Semi-supervised annotation of Transcranial Doppler ultrasound micro-embolic data

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

Emmanuel Roux
CREATIS Laboratory*
LYON, France
emmanuel.roux@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

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 a non-invasive ultrasound monitoring technique allowing real time measurements of the blood flow velocity mainly in the Middle Cerebral Artery. It is commonly used to monitor patients with stroke risk by detecting micro-emboli. This technique generates a considerable amount of data whose annotation is expensive and time-consuming. We propose a semi-supervised learning method to annotate a dataset of micro-embolic data with reduced requirements of manual annotation. First, we start by extracting features from TCD spectrograms in an unsupervised manner using an Auto-encoder. Then, we project those features in a 2D space using the t-SNE algorithm. Afterwards, the dataset is partially annotated by an expert based on the 2D projection and the overall data information. Finally, the labels of the annotated samples are propagated to a part of the unlabeled samples using a K-nearest-neighbors (KNN) strategy. We evaluate our annotation method through the annotation error and the percentage of labeled samples with respect to the unlabeled samples. Our results show that we are able to annotate all the unlabeled samples with an annotation error of 12 % in less than 1 second against around 2 hours for a human expert. This represents a time saving of a factor of 10^5, showing the interest of our method.**

***Index Terms*—Semi-supervised learning, Emboli classification, Stroke, Data Annotation**

I. Introduction

Stroke is the second leading cause of death in the world [1], and one of the leading causes of disability. The most common type of stroke is ischaemic stroke [2] caused by the blockage of an artery that supplies blood to the brain. Microembolic signals have been associated to ischaemic strokes [3], hence the importance of its detection to prevent and treat this medical condition.

Transcranial Doppler (TCD) is a non-invasive ultrasound monitoring technique allowing real time measurements over long periods of time of the blood flow velocity mainly in the Middle Cerebral Artery. It is commonly used to monitor patients with stroke risk by detecting micro-emboli through High Intensity Transient Signals (HITS). Many researchers have used signal processing techniques such as Fourier and Wavelet transforms [4] [5] [6], machine learning techniques such as SVMs [7] and deep learning techniques such as Convolutional Neural Networks (CNNs) to detect and classify HITS between Artifacts (Art.) and Emboli. However it remains difficult to differentiate between Gaseous Emboli (GE) and Solid Emboli (SE).

In this study we propose a semi-supervised learning method to annotate a dataset of micro-embolic TCD data with reduced requirements of manual annotation. Similar semi-automatic annotation methods have been proposed in other context by Benato et al. [8] and Zhu and Ghahramani [9]. The first team proposed to use of an Auto-Encoder (AE) to extract features from the original data and then project this data into a 2D space to manually annotate some samples and automatically annotate a part of the remaining samples using machine learning classifiers. The second team proposed a K-nearest-neighbor (KNN) approach to propagate the labels from a few labeled samples to the rest of the unlabeled samples.

In this paper, we propose a semi-automatic annotation method for TCD data based on feature extraction, dimensionality reduction (DR), manual annotation and label propagation. The proposed method is fast and accurate. To our knowledge, this is the first work to achieve semi-automatic TCD data annotation using machine and deep learning techniques.

II. Proposed method

The method that we use for semi-automatic data annotation is composed of four steps that are going to be detailed in the following subsections. Furthermore, the proposed method relies on three assumptions: the **structure assumption** (i.e. samples that are in the same structure, cluster or manifold, are likely to have the same labels) [10], the **preservation of the local structure** during projection and the **annotation space**

coverage (i.e. the labeled samples should cover the whole 2D annotation space).

A. Feature extraction

The first step of our method consists in extracting data specific features from the images representing the spectrograms of the HITS using a convolutional AE. The architecture that we use to do this is shown in figure 1. The AE has two parts: the *encoder* that encodes the input into a latent feature space, and the *decoder* that uses the latent representation to reconstruct the original input.

As our objective is to annotate data thanks to the learned representations of the AE (and not to reconstruct images using the latent space), we use all the available data for training in order to improve the learned representations of the *encoder*.

B. Dimensionality Reduction

The second step of our method consists in reducing the dimension of the latent space obtained by the *encoder* of the AE. Indeed, even if the chosen AE architecture allows to considerably reduce the dimension of the original images, it remains too high for visualization and annotation purposes. To, tackle this problem, we follow the steps of Benato et al. [8] an use the *t-SNE* algorithm [11] to project the *encoder* latent feature space into a 2D space. Benato et al. showed that, *t-SNE* is preferred over *global* DR techniques such as *UMAP* or *PCA* because the preservation of the local structure of the samples during projection is important for label propagation.

