Knowledge Distillation-based Channel Reduction for Wearable EEG Applications

Velu Prabhakar Kumaravel [1], Una Pale [1], Tomas Teijeiro [1], Elisabetta Farella [1], and David Atienza Alonso [1]

[1]Affiliation not available

October 30, 2023

Abstract

Wearable EEG applications demand an optimal trade-off between performance and system power consumption. However, high-performing models usually require many features for training and inference, leading to a high computational and memory budget. In this paper, we present a novel knowledge distillation methodology to reduce the number of EEG channels (and therefore, the associated features) without compromising on performance. We aim to distill information from a model trained using all channels (teacher) to a model using a reduced set of channels (student). To this end, we first pre-train the state-of-the-art model on features extracted from all channels. Then, we train a naive model on features extracted from a few task-specific channels using the soft labels predicted by the teacher model. As a result, the student model with a reduced set of features learns to mimic the teacher via soft labels. We evaluate this methodology on two publicly available datasets: CHB-MIT for epileptic seizure detection and BCI competition IV-2a dataset for motor-imagery classification. Results show that the proposed channel reduction methodology improves the precision of the seizure detection task by about 8% and the motor-imagery classification accuracy by about 3.6%. Given these consistent results, we conclude that the proposed framework facilitates future lightweight wearable EEG systems without any degradation in performance.

Knowledge Distillation-based Channel Reduction for Wearable EEG Applications

Velu Prabhakar Kumaravel, Una Pale, Tomás Teijeiro, Elisabetta Farella, and David Atienza

***Abstract*—Wearable EEG applications demand an optimal trade-off between performance and system power consumption. However, high-performing models usually require many features for training and inference, leading to a high computational and memory budget. In this paper, we present a novel knowledge distillation methodology to reduce the number of EEG channels (and therefore, the associated features) without compromising on performance. We aim to distill information from a model trained using all channels (teacher) to a model using a reduced set of channels (student). To this end, we first pre-train the state-of-the-art model on features extracted from all channels. Then, we train a naive model on features extracted from a few task-specific channels using the soft labels predicted by the teacher model. As a result, the student model with a reduced set of features learns to mimic the teacher via soft labels. We evaluate this methodology on two publicly available datasets: CHB-MIT for epileptic seizure detection and BCI competition IV-2a dataset for motor-imagery classification. Results show that the proposed channel reduction methodology improves the precision of the seizure detection task by about 8% and the motor-imagery classification accuracy by about 3.6%. Given these consistent results, we conclude that the proposed framework facilitates future lightweight wearable EEG systems without any degradation in performance.**

***Index Terms*—EEG, Channel Reduction, Knowledge Distillation, Machine Learning, Seizure Detection, Motor Imagery.**

I. Introduction

Electroencephalography (EEG) is a non-invasive neuroimaging technique for investigating brain function and pathology [1]. EEG measures the brain's electrical activities via electrodes placed on the scalp. As EEG offers high temporal resolution (in the order of ms), it helps precisely detect the onset of abnormal electrical activities. EEG is used in various clinical and non-clinical applications, including epilepsy monitoring, Brain-Computer Interface (BCI) based rehabilitative technologies, and cognitive studies for neuroscientific research.

Artificial Intelligence (AI) has made breakthroughs in several health applications in the past two decades, including the EEG domain [2], [3]. The increasing availability of datasets facilitates Machine Learning (ML) algorithms to successfully identify the salient (hidden) features crucial to solve a given task. Such algorithms significantly reduce the burden on human experts who processes data by visual analysis or statistical signal processing tools. However, the best-performing AI algorithms are often cumbersome, presenting deployment challenges as they typically demand a large computational and memory budget. This drawback hinders progress, especially in realizing wearable EEG designs that can monitor patients continuously outside the clinical environment. Despite the progress made in current hardware design strategies to address this challenge, we are still far behind the goal [4].

Moreover, learning classification models relies on data labeled by a human expert. This poses two practical limitations: 1) Manual labeling of EEG data is time-consuming; 2) For some applications, it is not possible to achieve consensus among the experts. For example, EEG technicians often find it challenging to define the exact start/end of an epileptic seizure event. Such uncertain labeling can explain why intelligent models occasionally fail to discriminate between the classes or detect unlabeled seizures [5]. Further, such crisp labels in clinical EEG data (i.e., hard targets) do not reflect real-life events as the switch between classes is unnaturally abrupt (sudden transition from positive to negative class or vice versa). Therefore, to improve the performance of the ML models, it is worth investigating alternative ways of labeling the data.

