LABELING EEG COMPONENTS WITH A BAG OF WAVEFORMS FROM LEARNED DICTIONARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalograms (EEGs) are useful for analyzing brain activity, and spatiotemporal patterns in the EEG signal have clinical value, serving for example as biomarkers of diseases such as epilepsy. EEGs are a combination of components from multiple sources within the brain, the electrical activity of muscles, including the heart, and artifacts due to movement and external signals (e.g, line noise). Separating and classifying the sources of these components is important for analyzing the brain patterns in the EEG data. We propose *bag-of-waves* (BoWav), a new feature for the classification of EEG independent components (ICs). BoWav represents the IC time series through the distribution of counts of waveforms from a learned shift-invariant dictionary based reconstruction. We found that BoWav has a promising predictive performance, outperforming the state-of-the-art method for IC classification, ICLabel, in two of three classes of interest.

1 INTRODUCTION AND RELATED WORKS

EEG time series data are noisy due to various artifacts in the signal such as muscle twitching and eye blinking. Separating out these artifacts along with decomposing the remainder into constituent brain sources can be accomplished using independent component analysis (ICA) Hyvärinen & Oja (2000), which can extract statistically independent components (ICs) from EEG time series. For use in clinical settings, these components still need to be manually examined by experts to determine their provenance. Several works have proposed the use of spatial, spectral, and temporal features to automatically classify ICs, like the difference in average activation between the frontal and posterior areas of the brain to detect eye blinks Mognon et al. (2011), the power in standard EEG spectral bands like the α-band Winkler et al. (2011); Frølich et al. (2015), and summary statistics like kurtosis, differential entropy, and amplitude range Sai et al. (2018); Mognon et al. (2011); Frølich et al. (2015); Winkler et al. (2011). Recently, Pion-Tonachini et al. (2019) proposed ICLabel to automatically classify component signals into seven different categories such as brain, eye artifact, noise, muscle artifact, power line noise, etc. IClabel relies on second-order statistics of each IC's time series, namely, the power spectral density (PSD) estimate and the auto-correlation sequence, along with the topographical scalp map obtained from ICA over the known electrode placement. These approaches are typically feature-engineered and require expert domain knowledge. They also ignore the higher-order statistics of the time series, like the morphology and occurrence rate of recurrent waveforms. Additionally, in some scenarios such as electrocorticographic (ECoG) data, the spatial data differs between subjects due to different electrode placement, limiting the applicability of current approaches. The hypothesis is that the waveform morphology evident in the time series is crucial to expert analysis Cole & Voytek (2017), and relying on second-order statistics is insufficient for distinguishing different types of components with similar spectra. Inspired by the bag-of-words model from the natural language processing (NLP) literature Eisenstein (2019), our approach, *bag-of-waves* (BoWav), *learns* a dictionary of waveforms representative of each IC class, and then uses as features for classification the counts of the waveforms that best reconstruct windows of the time series. We train two linear classifiers, one using BoWav, and the other using the PSD and autocorrelation sequence, and compared their performance with ICLabel on the three physiological signals of the seven IC classes that ICLabel defines. Our results on IC classification show that BoWav provides a feature representative for times series that has both predictive power and interpretability.

2 METHODS AND DATA

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{N_{\text{exp}}}, y_{N_{\text{exp}}}), (\mathbf{x}_{N_{\text{exp}}+1}, y_{N_{\text{exp}}+1}), \ldots, (\mathbf{x}_N, y_N)\}$ denote a dataset of N instances, with N_{exp} expert-annotated ICs, $\mathbf{x}_i \in \mathbb{R}^{n_i}$ the i-th IC time series of length n_i, with label $y_i \in \mathcal{Y}_{\text{exp}}$ if $1 \leq i \leq N_{\text{exp}}$, and $y_i \in \mathcal{Y}$ if $N_{\text{exp}} < i \leq N$, for $i \in [N]$, and $\mathcal{Y}_{\text{exp}} \subset \mathcal{Y}$. We evaluate BoWav in the classification of ICs into the seven IC classes defined by ICLabel Pion-Tonachini et al. (2019): $\mathcal{Y} = \{\texttt{brain}, \texttt{muscle}, \texttt{eye}, \texttt{heart}, \texttt{line noise}, \texttt{channel noise}, \texttt{other}\}$. We use a dataset of multichannel EEG recordings from 34 subjects, sampled at 256 Hz, with expert labels for three of the seven IC classes: $\mathcal{Y}_{\text{exp}} = \{\texttt{brain}, \texttt{muscle}, \texttt{eye}\}$ Onton & Makeig (2022). There are in total $N = 7255$ ICs, and $N_{\text{exp}} = 1162(16\%)$ of those have expert labels. Since the comparison of a BoWav-based classifier with ICLabel is only meaningful in the three classes where expert labels are available, we need to train our classifier (see below) in the full set of seven IC classes that ICLabel was trained for, to have a fair comparison. We thus follow a weakly-supervised learning approach Zhou (2018) and annotate the rest of ICs using the EEGLAB Delorme & Makeig (2004) plugin of ICLabel, which might introduce label noise to the training, as some ICs could still be labeled as `brain`, `muscle`, or `eye`.