As in the previous step and for the same reasons, all the samples are going to be used for training in this step.

C. Manual annotation

The third step of our method consists in manually annotating a part of the available samples using the 2D space obtained in the previous step and the overall data information. The idea is to manually annotate data so that the labeled samples cover the whole 2D space (i.e. avoid wide areas without labels).

Moreover, to do the annotation the expert has access to: image of the HITS, audible (Doppler) signal of the HITS, pathology of the subject, use of contrast agent, source hospital and hospital service (all this information is not always available for all the subjects). The expert annotator will use this information to decide the *membership score* of a sample to a certain class. The final label of a HITS corresponds to the class having the highest *membership score*.

Furthermore, at the end of the step we dispose of a dataset $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, where $\mathcal{L}$ corresponds to the labeled samples (C classes) and $\mathcal{U}$ to the unlabeled samples ($|U| >> |\mathcal{L}|$, where $|.|$ is the cardinal operator).

D. Label propagation

The fourth step consists in propagating the labels from the labeled samples, $\mathcal{L}$, to some (or all) of the unlabeled samples, $\mathcal{U}$.

Let's denote $\mathcal{N}_K(s)$ the K-neighborhood of the 2D representation of sample $s \in \mathcal{D}$. Let's also denote $\mathcal{C}(s) = [p_1^s, p_2^s, ..., p_C^s]$ the *membership scores* of a sample $s \in \mathcal{D}$, where $\forall i \in \{1, ..., C\}, s \in \mathcal{D}, p_i^s$ can be interpreted as the probability that sample s belong to the class C_i. To propagate the labels from the labeled samples to unlabeled samples, we proceed using algorithm 1.

Algorithm 1: K-nearest-neighbor label propagation

```
Input: $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, $K$
Output: Set of newly labeled samples $\tilde{\mathcal{L}}$
Initialization: $\tilde{\mathcal{L}} = \emptyset$
Iterations:
while exist samples to label do
    $\mathcal{A} = \emptyset$;
    for $s \in \mathcal{L} \cup \tilde{\mathcal{L}}$ do
        $\mathcal{V}_K(s) = \mathcal{N}_K(s) \cap \mathcal{U}$;
        for $u \in \mathcal{V}_K(s)$ do
            $\mathcal{C}(u) = \mathcal{C}(s)$;
        end
        $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{V}_K(s)$;
        $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{A}$
    end
    $\tilde{\mathcal{L}} \leftarrow \tilde{\mathcal{L}} \cup \mathcal{A}$
end
```

The algorithm finishes when: (1) all the samples are labeled or (2) the only remaining unlabeled samples are the ones that are not in the K-neighborhood of any labeled samples.

The main difference between our method and the one proposed by Benato et al. [8] is that we use a *KNN* strategy to propagate the labels from the labeled samples to the unlabeled samples whereas they use *Laplacian SVM* and *Optimum Path Forest*.

E. Data description

TCD recordings were performed on 39 subjects of 11 different centers using an Atys Medical TCD Robotized Holter (TCD-X, probe frequency of 1.5 MHz) allowing recordings between 30 and 180 minutes. The spectrograms were computed from the TCD signals and High Intensity Transient Signals (HITS) were detected (9 dB threshold) resulting in 1545 extracted HITS, each of duration 250 ms. Finally, the spectrograms of each HITS were transformed into images used to train the AE for feature extraction.

III. Results

To evaluate our semi-automatic data annotation method, we proceed as follows. First, we apply the first three steps of our method in order to get a dataset $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ where $|\mathcal{L}| = 0.1 \times |\mathcal{D}|$. Then, we apply the fourth step to propagate the labels from the $\mathcal{L}$ samples to $\mathcal{U}$ samples (50 repetitions). If we note, $\mathcal{G}$ the set of correctly labeled samples, we can measure the performances of the method using the *annotation error* ϵ and the *percentage of labeled samples* p:

$$\epsilon = 1 - \frac{|\mathcal{G}|}{|\tilde{\mathcal{L}}|} \text{ and } p = \frac{|\tilde{\mathcal{L}}|}{|\mathcal{U}|} \tag{1}$$

