

Soft labels noise tolerant loss functions for transcranial Doppler ultrasound signal classification.

Yamil Vindas
CREATIS Laboratory*
LYON, France
yamil.vindas@creatis.insa-lyon.fr

Blaise Kévin Guépié
Laboratoire Informatique et Société Numérique
Université de Technologie de Troyes
Troyes, France
blaise_kevin.guepie@utt.fr

Marilys Almar
Atys Medical
Soucieu-en-Jarrest, France
marilys.almar@atysmedical.com

Emmanuel Roux
CREATIS Laboratory*
LYON, France
emmanuel.roux@creatis.insa-lyon.fr

Philippe Delachartre
CREATIS Laboratory*
LYON, France
philippe.delachartre@creatis.insa-lyon.fr

*Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69XXX, LYON, France

Abstract—Transcranial Doppler (TCD) is an ultrasound monitoring method that enables real-time measurements of blood flow velocity, primarily in the Middle Cerebral Artery. Its common application lies in monitoring patients at risk of stroke by effectively detecting micro-emboli. This is done through the detection of high intensity transient signals (HITS), which can be categorized between artifacts, gaseous emboli, and solid emboli. State-of-the-art methods for HITS classification are not able to capture the uncertainty of HITS soft-annotation, nor the noise in their soft-labels (soft-noise), both coming from the expert annotation doubt. To better handle this, we propose to train deep learning models using soft labels cost functions, such as the soft cross entropy and the Jensen-Shanon divergence (JSD), which directly approximates the soft distribution of the input samples instead of a hard label proxy. We evaluate the robustness of our approach against symmetrical soft label noise in terms of final hard classification (using the Matthews correlation coefficient, MCC) and human expert uncertainty capturing (using the Hellinger distance). The obtained models trained with JSD soft-labels were robust against soft-noise with improvements of up to 24% in terms of MCC. At last, these models were able to better capture the human expert uncertainty of the true labels, achieving Hellinger distance improvements up to 0.10 (relative gap of 16%).

Index Terms—Transcranial Doppler, Soft labels, Supervised learning, Emboli classification, Stroke

I. INTRODUCTION

Stroke ranks as the second most prominent contributor to global mortality [1], while also standing out as a key factor behind disability. Ischemic stroke, resulting from the blockage of a brain-supplying artery, represents the prevailing form of stroke [2]. The presence of transcranial Doppler (TCD) microembolic signals has been associated to ischemic stroke risks [3], making their detection important to help clinicians prevent strokes.

TCD is a non-invasive ultrasound monitoring technique enabling real-time, long-term measurements of blood flow velocity principally in the middle cerebral artery. One of

its main applications involves the monitoring of stroke-prone patients by identifying micro-emboli through high-intensity transient signals (HITS), which can be then classified in three categories: artifacts (Art.), gaseous emboli (GE), and solid emboli (SE). Several works have applied signal processing techniques to create handcrafted features used to classify HITS, based on wavelet and Fourier transforms [4] [5] [6] [7]. More recent works have used deep learning methods to automatically extract feature from the Doppler signal and/or a time-frequency representations [8] [9], pushing further the limits of HITS classification. However, these methods are not able to capture the uncertainty of HITS soft-annotation [10], nor the noise in their soft labels (soft-noise), both coming from the expert annotation doubt.

In this work, we propose to take into account the uncertainty of HITS annotation by directly training deep learning models using soft annotation through adapted loss functions. Similar approaches have been adopted in other contexts by Wang et al. [11] and Engleson et Azizpour [12]. The first team proposed a noise-tolerant loss function, the symmetric cross entropy (SymCE) in the context of hard-label classification, partially solving the label-noise sensibility of classical cross entropy (CE), while having relatively fast convergence. The second team, proposed another hard-label noise robust loss function, a generalization of the Jensen-Shannon divergence (JSD), interpolating between the CE (nonrobust with fast convergence) and the mean absolute error (robust with slow convergence) loss functions.

Following previous works, here we proposed to adapt different noise-tolerant loss functions in the context of noisy soft labels, to improve the HITS classification while taking into account the uncertainty of manual annotation. To our knowledge, this is the first work to perform noisy soft label classification on TCD data deep learning techniques.