Shift-invariant k-means Let q be the number of ICs on a given class, and $\check{\mathbf{X}}_i = [\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_{M_i}]$ be a matrix whose columns are the M_i non-overlapping windows of length L extracted from the IC time series $\mathbf{x}_i$. For each class, we randomly sample the data across ICs and time, by taking a subset of the ICs that belong to a class, and a subset of all possible non-overlapping windows from each chosen IC. We thus obtain the set $\check{\mathcal{X}} = \{\check{\mathbf{x}}_{\pi_j}\}_{j=1}^{\check{N}}$, with $\pi_j \in \pi$, and π being a random permutation of $[\check{N}_{\text{tot}}]$, with $\check{N}_{\text{tot}} = \sum_{i=1}^{q} M_i$ being the total number of non-overlapping windows available on a given IC class. We then use a shift invariant k-means algorithm Mendoza-Cardenas & Brockmeier (2021), abbreviated here as *sikmeans*, to learn $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k] \in \mathbb{R}^{P \times k}$, a class-specific codebook of k centroids of length P, with $L > P$. We solve

$$\min_{[\mathbf{c}_1,\ldots,\mathbf{c}_k] \in \mathbb{R}^{P\times k}} \frac{1}{\check{N}} \sum_{i=1}^{\check{N}} \min_{\substack{\tau \in \{0,\ldots,L-P\} \\ \nu \in \{1,\ldots,k\} \\ a \in \mathbb{R}_{\geq 0}}} \frac{1}{2} \|T_P(\check{\mathbf{x}}_i, \tau) - a\mathbf{c}_\nu\|_2^2, \tag{1}$$

where $T_P(\check{\mathbf{x}}_i, \tau)$ is the P-length window extracted from $\check{\mathbf{x}}_i$ at time shift τ, with $\tau \in \{0, \ldots, L-P\}$. This optimization problem is solved by iteratively updating the codebook and the shift-invariant cluster assignments in an alternating fashion, similar to the classical k-means algorithm. First, with a fixed codebook, we find the centroid closest to each non-ovelapping window, i.e., the best tuple $\{\tau, \nu, a\}$. Then, each centroid in the codebook is updated by averaging the signals in its cluster, after extracting the P-length windows at the best τ found in the previous step. In the end, centroids in $\mathbf{C}$ are an approximation of recurrent waveforms that are representative of the IC time series.

For a given non-overlapping representation of an IC, $\check{\mathbf{X}}$, let $g : \mathbb{R}^{L\times M} \times \mathbb{R}^{P\times k} \to [k]^M$ be a shift-invariant cluster assignment operator such that $g(\check{\mathbf{X}}, \mathbf{C}) = \boldsymbol{\nu} = [\nu_1, \ldots, \nu_M]^{\text{T}}$, with $\nu_i \in [k]$ being the index of the centroid in $\mathbf{C}$ closest to the i-th window in $\check{\mathbf{X}}$, for $i \in [M]$.

Let $\mathcal{Y} \in \{1, 2, \ldots, J\}$ be the set of J IC class labels, and $\mathbf{C}_j \in \mathbf{R}^{P\times k}$ be a codebook of k representative waveforms of length P for class $j \in \mathcal{Y}$. The codebook $\mathbf{C}_j$ is learned by running sikmeans on non-overlapping windows of length L taken at random from a set of ICs in $\mathcal{D}_j$, where

$$\mathcal{D}_j = \begin{cases} \{(\mathbf{x}_i, y_i) : y_i = j, i = 1, \ldots, N_{\text{exp}}\}, & j \in \mathcal{Y}_{\text{exp}} \\ \{(\mathbf{x}_i, y_i) : y_i = j, i = N_{\text{exp}} + 1, \ldots, N\}, & j \in \mathcal{Y} \setminus \mathcal{Y}_{\text{exp}}. \end{cases} \tag{2}$$

In other words, the codebooks for classes with ground-truth labels are learned using only ICs with expert labels, and for the rest of the classes we use the ICs with noisy labels.

Bag-of-Waves The feature representation for BoWav is a vector, $\mathbf{z} = [\texttt{BoW}(g(\check{\mathbf{X}}, \mathbf{C}_1))^{\text{T}}, \ldots, \texttt{BoW}(g(\check{\mathbf{X}}, \mathbf{C}_J))^{\text{T}}]^{\text{T}} = [\texttt{BoW}(\boldsymbol{\nu}_1)^{\text{T}}, \ldots, \texttt{BoW}(\boldsymbol{\nu}_J)^{\text{T}}]^{\text{T}}$, with $\texttt{BoW} : [k]^M \to \{0, \ldots, M\}^k$ denoting the bag-of-waves mapping such that $\texttt{BoW}(\boldsymbol{\nu}) = \mathbf{z}' = [z'_1, \ldots, z'_k]^{\text{T}}$ are the counts of the labels in $\boldsymbol{\nu}$, with $z'_i = \sum_{j=1}^{M} [\![\nu_j = i]\!]$, for $i \in [k]$, and $[\![\texttt{statement}]\!] = 1$ when `statement` is

true, and 0 otherwise. Finally, we scale the BoWav features by their inverse document-frequency (idf) Robertson (2004) to downweight overly frequent waveforms.

PSD and autocorrelation We also consider the feature vector $\mathbf{z} = [P_{\mathbf{x}}^{\mathrm{T}},\ R_{\mathbf{xx}}^{\mathrm{T}}]^{\mathrm{T}}$, with $P_{\mathbf{x}}$ and $R_{\mathbf{xx}}$ being the PSD and autocorrelation of an IC, respectively. These two time-frequency features are computed in the same way as in ICLabel Pion-Tonachini et al. (2019).

Classifier Let $h_{\boldsymbol{\Theta}} : \mathcal{Z} \to [0,1]^J$ be a multi-class classifier with parameters $\boldsymbol{\Theta} \in \mathbb{R}^{d \times J}$ such that $h_{\boldsymbol{\Theta}}(\mathbf{z}) = \mathbf{p} = [p_1,\ p_2,\ \ldots, p_J]^{\mathrm{T}}$ is a vector of class probabilities, with $\mathbf{z} \in \mathcal{Z}$ being the feature vector extracted from the IC time series $\mathbf{x}$, $p_j = \mathbb{P}(Y = j \mid \mathbf{z})$, and $Y \in \mathcal{Y} = \{1, 2, \ldots, J\}$ denoting a random variable corresponding to the class label. Let $\mathcal{L}_i(\boldsymbol{\Theta})$ be the cross-entropy loss of $h_{\boldsymbol{\Theta}}$ on the i-th sample, $\mathbf{z}_i$. We train $h_{\boldsymbol{\Theta}}$ by finding the $\boldsymbol{\Theta}$ that minimizes