Figure 2 shows ϵ and p with respect to K. First, we can see that p increases with the value of K to converge to 100% of labeled samples, and this, with $\epsilon \leq 15\%$. Secondly, we can

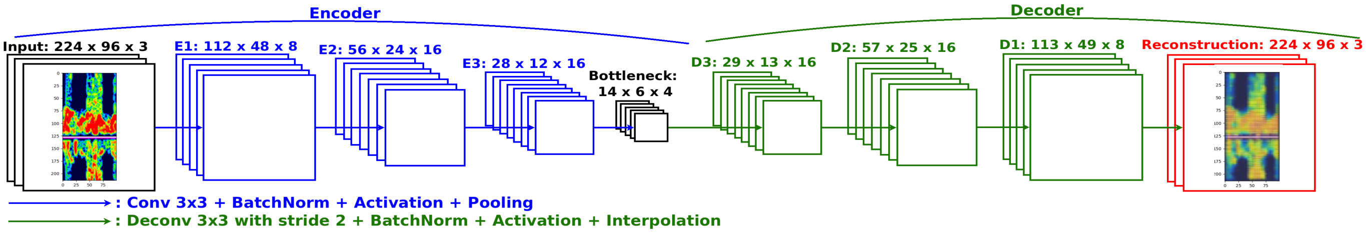

Fig. 1: Autoencoder architecture used in the unsupervised feature extraction step of our proposed method (section II.B)

notice that ϵ increases with the value of K (for $K \leq 10$), we have (in average) $\epsilon \leq 12.5\%$. Finally, we can identify two regimes: the *dynamic* regime where for small values of K, p increases; and the *permanent* regime where, for higher values of K, p converges to 100%.

Moreover, we measure the time needed by our method and by an expert to annotate one sample. Our method was able to annotate one sample in 44.9 ± 3.9 µs with $\epsilon = 11.92 \pm 2.08\%$ against around 8 s for a human annotator (we suppose that the annotator does not make errors). This represents a gain in time of a factor of 10^5.

IV. Discussion

Our method was able to annotate much faster (factor of 10^5) the same amount of data as a human expert at the expense of a higher annotation error. Even though no human annotator can achieve the speed annotation of our method, a more precise measurement of the human annotation speed should be done using different expert annotators.

Furthermore, we showed that there is a negative correlation between the annotation error and the percentage of labeled samples. There is then a trade-off between the annotation error that one can tolerate and the percentage of labeled samples needed in the database. In our case, as the percentage of annotated samples converges to 100% and the annotation error increases with K we focus on $K \leq 10$, with $K = 10$ being the first value of K allowing to annotate all the available samples. Moreover, a good compromise between annotation error and percentage of labeled samples is obtained for $K = 5$ where more than 90% of the samples are labeled with an annotation error smaller than 11% (c.f. figures 2 and 3).

Finally, annotation errors are inevitable with our proposed method, and it can disrupt classifiers trained on this type of data (c.f. figure 3). Most of the annotation errors are at the boundaries of two clusters of different classes, especially at the interface between the solid emboli and gaseous emboli. A good solution to reduce the negative impact of wrongly labeled samples is the use of *robust loss functions* such as *Generalized Cross-Entropy Loss* [12] to compensate the noise in the labels.

V. Conclusion

This work proposes a semi-automatic annotation method for TCD micro-embolic data. Our proposed method is composed of four steps: feature extraction, dimensionality reduction, manual annotation of a small amount of data and automatic label propagation. This pipeline allows to decrease considerably the annotation time (gain in time of a factor of 10^5) while keeping the annotation error smaller than 12%. We plan to apply our method to a larger TCD dataset and use the obtained dataset for a classification task. Additionally, to reduce the annotation error, we plan to use local quality projection metrics in order to identify the good candidates to propagate the labels.

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 Universite 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] Serena J, Segura T, Castellanos M, Dávalos A: Microembolic Signal Monitoring in Hemispheric Acute Ischaemic Stroke: A Prospective Study. Cerebrovasc Dis 2000;10:278-282. doi: 10.1159/000016070

[4] Markus HS, Punter M. Can transcranial Doppler discriminate between solid and gaseous microemboli? Assessment of a dual-frequency transducer system. Stroke. 2005 Aug;36(8):1731-4. doi: 10.1161/01.STR.0000173399.20127.b3. Epub 2005 Jul 14. PMID: 16020767.

