy\ 0,\ Target\ 1\ (Noisy)$$
$$Weak\ Proxy\ 0,\ Target\ 0\ (Clean)$$

We will consider X as the total set of samples that are available for our semi-supervised framework. Within this, there are two subsets: X_F, which are fully labeled samples, and X_W, which are samples that only have the weak label. For this study, we can assume that each sample in X_F has both the weak proxy and manually-acquired full label.

C. Classification Models

Our framework utilized two models that were trained in a decoupled fashion. The first was a deep learning model $g_\theta \rightarrow \mathbb{R}^d \rightarrow P$ that served as both a feature extractor and target label classifier, where $\mathbb{R}^d$ represents extracted features and P indicates the final binary classification for the target variable. The convolutional backbone was based on ResNet-18, as that was determined to be the optimal architecture from previous stroke MRI classification studies.

The second model was a discriminator $D_S \rightarrow L$ that classified samples based on the relationship between their weak proxy and target labels. That is, the model determined a sample's likelihood that the weak proxy matched the target. The model took features extracted from g_θ as input and computed cosine similarity to those features extracted for each of the four data categories outlined in II-B.

D. Label Refining Framework

Our framework consisted of two stages. In the first stage, we trained our deep learning model. Once the deep learning architecture was sufficiently trained, we froze the network for it to serve as a feature extractor. In the second phase, we extracted features for both X_F and X_W. Using these extracted features, we trained D_S to classify samples in X_W based on their cosine similarity to samples in X_F. We then used these classifications to infer labels for samples in X_W for which there was high confidence that the pseudolabel was correct. In this respect, we only included samples for which the weak proxy label for X_W^i matched the classification determined by D_S. These high-confidence samples were then incorporated into the training set, and the process iterates again starting at stage 1. Our framework algorithm is depicted in Algorithm 1.

Algorithm 1: Proxy Learning Framework

```
for iter ∈ 1, 2, 3, ..., 10 do
    X̂_t = augment(X_t);
    train g_θ using X̂_t
    while g_θ fixed do
        extract features g_θ → ℝ^d
        train D_S
        X̂_w = sample(X_w)
        for i ∈ X̂_W do
            Compute s(i, C) = (i · C) / (||i|| · ||C||)
            Assign cluster label L_i = min(s_i, C)
            if s_i > 0.5
            & L_i^proxy = i^proxy then
                if iter in early stage then
                    L_i == "Clean"
                else
                    L_i == "Clean" | "Noisy"
                add L_i to X_t
            else
                continue
test g_θ on X_E: g_θ → ℝ^d → P_E
```

E. Experiments

For all experiments, we used the same initial training set X_F, comprising 72 patients from UCLA with both proxy and target labels. A total of 345 patients X_W, all from the UCLA Stroke dataset, were used as potential pseudo-labeled data in the semi-supervised stage of training. All models and experiments were tested on a set of patients X_E, which comprised 56 patients from Asan Medical Center. For the deep learning model g_θ, hyperparameters were tuned in accordance with previous experiments for this architecture.

We hypothesize that proxy information can supplement the semi-supervised framework training process in the following ways: using proxy labels to select weakly labeled samples for inclusion in the training set, and incorporating samples of varying noisiness at different training iterations. We compared our proposed framework to a baseline, fully-supervised network trained on fully-supervised samples alone. We also compare to the current state-of-the-art in semi-supervised learning, a method performed by Berthelot et al. that involves consistency regularization on augmented samples. [17] To test the effects of our proposed techniques, we also completed ablation studies for each methodological adaptation implemented in our framework. The primary metric used to perform this study was receiving-operator characteristic (ROC) area under the curve (AUC), though we also report detection sensitivity and specificity. Each model was run ten times to report mean performance across metrics.

III. Results

Our experimental results are summarized in 1 and Table I. When tested on an unseen external validation dataset, our semi-supervised framework was able to achieve an average ROC-AUC of 74.30±1.9%. This model outperformed the current state of the art in SSL for our DWI-FLAIR mismatch detection task, and it also achieved the lowest variance when run in replicate. Our ablation experiments illustrated that both methods to utilize proxy information enhanced model performance.

TABLE I

Performance metrics across experiments and our proposed framework.