$$\mathcal{L}(\boldsymbol{\Theta}) = w_{\text{exp}} \sum_{i=1}^{N_{\text{exp}}} b_i \mathcal{L}_i(\boldsymbol{\Theta}) + \sum_{i=N_{\text{exp}}+1}^{N} b_i \mathcal{L}_i(\boldsymbol{\Theta}) + \frac{1}{2}\alpha \|\text{vec}(\boldsymbol{\Theta})\|_2^2 + \beta \|\text{vec}(\boldsymbol{\Theta})\|_1, \tag{3}$$

with $w_{\text{exp}} > 0$ being a sample weight used to emphasize samples with ground-truth labels, $\text{vec}(\boldsymbol{\Theta})$ denoting the result of flattening the matrix $\boldsymbol{\Theta}$ into a vector, and $\|\cdot\|_2$ and $\|\cdot\|_1$ denoting the l_2 and l_1 norms, respectively, with corresponding $\alpha, \beta > 0$ scalar factors that determine the regularization strength. In order to account for class imbalance in the data set, we scale each sample loss by a class weight, b_i, that is inversely proportional to the frequency of y_i, the class the sample belongs to. Formally, $b_i = \frac{N}{\sum_{l=1}^{N} [\![y_l = y_i]\!] J}$.

3 Experiments and Results

We split the data into two disjoint sets of 27 and 7 subjects for training and test, respectively, which corresponds to an approximate 80:20 ratio of the number of ICs. To extract the BoWav feature vector, we first need to learn a codebook of waveforms from each of the seven IC classes. We ran sikmeans on windows of length $L = 384$ (1.5 seconds) to a learn a codebook of $k = 128$ centroids with length $P = 256$ (1 second), using random initialization, and taking the best of three runs in terms of the sum of distances of all the windows to their closest centroid. In order to learn the codebooks in a reasonable amount of time, we sampled the training set across ICs and across time: we used a maximum of two ICs per subject and 50 minutes per IC, totaling 45 hours worth of training in most IC classes. Once all the class-specific codebooks are learned, we extract the BoWav features by performing shift-invariant cluster assignment on 1.5-second non-overlapping windows of the IC time series, using all the ICs avaliable in the training set, and taking 50 minutes per IC. For the feature vector composed of the PSD and autocorrelation, we used all the training data available.

We used scikit-learn Pedregosa et al. (2011) to train two multinomial logistic regression classifiers (see Eq. 3), one using the BoWav feature, and the other one using the concatenation of the PSD and autocorrelation sequence. The regularization parameters `C` and `l1_ratio` from scikit-learn are related to α and β as $\alpha = (1 - \texttt{l1_ratio})/\texttt{C}$ and $\beta = \texttt{l1_ratio}/\texttt{C}$. We split the training data using a leave-one-subject-out cross-validation, and compute balanced accuracy (our validation metric, see App. A) only on the ICs of the held-out subject that have ground-truth labels. To find the best parameters for each classifier, we search among the following set of candidates: `C`= {0.001, 0.01, 0.1, 1, 10, 100, 1000}, `l1_ratio`= {0, 0.2, 0.4, 0.6, 0.8, 1}, and $w_{\text{exp}} = \{1, 2, 4, 8, 16, 32\}$.

Figure 1 shows the confusion matrices of ICLabel and the two linear classifiers trained with BoWav and PSD+autocorrelation, evaluated on the test set with expert-annotated data. The rows of these matirices are normalized to sum to 1, and thus the numbers in the main diagonal correspond to the recall of the classifier on each class. Similarly, Table 1 shows the precision, recall, and F1 scores of each classifier, as well as the support (number of ICs) of each class. The BoWav classifier outperforms both baselines (ICLabel and the PSD+autocorrelation classifier) on the `brain` class, with a significant increase in recall. Furthermore, the BoWav classifier also has a significantly better performance in the `muscle` class compared to ICLabel.

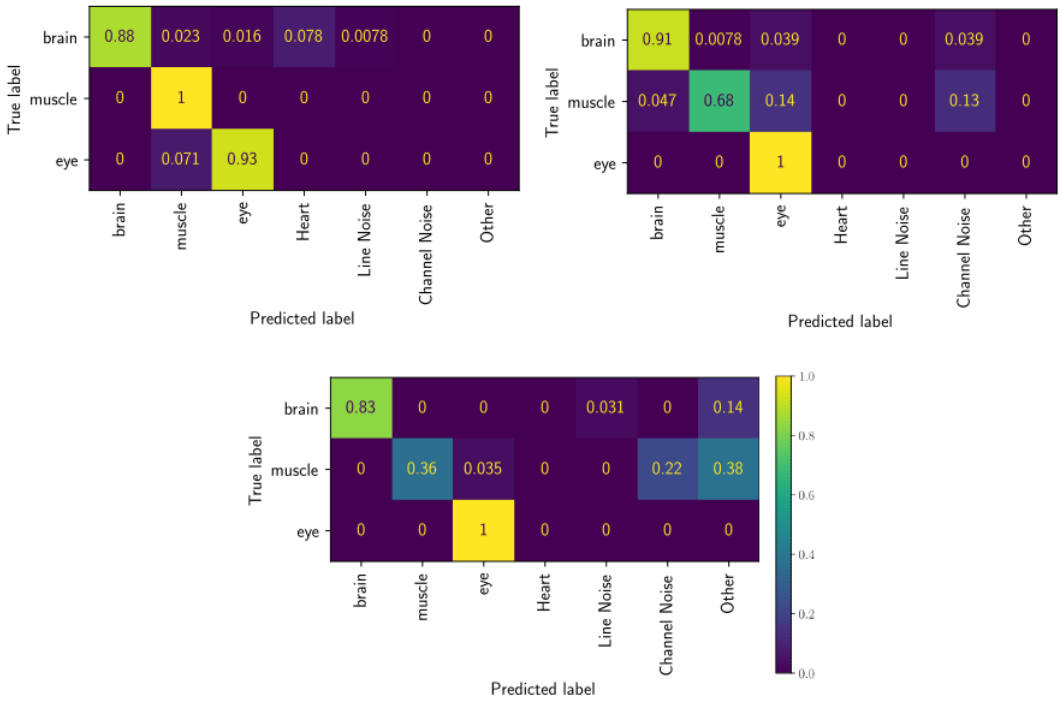

Figure 1: Left: Confusion matrix of the PSD and autocorrelation feature vector computed on data annotated by experts. Right: Confusion matrix of the BoWav feature vector. Bottom: Confusion matrix of IClabel.