In this work, we aim to address both challenges entirely from the software perspective by leveraging the Knowledge Distillation (KD) framework [6], [7], which is traditionally a model compression strategy. KD refers to transferring knowledge from a complex high-performing model (teacher) to a smaller one (student) without any significant loss in performance. There are several approaches to distilling knowledge from teacher to student in the literature (see [8] for a survey on knowledge distillation approaches). Here, we employ the offline distillation strategy in which we first train the teacher model using features extracted from all EEG channels; then, we train the student model using features extracted from a reduced set of channels with the soft labels (or predicted probabilities) estimated by the pre-trained teacher model. Using the proposed methodology, first, we reduce the computational burden by decreasing the number of EEG channels (and the associated feature vectors) without compromising on performance. Second, we aim to resolve the abovementioned labeling issue by replacing the original crisp labels with soft labels (which provide more realistic transitioning between classes).

This work has been partially supported by the ML-Edge Swiss National Science Foundation (NSF) Research project (GA No. 200020182009/1) and the PEDESITE Swiss NSF Sinergia project (GA No. SCRSII5 193813/1).

V.P. Kumaravel is with Energy-Efficient Embedded Digital Architectures (E3DA) Unit, Fondazione Bruno Kessler, Italy, and Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy (e-mail: vkumaravel@fbk.eu).

U. Pale is with Embedded Systems Laboratory (ESL), École Polytechnique Fédérale de Lausanne, Switzerland.

T. Teijeiro is with Basque Center for Applied Mathematics (BCAM), Spain.

E. Farella is with Energy-Efficient Embedded Digital Architectures (E3DA) Unit, Fondazione Bruno Kessler, Italy.

D. Atienza is with Embedded Systems Laboratory (ESL), École Polytechnique Fédérale de Lausanne, Switzerland.

To prove the generalization of the proposed channel reduction strategy, we validate the approach on two publicly available EEG datasets, each collected for different applications: 1) CHB-MIT Corpus labeled for epileptic seizures [9]; 2) BCI Competition IV-2a labeled for motor-imagery movements [10]. Our experiments prove that the KD-based channel reduction strategy is successful in both test cases. Precisely, our analysis reveals that it is possible to obtain similar (or sometimes even better) performance using the proposed methodology despite a significant reduction of input data channels/features.

Our contributions are summarized as follows: i) To the best of our knowledge, we propose an EEG channel reduction methodology using the teacher-student framework for the first time; ii) Unlike existing channel reduction/selection approaches that are suitable only for specific applications, our proposed method can be applied to any multi-channel EEG dataset irrespective of the application domain; iii) We demonstrate that student models can perform better than the baseline models (without knowledge distillation) in both applications, with/without channel reduction.

The rest of the article is organized as follows: Section II provides a brief overview of the existing channel selection methods for both considered applications, namely, seizure detection and motor-imagery classification; Section III describes the proposed methodology using knowledge distillation framework; Section IV describes the experimental procedure to validate the proposed method; Section V presents the obtained results and Section VI provides a comprehensive discussion of the proposed work; finally, Section VII concludes the paper.

II. Related Work

Conventional scalp EEG devices have different channels for acquiring signals from different brain regions. In most applications, selecting a subset of channels in which task- or pathology-relevant features are present would be beneficial [11]. As the computational complexity of the algorithms increases as a function of the number of channels, optimal channel selection helps realize low-power systems with faster response rates (or inference time) [12]. Further, in some applications, channel selection improves the performance of the system as we exclude redundant channel information from processing [13]. From the end-user perspective, reducing the number of channels would improve the comfort level and reduce the setup time [12].

Most of the existing EEG channel reduction/selection approaches for seizure detection fall into either of these three categories: i) In the first category, channels are ranked based on certain features (e.g., channel variance) and the top-most channels are chosen for further processing [14], [15]; ii) In the second category, different combinations of channels are tested, and the best combination of channels which improves the performance is chosen [16], [17]; iii) The third category is called as the recursive channel selection by backward elimination or forward selection that aims at estimating which channels are most helpful to discriminate the classes of interest. The goal is to find the smallest number of channels, such that the average classification performance is at least as good as the performance obtained using all channels. [18].