Model	AUC[a]	Sens.[b]	Spec.[c]
Fully Supervised	64.36±6.63%	64.15±29.8%	45.16±27.7%
Baseline SSL	67.11±3.58%	60.97±14.53%	70.45±14.26%
+ Noise Selection	67.35±3.82%	70.65±11.13%	60.91±10.84%
+ Proxy Selection	70.82±3.52%	75.81±9.49%	61.36±8.92%
+ Our Model	**74.30±1.9%**	73.87±10.74%	69.08±8.81%
MixMatch	63.9±2.4%	72.4±5.4%	70.1±10.3%

[a]Receiving Operator Characteristic Area Under the Curve

[a]Sensitivity [a]Specificity

IV. Discussion

Determining the age of an acute ischemic stroke lesion is essential to informing stroke treatments. The relationship between time clock and tissue clock has been long studied, identifying TSS as a surrogate proxy for the progression of ischemic tissue. Clinical imaging has illustrated tissue changes underlying this progression. Signal intensity differences between DWI and FLAIR imaging were originally proposed as a method to identify patients within 3 hours of stroke onset. [6] More recently, it has also been incorporated into stroke treatment guidelines, as the presence of DWI-FLAIR mismatch can be used to give patients thrombolytics when onset time is unknown. This imaging biomarker has also been clinically correlated with better outcomes for other stroke treatments such as mechanical thrombectomy. [18], [19] Detecting DWI-FLAIR mismatch has two challenges: it requires assessment by an expert neuroradiologist, and it is an inherently subjective assessment prone to inter-reader variability. [20] This proxy-target paradigm could be extended to other medical tasks, where an easily-collected clinical variable serves as a surrogate proxy for an underlying label that requires expert annotation. [21] A common problem among medical tasks is labels are expensive to acquire, leading to training on small datasets with questionable generalization when evaluated solely on data from one institution. One way to address this is by evaluating generalizabilty using an unseen, external dataset.

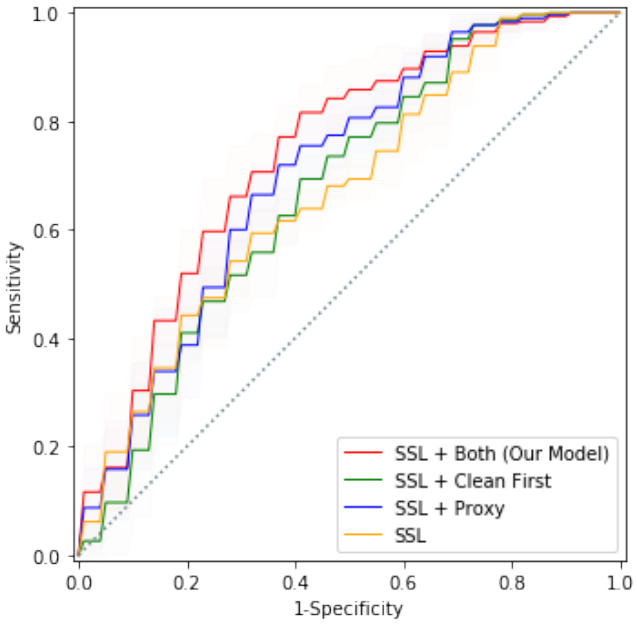

Fig. 1. ROC Curves illustrating the average performance of our proposed method alongside ablation experiments.

Many areas of machine learning research have explored methods to incorporate prior knowledge into their models, as this can be particularly informative for medical detection and segmentation tasks. Our approach utilizes prior information in a semi-supervised framework in two ways: to stratify unlabeled examples into clinically meaningful categories, and to classify samples according to the level of noise. Both of these methods have the goal of enhancing confidence in pseudo-labeled samples. Combining these two strategies yielded

both higher detection performance and lower variability across replicates, the latter addressing instability of machine learning models trained on small datasets. Moreover, our proposed method outperformed the current state of the art in semi-supervised learning, which involves calculating the consistency of predictions on augmented samples. The image registration pipeline creates spatially aligned images such that stroke location is information contained within the image. The types of augmentation used in natural image datasets, e.g., flipping, rotation, and translation, may be less useful in a model where spatial location could be informative. [22] Given that medical datasets, even without labels, are often small, learning a good feature representation that is reliant on data augmentation method may not be as effective.

Our study sought to leverage TSS as a surrogate to detect DWI-FLAIR mismatch in MR imaging performed on acute ischemic stroke patients. The experiments show that proxy information used in a semi-supervised learning framework can enhance performance both in terms of increasing classification accuracy and model stability. Our study has a few limitations, namely the small sample size for both the development and evaluation datasets, and the small number of experts that generated the annotated labels. Future research directions could include evaluating this approach on a larger multi-institutional dataset, which could enable stratification of patients by clinical factors such as demographics and medical history. Clinical segmentation of patient populations could elucidate clinical factors that influence tissue progression during ischemic stroke. A wider evaluation of this approach could produce a model capable of automatically detecting DWI-FLAIR mismatch from MRI taken at imaging, which could help inform treatment options for AIS patients.

REFERENCES

[1] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 03 2019.