II. PROPOSED METHOD

Let us suppose that we have a dataset $\mathcal{D} = \{(X_1, y_1), \dots, (X_N, y_N)\}$ compose of N soft labeled samples. Without loss of generality, we suppose that for all $i \in [1, N]$, $X_i \in \mathbb{R}^M$ and $y_i \in \mathbb{R}^{1 \times C}$ where M is the dimension of the input space, and C is the number of classes. Moreover, for $i \in [1, N]$, $k \in [1, M]$, $p \in [1, C]$, we denote as X_i^k the k^{th} element of X_i and y_i^p the p^{th} element of y_i . Moreover, We suppose that we have a function $OH : [0, 1]^C \rightarrow [0, 1]^C$ creating a hard label one-hot encoding from one soft label $y \in [0, 1]^C$: $OH(y)_p = 1$ if $p = \arg\max(y)$, else $OH(y)_p = 0$. At last, let us suppose that we have a model $\mathcal{M}$ taking as input a sample X_i with $i \in [1, N]$ and giving as output a prediction $\tilde{y}_i = \mathcal{M}(X_i) \in \mathbb{R}^{1 \times C}$.

A. Preliminaries

We introduce the two main loss functions used for hard label classification, the CE and the SymCE [11]. If we denote as $\mathcal{L}_{CE}$ the CE loss function, and $\mathcal{L}_{SymCE}$ the SymCE loss function, for all $(X, y) \in \mathcal{D}$:

$$\mathcal{L}_{CE}(OH(y), \tilde{y}) = H(OH(y), \tilde{y}) \quad (1)$$

$$\mathcal{L}_{SymCE}(OH(y), \tilde{y}) = \alpha \times H(OH(y), \tilde{y}) + \beta \times H(\tilde{y}, OH(y)) \quad (2)$$

where H is the cross entropy function, α, β are hyperparameters of $\mathcal{L}_{SymCE}$.

B. Working with soft labels

Contrary to hard labels where one single class $i \in [1, C]$ is associated to one sample X_i (i.e. $\forall p \in [1, C], y_i^p = 0$ if $p \neq i$ and $y_i^p = 1$ if $p = i$), soft labels assigns membership scores for each class, often normalized in order to sum to one:

$$\forall i \in [1, N], p \in [1, C], y_i^p \in [0, 1] \text{ and } \sum_{l=1}^C y_i^l = 1$$

In fact, hard labels are just a particular case of soft labels, where all the labels are certain to belong to only one class. Therefore, several loss functions, originally made for hard labels, can be easily extended to soft labels.

C. Soft cross entropy and Jensen-Shannon divergence

Here we introduce the two main loss functions used in this work, the soft CE and the JSD [12], used in the context of soft labels classification. If we denote as $\mathcal{L}_{SoftCE}$ the soft CE loss function, and $\mathcal{L}_{JSD}$ the JSD loss function, for all $(X, y) \in \mathcal{D}$:

$$\mathcal{L}_{SoftCE}(y, \tilde{y}) = H(y, \tilde{y}) \quad (3)$$

$$\mathcal{L}_{JSD}(y, \tilde{y}) = \frac{1}{2} \times (KL(y||m) + KL(\tilde{y}||m)) \quad (4)$$

where H is the cross entropy function, KL is the Kullback-Leibler divergence, and $m = \frac{1}{2} \times (y + \tilde{y})$.

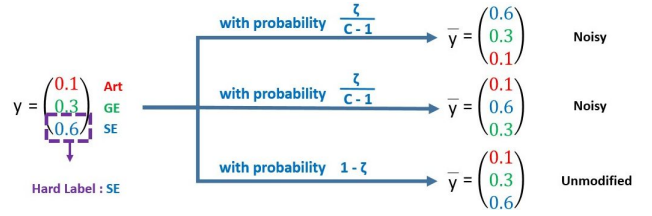


Fig. 1: Approach to simulate symmetric noise on soft labels. The probability of changing of hard label is the same for all the other labels, but the change is done in terms of soft score.

D. Simulation of symmetric soft label noise

In order to test the tolerance of different loss functions for soft noisy labels, we have to adapt the definition of symmetric noise for hard labels, to the soft labels case. Indeed, when working with hard labels, noise is called *symmetric* of noise rate ζ , if for all $i \in [1, N]$, $p, l \in [1, C]$,

$$P(\bar{y}_i = e_k | y_i = e_l) = \begin{cases} 1 - \zeta & \text{if } p = l \\ \frac{\zeta}{C-1} & \text{if } p \neq l \end{cases} \quad (5)$$

where $\bar{y}_i$ is the corrupted noisy version of y_i , and for all $c \in [1, C]$ e_c denotes a vector of the standard basis of $\mathbb{R}^{1 \times C}$ such that the c^{th} is equal to 1 and the rest are equal to 0.