Table 1: Classification results

	BoWav			PSD+autocorr			ICLabel			
	prec.	**recall**	**f1**	**prec.**	**recall**	**f1**	**prec.**	**recall**	**f1**	**supp.**
brain	0.97	0.91	**0.94**	1.00	0.88	0.93	1.00	0.83	0.91	128
muscle	0.98	0.68	0.81	0.96	1.00	**0.98**	1.00	0.36	0.53	85
eye	0.45	1.00	0.62	0.87	0.93	**0.90**	0.82	1.00	**0.90**	14

4 DISCUSSION

We proposed a new feature, the bag-of-waves (BoWav), for the classification of EEG independent components. Using a simple linear classifier, BoWav outperformed two strong baselines in the brain-related IC class, which is a very interesting and promising result. The codebook of waveforms for the `brain` class was learned from only 27 subjects, from which the BoWav classifier was able to generalize to 7 unseen subjects. Moreover, brain activity has much more variability and diversity in waveform morphology in comparison to the other six artifact-related classes. In addition to its good predictive power, BoWav is highly interpretable, as the codebooks learned have waveforms that are representative of each condition, like the QRS complex Malmivuo & Plonsey (2012) in the `heart` codebook, and the typical low-frequency spikes due to eye blinks in the `eye` codebook (see App. B).

There are multiple directions in which our work could be extended or improved. First, BoWav could be used to provide a more fine-grained classification of brain-related ICs, e.g., classifying brain ICs into subclasses like α-rhythm IC, or mu-rhythm IC. Second, given the performance delivered by the PSD and autocorrelation, it would be interesting to test the performance of a classifier trained with the concatenation of BoWav and those two other features. Finally, the problem of learning codebooks of representative waveforms from long time series remains a challenging and open problem, and research efforts in that direction will have an important impact in the improvement of downstream tasks, like the classification of time series through a bag-of-waves or, in general, an NLP-based approach.

REFERENCES

Scott R. Cole and Bradley Voytek. Brain Oscillations and the Importance of Waveform Shape. *Trends Cogn. Sci.*, 21(2):137–149, 2017. ISSN 1879307X. doi: 10.1016/j.tics.2016.12.008.

Arnaud Delorme and Scott Makeig. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*, 134(1):9–21, 2004. ISSN 01650270. doi: 10.1016/j.jneumeth.2003.10.009.

Jacob Eisenstein. *Introduction to natural language processing*. MIT press, 2019.

Laura Frølich, Tobias S. Andersen, and Morten Mørup. Classification of independent components of EEG into multiple artifact classes. *Psychophysiology*, 52(1):32–45, 2015. ISSN 00485772. doi: 10.1111/psyp.12290. URL https://onlinelibrary.wiley.com/doi/10.1111/psyp.12290.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, jun 2000. ISSN 08936080. doi: 10.1016/S0893-6080(00)00026-5.

Jaakko Malmivuo and Robert Plonsey. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, 2012. ISBN 9780199847839. doi: 10.1093/acprof:oso/9780195058239.001.0001.

Carlos H. Mendoza-Cardenas and Austin J. Brockmeier. Shift-invariant waveform learning on epileptic ECoG. In *43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 4, aug 2021. URL http://arxiv.org/abs/2108.03177.

Andrea Mognon, Jorge Jovicich, Lorenzo Bruzzone, and Marco Buiatti. ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48 (2):229–240, feb 2011. ISSN 00485772. doi: 10.1111/j.1469-8986.2010.01061.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8986.2010.01061.x.

Julie Onton and Scott Makeig. *Imagined Emotion Study*. OpenNeuro, 2022. doi: 10.18112/openneuro.ds003004.v1.1.1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011. ISSN 15324435.

Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage*, 198:181–197, 2019. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.05.026.

Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. *J. Doc.*, 60(5):503–520, 2004. ISSN 00220418. doi: 10.1108/00220410410560582.

Chong Yeh Sai, Norrima Mokhtar, Hamzah Arof, Paul Cumming, and Masahiro Iwahashi. Automated classification and removal of EEG artifacts with SVM and wavelet-ICA. *IEEE J. Biomed. Heal. Informatics*, 22(3):664–670, 2018. ISSN 21682208. doi: 10.1109/JBHI.2017.2723420.

Irene Winkler, Stefan Haufe, and Michael Tangermann. Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. *Behav. Brain Funct.*, 7(1):1–15, 2011. ISSN 17449081. doi: 10.1186/1744-9081-7-30.

Zhi Hua Zhou. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.*, 5(1):44–53, 2018. ISSN 2053714X. doi: 10.1093/nsr/nwx106.

A BALANCED ACCURACY

The validation metric that we use is balanced accuracy, defined as

$$\texttt{balanced-accuracy}(y, \hat{y}, w_{\text{exp}}) = \frac{1}{\sum_j \hat{w}_j} \sum_i [\![\hat{y}_i = y_i]\!] \hat{w}_i, \tag{4}$$

with y_i and $\hat{y}_i$ denoting the true and predicted labels for the i-th sample, for $i \in \{1, \ldots, N_{\text{exp}}, N_{\text{exp}} + 1, N\}$, and

$$\hat{w}_i = \begin{cases} \frac{w_{\text{exp}}}{\sum_{j=1}^{N_{\text{exp}}} [\![y_j = y_i]\!] w_{\text{exp}} + \sum_{j=N_{\text{exp}}+1}^{N} [\![y_j = y_i]\!]} & i = 1, \ldots, N_{\text{exp}} \\ \frac{1}{\sum_j [\![y_j = y_i]\!]} & i = N_{\text{exp}} + 1, \ldots, N, \end{cases} \tag{5}$$

being a scalar that combines the expert and class weights to account for the mixed label source and the class imbalance in the dataset.