[5] M. Gençer, G. Bilgin, and N. Aydin, Embolic Doppler ultrasound signal detection via fractional Fourier transform, 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.3050-3053, 2013. DOI : 10.1109/EMBC.2013.6610184

[6] Serbes G, Aydin N. Denoising performance of modified dual-tree complex wavelet transform for processing quadrature embolic Doppler signals. Med Biol Eng Comput. 2014 Jan;52(1):29-43. doi: 10.1007/s11517-013-1114-x. Epub 2013 Sep 19. PMID: 24048958.

[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] Bárbara C. Benato, Jancarlo F. Gomes, Alexandru C. Telea, Alexandre X. Falcão, Semi-automatic data annotation guided by feature space projection, Pattern Recognition, Volume 109, 2021, 107612, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2020.107612.

[9] X.Zhu and Z.Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report,Carnegie Mellon University,2002

[10] O. Chapelle, B. Scholkopf and A. Zien, Eds., "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]," in IEEE Transactions on Neural Networks, vol. 20, no. 3, pp. 542-542, March 2009, doi: 10.1109/TNN.2009.2015974.

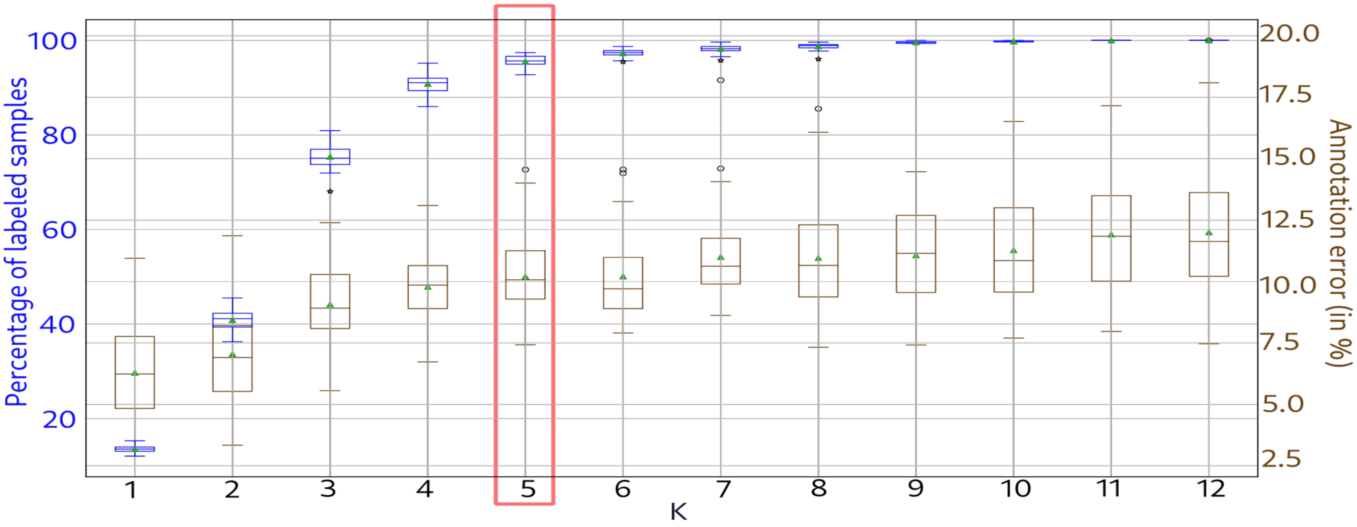

Fig. 2: Label propagation metrics with respect to the neighborhood size considered for label propagation. The blue curve represents the percentage of unlabeled samples that were annotated (left axis) by our method. The brown curve represents the annotation error in % (right axis). As the percentage of annotated samples converges to 100% and the annotation error increases with K, we focus on $K \leq 12$. There is a trade-off between the percentage of labeled samples and the annotation error. For $K = 5$ we have a good trade-off with an annotation error of $10.20 \pm 1.63\%$ and a percentage of labeled samples of $95.52 \pm 1.23\%$.