The primary drawback of the first category is that it is application-specific, as the chosen features are usually specific to a particular application domain. Notably, care must be taken to select the appropriate feature(s) - which typically requires extensive analysis. Further, if the selected features are sensitive to EEG noise or artifacts, the resultant chosen channels might represent noise more than neural information. The second approach is exhaustive, as numerous combinations of channels should be evaluated. Since the evaluation is usually based on the classifier's performance, the selected channels might be optimal only for the evaluated model, and retraining is required for a new model. The third approach is exhaustive as well since, in both forward selection and backward elimination, the model is trained and validated every time a channel is added or eliminated in each iteration. Also, in this case, the evaluation criterion is based on the classifier's performance. In sum, the above-mentioned approaches are application/classifier-dependent and computationally expensive.

Regarding the motor-imagery BCI classification task, in most studies, channels were chosen manually based on domain knowledge. For example, channels C3, C4, and Cz are considered important as they are located over the motor cortex [10] and relevant for the motor-imagery task. The most widely used automated techniques are based on Common Spatial Pattern (CSP; [19]) and variants [20], [21]. CSP uses spatial filters that lead to new time series whose variances are optimal to discriminate between two classes. CSP-based approaches are computationally efficient, yet, they do not achieve satisfactory performance [22]. Alternate solutions include the Sequential Floating Forward Selection (SFFS) algorithm [23] and Support Vector Machine-Recursive Feature Elimination (SVM-RFE), which come under the third category of channel selection algorithms discussed above. As stated, such algorithms are both time-consuming and classifier dependent.

This study aims to develop a generic channel reduction methodology that can be applied to any EEG application domain. Thereby, we target removing the manual and computational efforts required to identify and analyze domain-specific features. Further, we focus on improving the performance using only a few channels, irrespective of the classifier being utilized. Thus, we propose an application/classifier-independent EEG channel reduction method based on the Knowledge Distillation (KD) framework [7]. To the best of our knowledge, this is the first time that the KD-based approach has been employed for EEG channel reduction.

III. Knowledge Distillation Framework

Knowledge Distillation (KD) is a model compression technique used to transfer knowledge from a highly complex teacher model to a lightweight student model. In KD, the student learns to mimic the teacher model by utilizing the embedded knowledge to achieve similar or even better performance. Such knowledge comes from the output class probabilities or soft labels estimated by the teacher models.

As the computed soft labels have high entropy, they provide much more information per training class than hard targets

Fig. 1: An example scenario demonstrating the effectiveness of soft labels. The probabilities/soft labels assigned for all classes are reported next to the input sample/image.

could possibly provide. Further, the soft labels provide much less variance in the gradient between training classes. As a result, student models can be trained on much less data than teacher models without compromising on performance. Figure 1 shows a toy example demonstrating the usefulness of soft labels using samples from the MNIST dataset [24]. In this example, we consider four hand-written numbers (4-class problem), namely, 7, 2, 4, and 9. Any ML model trained for recognizing different classes predicts the unseen input samples with a probability assigned for each class. For example, the model predicts the input image "7" with a probability of 0.80, and since class "2" shows similar features as class "7", the model assigns a probability of 0.15. As can be inferred from these values, the soft labels provide lower variance (and higher entropy) between classes than the traditional one-hot encoded hard targets. Given this dark knowledge transferred from teacher models, student models achieve similar or better performance even with fewer feature sets.

In the EEG domain, cross-modal knowledge distillation strategy is successfully applied for various tasks such as emotion recognition, sleep scoring, and seizure detection, where knowledge from one data modality (e.g., EEG) is transferred to another (e.g., ECG) [3]. In this work, we use the KD framework to reduce the number of EEG channels - as such, we transfer the knowledge from many EEG channels in the source domain to the few channels in the target domain. Figure 2 explains the proposed KD-based EEG channel reduction methodology.

A. Pre-training Teacher Model

For a given dataset, we first train the state-of-the-art model using features extracted from many EEG channels. We employ the leave-one-out cross-validation (LOOCV) strategy, where all but one file are used for training, and the left-out file is used for testing [25]. The obtained predicted probabilities (i.e., soft labels) for each test file are stored in separate files for transferring the knowledge later. This process is depicted in Figure 2 (green panel). In this work, for the epileptic seizure and the motor-imagery BCI datasets, we used Random Forest (RF) and Linear Support Vector Machine Classifier (SVC) models as the teacher, respectively.