Therefore, to add symmetric noise to soft labels, we proceed as follows (see figure 1).

$$P(\arg\max(\bar{y}_i) = k | \arg\max(y_i) = l) = \begin{cases} 1 - \zeta & \text{if } p = l \\ \frac{\zeta}{C-1} & \text{if } p \neq l \end{cases} \quad (6)$$

In this way, a sample can change of (hard) class with equal probability to the other classes, but the change is made in terms of soft scores instead of hard labels.

III. EXPERIMENTAL SETUP

A. Data description

A total of 39 subjects from 11 different centers underwent TCD recordings using an Atys Medical TCD Robotized Holter device (TCD-X) with a probe frequency of 1.5 MHz. The recordings were conducted for durations ranging from 30 to 180 minutes. From these TCD signals, spectrograms were generated and High Intensity Transient Signals (HITS) were identified using a threshold of 9 dB. In total, 1 541 HITS were detected, each lasting 250 ms. Subsequently, the spectrograms of each HITS were transformed into images, which were then used to train 2D time-frequency convolutional neural networks (CNNs).

Finally, the 1 541 HITS were subject-wise divided into two splits: 63% for training, and 37% for testing.

B. Used architectures

For the different experiments, we used the two models introduced in [9], namely a 2D convolutional neural network (CNN) model taking as input a time-frequency representation of the Doppler signal, and a 1D CNN-transformer taking as input the raw Doppler signal.

C. Experiments

We conduct two main experiments to evaluate our approach: (1) noise resistance in terms of classification performance, (2) noise resistance in terms of uncertainty capturing. For each experiment, we trained each model using four loss functions: $\mathcal{L}_{CE}$, $\mathcal{L}_{SoftCE}$, $\mathcal{L}_{SymCE}$, $\mathcal{L}_{JSD}$. For each experiment, we vary the noise rate (of the labels) going from 0 (no noise) to 50%.

For the training hyperparameters, all the 2D CNN models were trained using ADAM optimizer, with a batch size of 32 during 50 epochs, and a learning rate of 10^{-3} for all the loss functions, except for $\mathcal{L}_{JSD}$ where a learning rate of 10^{-4} was used. The 1D CNN-transformer models were trained using Noam optimizer with 150 epochs, batches of size 32, a learning rate of 10^{-1} , and 2000 warmup steps. At last, both experiments were repeated 10 times for statistical purposes, and the announced metrics are the mean and standard deviation, obtained on the testing set.

IV. RESULTS AND DISCUSSION

A. Experiment 1: classification label-noise resistance

In this experiment, we evaluate the classification performances (through the Matthews correlation coefficient, MCC, on the hard label prediction of the models) of different models based on the loss function used for training and on the level of symmetric noise presented in the training set. Results can be found in figure 2.

We observe that, for the 2D CNN model the most robust loss function, is the JSD soft label loss function. Indeed, compared to the hard labels loss functions CE and SymCE (noise robust), JSD outperforms them by a margin from 1.37% to 23.58% in terms of MCC. Moreover, this loss function also globally outperforms the SoftCE soft label cost function, by a margin up to 16.22% MCC. This can be justified by the fact that JSD tries to approximate the true soft label distribution instead of a hard label proxy. Therefore, especially for high levels of noise, we can achieve better classification performances, as some uncertainty is introduced in the models' predictions.

On the other hand, for the 1D CNN-transformer, the behavior is different, as for noise rates smaller than 30 % SymCE yields higher performances (up to 6.22% MCC gap). However, for higher noise rates, JSD outperforms all the other loss functions by a margin up to 2.76% MCC. An explanation for this is that, theoretically, SymCE is a robust loss function, so for small levels of noise, it can achieve great results.

B. Experiment 2: uncertainty capturing label-noise resistance

In this experiment, we evaluate the capacity of the different loss functions to train models capable of capturing the uncertainty of the human expert annotation uncertainty. To do this, we measure the Hellinger distance of the different trained models, based on the noise rate. The Hellinger distance is commonly used to compare probability distributions, higher values indicating more discrepancies between the two compared distributions. In our case, we compare the predicted

test distribution by the models, with respect to the test human annotator distribution. Results can be found in figure 3.

First, if we compare the soft label loss function JSD against the hard label loss function CE, we can see that for both models the former outperforms the latter as it achieves smaller distances with respect to the true distribution, with gaps up to 0.10 (relative gap of 16%) in terms of Hellinger distance. This can be justified by the fact that soft label loss functions tries to directly approximate the distribution of the true soft labels, which can be uncertain based on the human expert annotation. Thus, this uncertainty is better approximated by these type of functions, compared to hard labels cost functions using artificial hard (certain) labels.