B CLASS-SPECIFIC CODEBOOKS

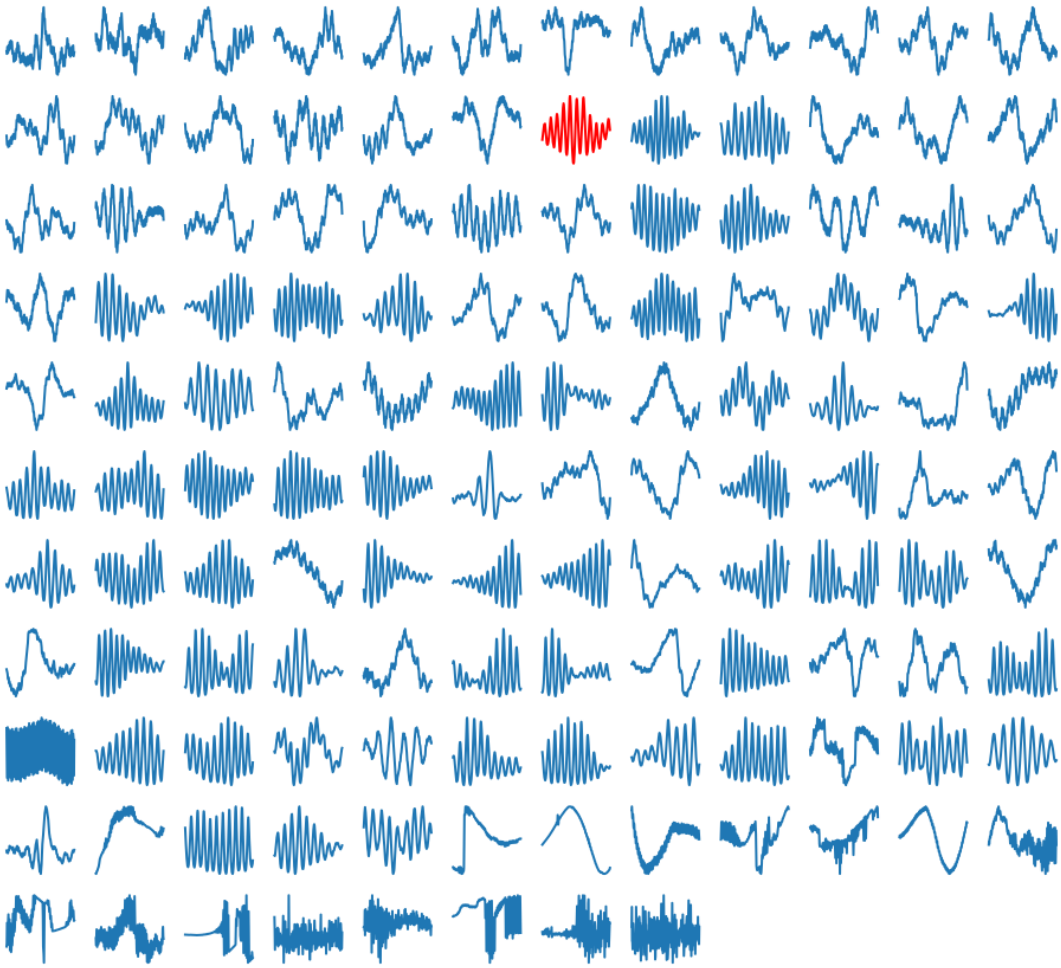

Figure 2: Codebook of brain centroids, ordered from left to right and top to bottom in descending order of their cluster size. In red, a waveform with a strong α-rhythm at 10 Hz.