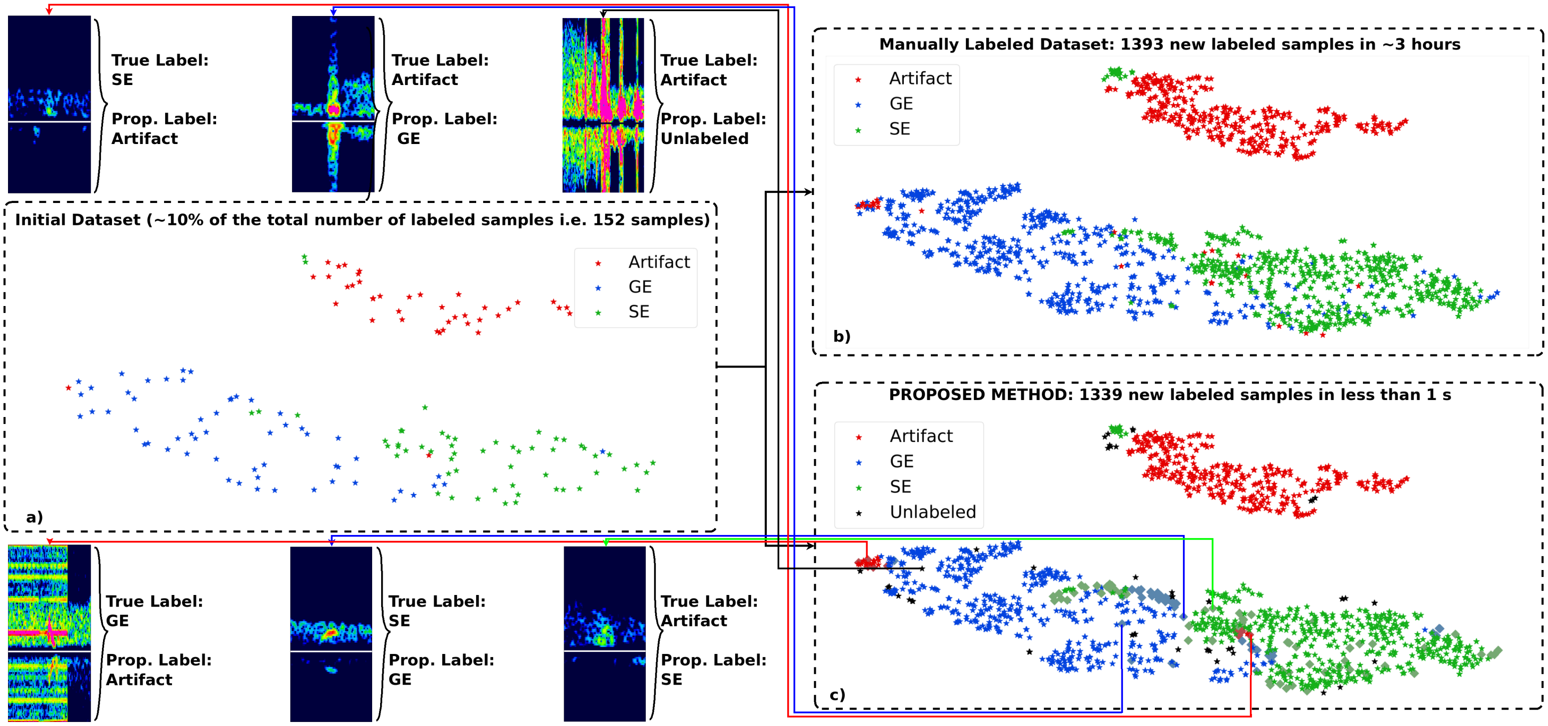

Fig. 3: Label propagation using $K = 5$. **a)** Initial dataset (10% of the total number of labeled samples). **b)** Manually labeled dataset (1545 samples) obtained by manually annotating the remaining unlabeled samples. **c)** Automatically labeled dataset (1491 samples) obtained using the initial dataset and our proposed method (annotation error of 9.11%, 96.12% of labeled samples and 3.88% of unlabeled samples). GE and SE stands for *Gaseous Emboli* and *Solid Emboli* respectively. The diamonds correspond to the wrongly labeled samples. The spectrograms of wrongly labeled and unlabeled samples are given as an example. Wrongly labeled samples are usually in the boundaries between clusters of samples of different classes.

[11] Van Der Maaten, L. Hinton, G. (2008). Visualizing Data using t-SNE . Journal of Machine Learning Research, 9, 2579–2605.

[12] Zhang, Z., Sabuncu, M. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. ArXiv, abs/1805.07836.