Second, compared to SymCE, JSD also seems to better capture human expert uncertainty for both models, especially for noise rates greater than 30%, with improvements up to 0.07 (relative gap of 12%) in terms of Hellinger distance. This high level noise is a plausible case, especially if working with semi-automatic data annotation methods [13], therefore JSD can be more interesting than SymCE in those cases.

V. CONCLUSION

This work proposes to do cerebral emboli classification by training deep learning models using soft label loss functions, which are relatively robust against noisy labels, and which are able to better capture the uncertainty of annotation of human experts. Two main experiments were performed to study these two points on artificially label noisy datasets, showing improvements up to 24% in terms of MCC for classification, and 0.10 (relative gap of 16%) in terms of Hellinger distance for uncertainty capturing.

As future work, we plan to develop a more complex noise-robust soft-label loss function, based on the geometric Jensen-Shannon divergence offering a trade-off between speed convergence and noise tolerance.

ACKNOWLEDGMENT

This work was carried out in the context of the CAREMB project funded by the Auvergne-Rhône-Alpes region, within the *Pack Ambition Recherche* program. This work was performed within the framework of the LABEX CELYA (ANR-10-LABX-0060) and PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

REFERENCES

- [1] Donkor ES. Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life. *Stroke Res Treat.* 2018;2018:3238165. Published 2018 Nov 27. doi:10.1155/2018/3238165
- [2] NHS website. (2019, August 20). Causes. Nhs.Uk. <https://www.nhs.uk/conditions/stroke/causes/>
- [3] J. Serena, T. Segura, M. Castellanos, et A. Dávalos, Microembolic signal monitoring in hemispheric acute ischaemic stroke: a prospective study, *Cerebrovasc Dis.*, vol. 10, n 4, p. 278-282, 2000, doi: 10.1159/000016070.
- [4] H. S. Markus et M. Punter, Can Transcranial Doppler Discriminate Between Solid and Gaseous Microemboli?: Assessment of a Dual-Frequency Transducer System, *Stroke*, vol. 36, n 8, p. 1731-1734, août 2005, doi: 10.1161/01.STR.0000173399.20127.b3.

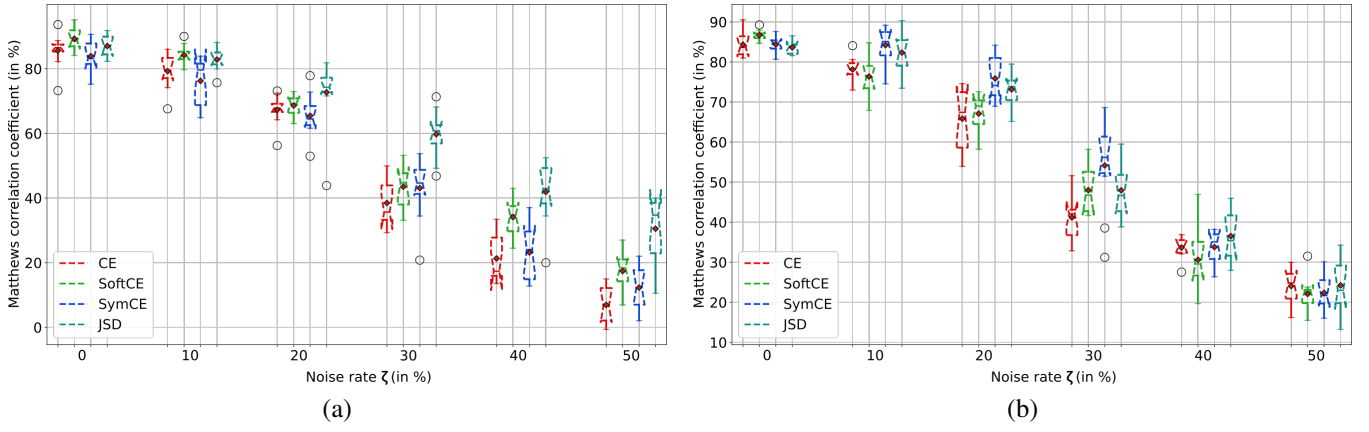


Fig. 2: Experiment 1: classification performance of two types of models trained using different soft and hard label loss functions. (a) 2D CNN [9]. (b) 1D CNN-transformer [9]. The classification performance is measured through the Matthews correlation coefficient (MCC) based on the hard label prediction of the models. The soft label cost function JSD generally outperforms the other cost functions, especially for high levels of noise (greater than 30%).