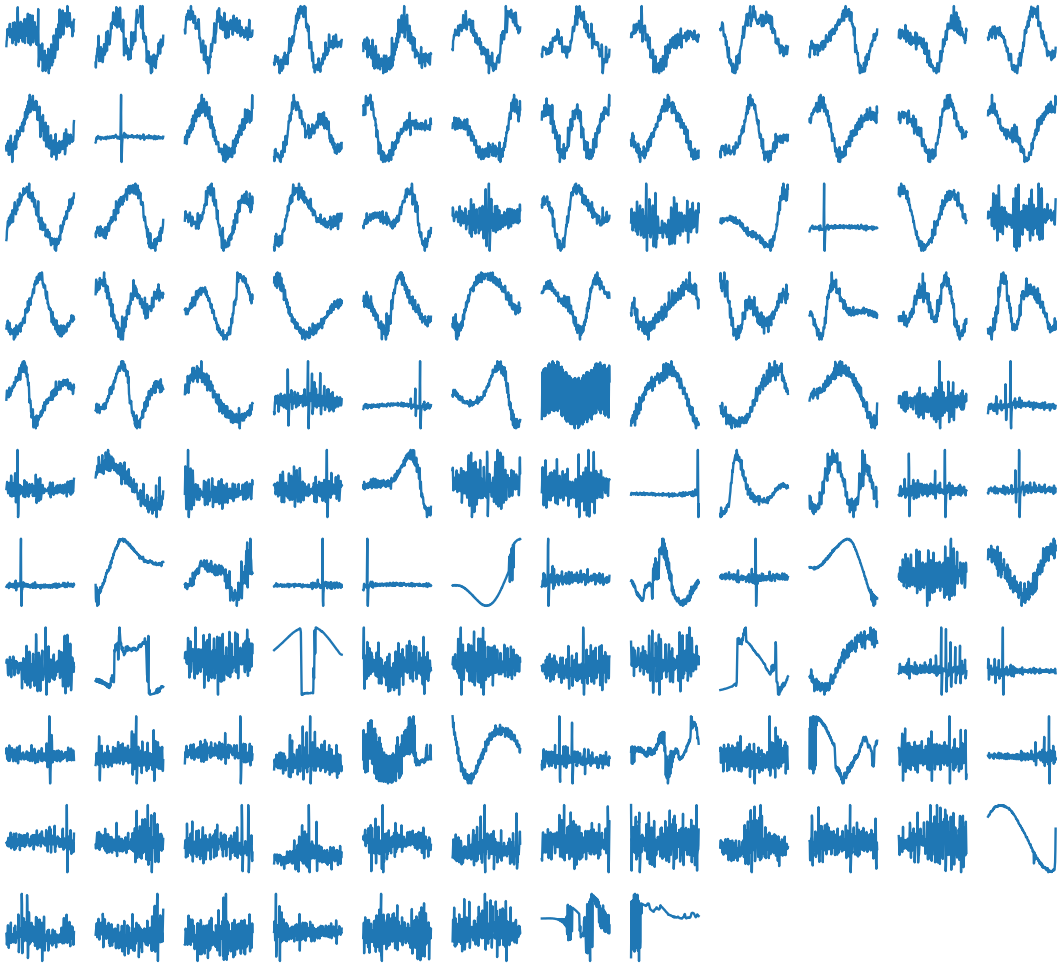

Figure 3: Codebook of muscle centroids.

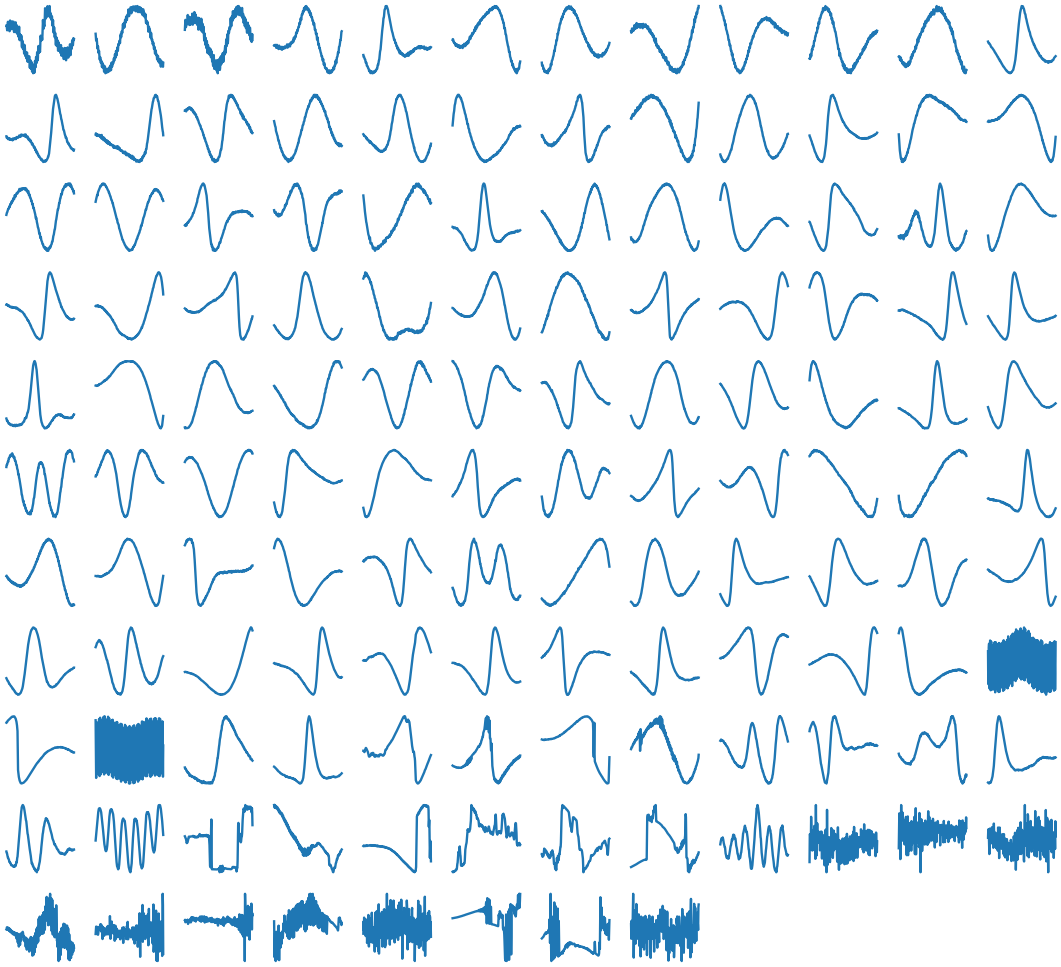

Figure 4: Codebook of eye centroids.