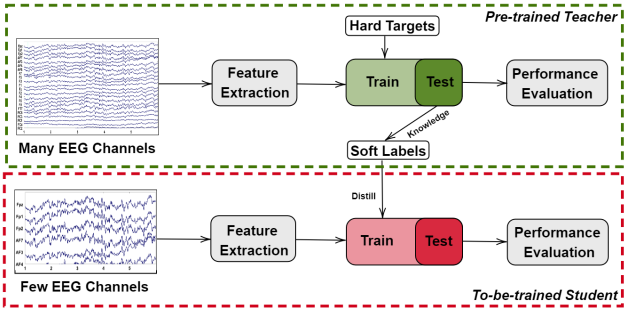


Fig. 2: Overview of the proposed KD-based channel reduction method.

B. Knowledge Distilled Student Model

Then, we train a naive model using features extracted from a few channels where we employ a similar cross-validation strategy (LOOCV) as before. The only difference is that this time, we train the model using soft labels obtained from the teacher. As a result, there is a knowledge distillation between the teacher and student after the pre-training. Precisely, the teacher model, which has learned from many EEG channels, teaches the student model that utilizes reduced channels. This process is depicted in Figure 2 (red panel). In this work, for the epileptic seizure and the motor-imagery BCI datasets, we used the eXtreme Gradient Boosting algorithm (XGBoost) and Linear Support Vector Machine Regressor (SVR) models as the student, respectively.

IV. Experimental Setup

This section describes the experiments performed to demonstrate the efficiency of the proposed KD-based EEG channel reduction in two publicly available datasets acquired for different applications.

A. Personalized Epileptic Seizure Detection

Epilepsy is one of the most common neurological diseases characterized by excessive hypersynchronous discharge of neurons in the brain (also known as seizures). Epileptic seizures are recurrent paroxysmal events characterized by stereotyped behavioral alterations reflecting the underlying neural mechanisms of the pathology [26]. EEG is the most common neuroimaging test to diagnose epilepsy by observing the significant deviation from the normal pattern of brain waves. A high-density EEG setup (i.e., with more than 64 electrodes) is useful for determining the affected areas of the brain. Once the regions of interest are identified for a given patient, a personalized low-density EEG design is possible, which supports continuous monitoring outside hospitals [12].

1) Data Description: We used CHB-MIT pediatric data, one of the most commonly used benchmarking datasets for seizure detection [9]. It contains data from 24 patients (aged 1.5 to 22 years). There are 183 seizures labeled in the data corpus, which makes around 7.6 ± 5.8 seizures per patient. We considered 18 channels that are common to all patients.

Several works in the literature assume a balanced class for seizure detection, which can lead to a non-realistic performance [27], [3]. To overcome this limitation, we performed similar data preparation in terms of class balance as done in [28]. For each patient, we created N number of files (where N represents the number of seizures), which contain one seizure epoch of length t (in seconds) and $10 * t$ seconds of non-seizure data randomly sampled from the same patient. Therefore, the class ratio for seizure and non-seizure is 1:10.

2) Feature Extraction: Before extracting the features, we applied a zero-phase, 4th order Butterworth band-pass filter between [1, 30] Hz. Then, we segmented the filtered data EEG data into windows of 4 seconds with an overlapping of 0.5 seconds. For each window and channel combination, we computed 19 features, namely Mean Amplitude, Line Length, and absolute and relative power values in different frequency-domain bands (such as Delta: [0.5-4] Hz, Theta: [4-8] Hz, Alpha: [8-12] Hz, and Beta: [12-30] Hz). We chose these features as they have shown high discriminative power for these datasets in our previous study [29].

3) Choice of Models: For the teacher model, we used an ensemble Random Forest (RF) classifier as it demonstrated the state-of-the-art performance in seizure detection [12]. As stated before, we used 19 features extracted from each channel (19 features x 18 channels = 342 features) to train and evaluate the RF model.