[2] G. Thomalla *et al.*, "A multicenter, randomized, double-blind, placebo-controlled trial to test efficacy and safety of magnetic resonance imaging-based thrombolysis in wake-up stroke (WAKE-UP)," *Int J Stroke*, vol. 9, no. 6, pp. 829–836, Aug 2014.

[3] V. C. Urrutia, R. Faigle, S. R. Zeiler, E. B. Marsh, M. N. Bahouth, M. C. Trevino, J. L. Dearborn, R. Leigh, S. Rice, K. Lane, M. O. Saheed, P. M. Hill, and R. H. Llinás, "Safety of intravenous alteplase within 4.5 hours for patients awakening with stroke symptoms," *PLoS ONE*, vol. 13, 2018.

[4] G. Thomalla *et al.*, "Mri guided thrombolysis for stroke with unknown time of onset," *The New England Journal of Medicine*, vol. 379, p. 611–622, 2018.

[5] M. R. Etherton, A. D. Barreto, L. H. Schwamm, and O. Wu, "Neuroimaging Paradigms to Identify Patients for Reperfusion Therapy in Stroke of Unknown Onset," *Front Neurol*, vol. 9, p. 327, 2018.

[6] G. Thomalla *et al.*, "DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4.5 h of symptom onset (PRE-FLAIR): a multicentre observational study," *Lancet Neurol*, vol. 10, no. 11, pp. 978–986, Nov 2011.

[7] W. J. Powers *et al.*, "Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association," *Stroke*, vol. 50, no. 12, pp. e344–e418, Dec 2019.

[8] A. Wouters, B. Cheng, S. Christensen, P. Dupont, D. Robben, B. Norrving, R. Laage, V. N. Thijs, G. W. Albers, G. Thomalla *et al.*, "Automated dwi analysis can identify patients within the thrombolysis time window of 4.5 hours," *Neurology*, vol. 90, no. 18, pp. e1570–e1577, 2018.

[9] K. C. Ho, W. Speier, H. Zhang, F. Scalzo, S. El-Saden, and C. W. Arnold, "A Machine Learning Approach for Classifying Ischemic Stroke Onset Time From Imaging," *IEEE Trans Med Imaging*, vol. 38, no. 7, pp. 1666–1676, 07 2019.

[10] K. C. Ho, W. Speier, S. El-Saden, and C. W. Arnold, "Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features," *AMIA Annu Symp Proc*, vol. 2017, pp. 892–901, 2017.

[11] H. Lee, E.-J. Lee, S. Ham, H.-B. Lee, J. S. Lee, S. U. Kwon, J. S. Kim, N. Kim, and D.-W. Kang, "Machine learning approach to identify stroke within 4.5 hours," *Stroke*, vol. 51, no. 3, pp. 860–866, 2020.

[12] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[13] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[14] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[15] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[16] H. Zhang, J. S. Polson, K. Nael, N. Salamon, B. Yoo, S. El-Saden, F. Scalzo, W. Speier, and C. W. Arnold, "Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time," *arXiv preprint arXiv:2011.03350*, 2020.

[17] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[18] S. Escalard, B. Gory, M. Kyheng, J.-P. Desilles, H. Redjem, G. Ciccio, S. Smajda, J. Labreuche, M. Mazighi, M. Piotin *et al.*, "Unknown-onset strokes with anterior circulation occlusion treated by thrombectomy after dwi-flair mismatch selection," *European journal of neurology*, vol. 25, no. 5, pp. 732–738, 2018.

[19] R. Fahed, A. Lecler, C. Sabben, N. Khoury, C. Ducroux, V. Chalumeau, D. Botta, E. Kalsoum, W. Boisseau, L. Duron *et al.*, "Dwi-aspects (diffusion-weighted imaging–alberta stroke program early computed tomography scores) and dwi-flair (diffusion-weighted imaging–fluid attenuated inversion recovery) mismatch in thrombectomy candidates: An intrarater and interrater agreement study," *Stroke*, vol. 49, no. 1, pp. 223–227, 2018.

[20] A. Ziegler, M. Ebinger, J. B. Fiebach, H. J. Audebert, and S. Leistner, "Judgment of flair signal change in dwi–flair mismatch determination is a challenge to clinicians," *Journal of neurology*, vol. 259, no. 5, pp. 971–973, 2012.

[21] A. Jamthikar, D. Gupta, N. N. Khanna, L. Saba, T. Araki, K. Viskovic, H. S. Suri, A. Gupta, S. Mavrogeni, M. Turk *et al.*, "A low-cost machine learning-based cardiovascular/stroke risk assessment system: integration of conventional factors with image phenotypes," *Cardiovascular diagnosis and therapy*, vol. 9, no. 5, p. 420, 2019.

[22] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.