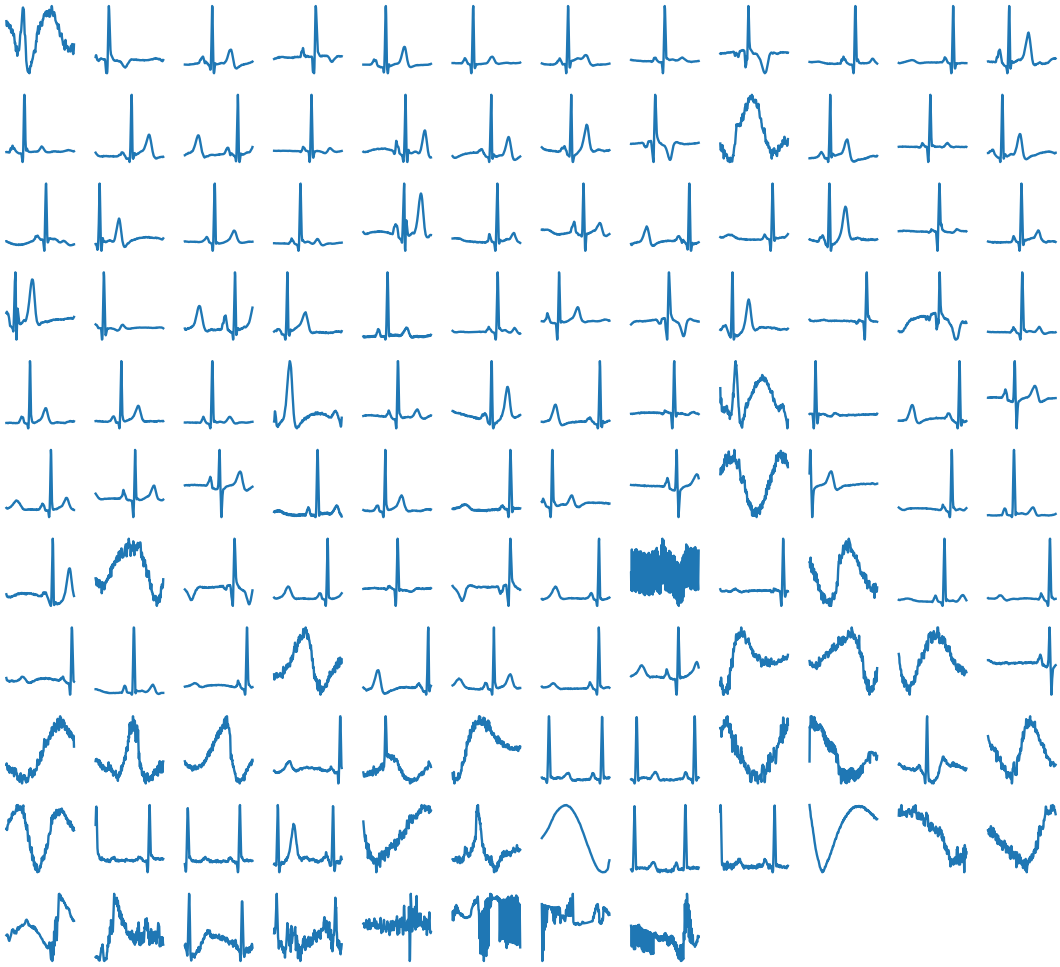

Figure 5: Codebook of heart centroids.

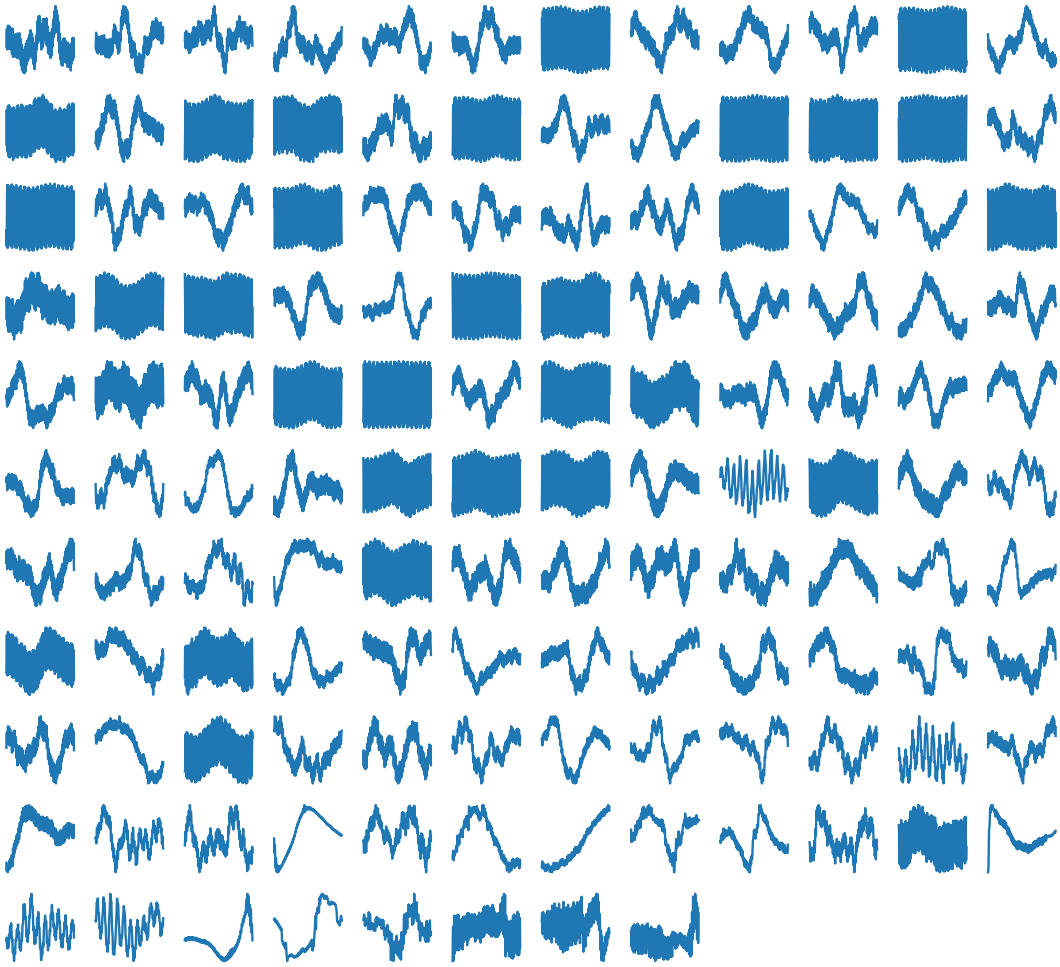

Figure 6: Codebook of line noise centroids.