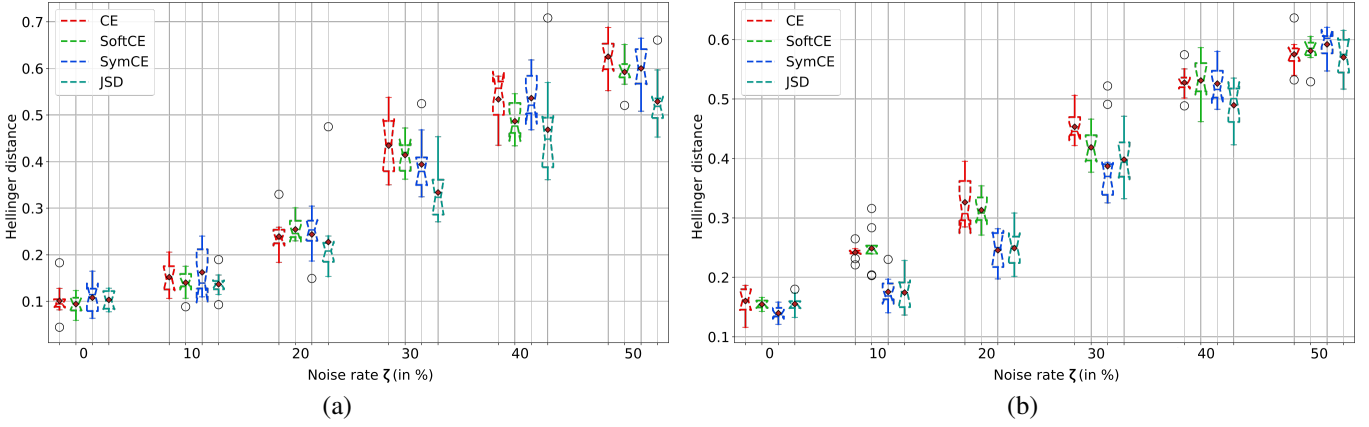


Fig. 3: Experiment 2: uncertainty capturing of two types of models trained using different soft and hard label loss functions. (a) 2D CNN [9]. (b) 1D CNN-transformer [9]. We used the Hellinger distance to compare the soft label distribution of the human expert annotations, with respect to the soft label predictions of the different models. The soft label cost function JSD generally outperforms the other cost functions, especially for high levels of noise (greater than 30%).

- [5] M. Gençer, G. Bilgin, et N. Aydın, Embolic Doppler ultrasound signal detection via fractional Fourier transform, *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2013, p. 3050-3053, 2013, doi: 10.1109/EMBC.2013.6610184.
- [6] G. Serbes et N. Aydın, Denoising performance of modified dual-tree complex wavelet transform for processing quadrature embolic Doppler signals, *Med Biol Eng Comput*, vol. 52, n 1, p. 29-43, janv. 2014, doi: 10.1007/s11517-013-1114-x.
- [7] B. K. Guépié, M. Martin, V. Lacrosaz, M. Almar, B. Guibert and P. Delachartre, Sequential Emboli Detection From Ultrasound Outpatient Data, in *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 334-341, Jan. 2019, doi: 10.1109/JBHI.2018.2808413.
- [8] P. Sombune, P. Phienphanich, S. Phuechpanpaisal, S. Muengtawepong, A. Ruamthanthong, et C. Tantibundhit, Automated embolic signal detection using Deep Convolutional Neural Network, *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2017, p. 3365-3368, juill. 2017, doi: 10.1109/EMBC.2017.8037577.
- [9] Y. Vindas, E. Roux, B. K. Guépié, M. Almar, et P. Delachartre, Guided deep embedded clustering regularization for multifeature medical signal classification, *Pattern Recognition*, vol. 143, p. 109812, nov. 2023, doi: 10.1016/j.patcog.2023.109812.
- [10] R. Du et al., Validation of soft labels in developing deep learning algorithms for detecting lesions of myopic maculopathy from optical coherence tomographic images, *The Asia-Pacific Journal of Ophthalmology*, vol. 11, n 3, p. 227, juin 2022, doi: 10.1097/APO.0000000000000466.
- [11] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, et J. Bailey, Symmetric cross entropy for robust learning with noisy labels, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, oct. 2019, p. 322-330. doi: 10.1109/ICCV.2019.00041.
- [12] E. Engleson et H. Azizpour, Generalized jensen-shannon divergence loss for learning with noisy labels, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, p. 30284-30297.
- [13] Y. Vindas, B. K. Guépié, M. Almar, E. Roux, and P. Delachartre, "Semi-automatic data annotation based on feature-space projection and local quality metrics: An application to cerebral emboli characterization," *Medical Image Analysis*, vol. 79, p. 102437, Jul. 2022, doi: 10.1016/j.media.2022.102437.