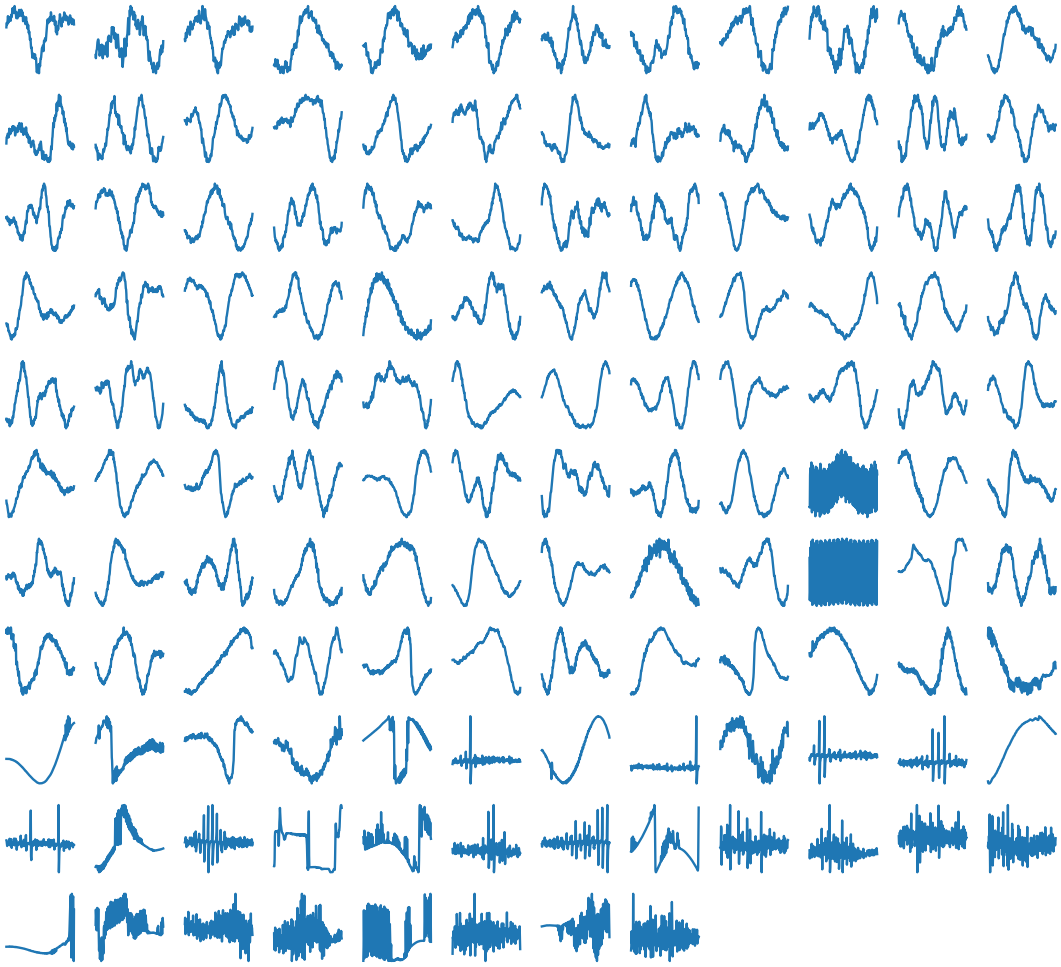

Figure 7: Codebook of channel noise centroids.

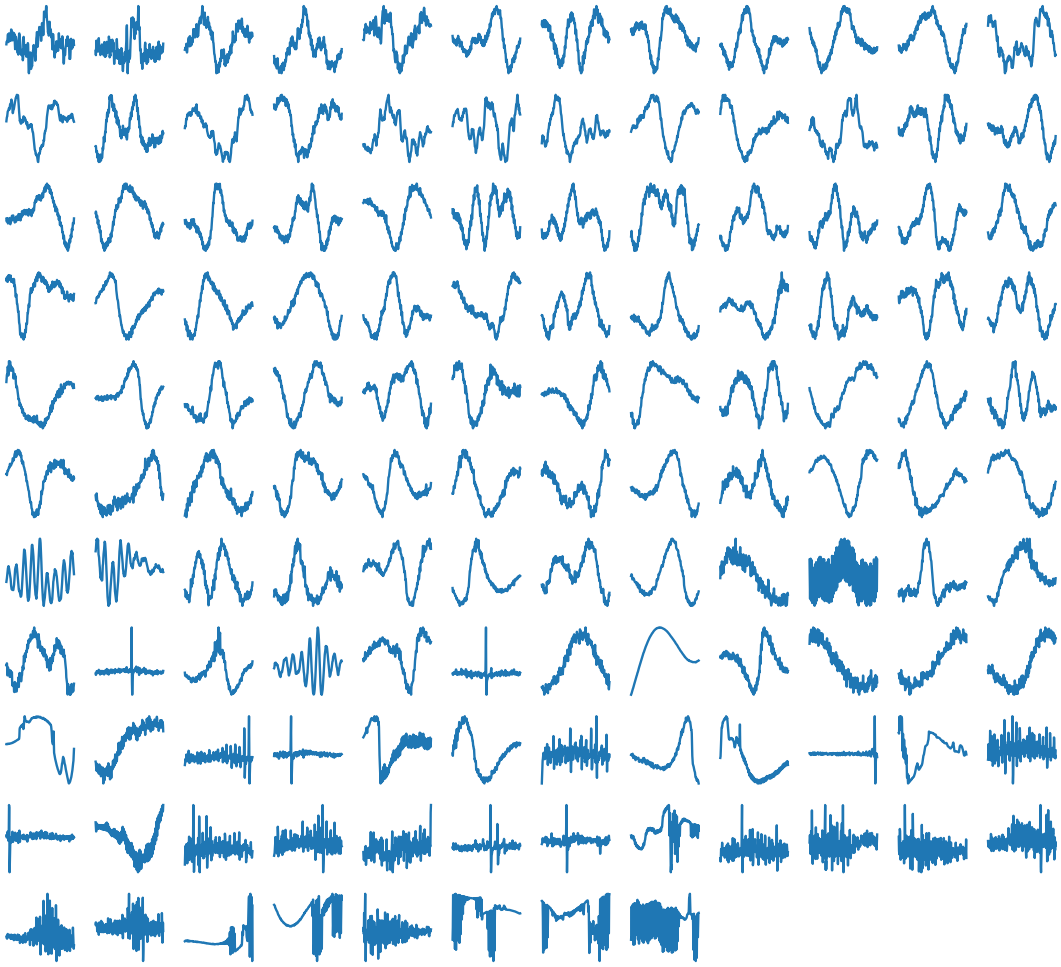

Figure 8: Codebook of ”other” centroids.