Since, in this work, the student model is trained using soft labels (i.e., continuous targets), our choices are limited as most classification algorithms do not support soft label training. For this reason, we chose the highly efficient eXtreme Gradient Boosting algorithm: XGBoost [30]. For reducing the channel size, we exploited the previous studies based on 2 bi-polar electrode configurations, also known as e-Glass setup (comprising of F7-T3, F8-T4 channels, see highlighted electrodes in Figure 3 for the location), demonstrating negligible performance loss compared to all electrodes [12]. As such, the student model is trained and evaluated only on a limited number of features (i.e., 19 features x 2 channels = 38 features), thereby reducing the input dimension up to 88%. Further, for comparison, we evaluated the student model when trained using hard labels (xGB Baseline), instead of soft labels.

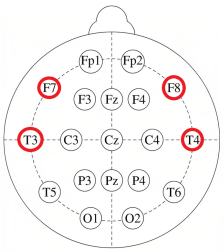


Fig. 3: Electrode montage configured in the 10-20 international system for the CHB-MIT dataset [9]. The highlighted e-Glass channels (in red) [12] are used for training the student model.

4) Cross-Validation Scheme: Since epileptic seizures can be very specific to each patient [5], we undertook a personalized approach (i.e., the training and test set for a given patient do not contain data from other patients). As stated in Section III-A, we performed the leave-one-out cross-validation (LOOCV) strategy for each patient. In other words, for each patient, one file was left out (for testing), and the remaining were used for training. The predicted probabilities (i.e., soft labels) of the test file were stored for training the student model later. A sample case is shown in Figure 4, in which the predicted soft labels provide richer information than the hard binary targets. Further, the soft labels show a realistic transition from non-seizure to seizure class and vice-versa.

5) Performance Evaluation: For a detailed evaluation, we considered three measures used by the authors in [27], namely Sensitivity (True Positive Rate or TPR), Precision (Positive Predictive Value or PPV), and F1 Score at two different levels: i) Episode, and ii) Duration. The Episode level detects seizure blocks (i.e., the start and end of the seizure event), also called as Block level. Even if the predicted seizure block does not cover the entire duration of the ground truth, the prediction is still considered a true positive. This metric is easy to interpret and is usually what the clinicians care about. On the other hand, the Duration level also considers the predicted duration within the detected episodes. More specifically, it corresponds to a standard sample-by-sample performance metric. In the end, another metric with a strong practical impact, especially in the clinical domain, that is used is the number of false positives (False Alarm Rate - FAR)/day.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

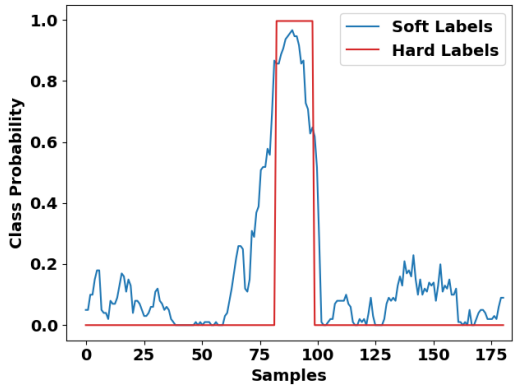
$$F1Score = \frac{2 * TPR * PPV}{TPR + PPV} \tag{3}$$

B. Personalized Motor Imagery BCI Classification

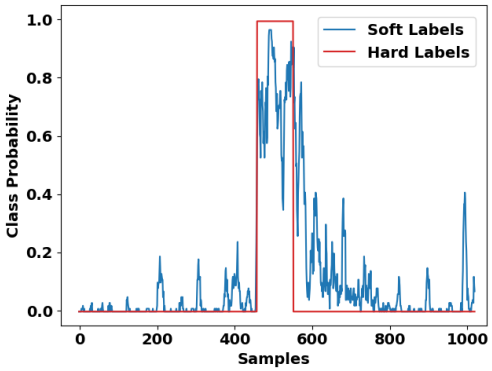
Traditional Brain-Computer Interface (BCI) systems aim to restore communication and control in severely paralyzed patients [31] and find its applications mainly for rehabilitation purposes. However, in recent years, BCI has found a wide range of applications for healthy people, as well [32]. Users prefer EEG-based non-invasive systems, due to their portability, safety, comfort, and relatively low cost. In such systems, multiple electrodes are placed on the scalp for acquiring the EEG signals. Since motor movements are specific to certain regions of the brain (e.g., motor cortex), motor-imagery BCI applications can benefit from channel reduction strategies.

1) Data Description: We used the BCI competition IV 2a dataset [10], which comprises EEG data from 9 subjects performing the cue-based BCI paradigm. The goal is to decode EEG signals in motor-sensory brain areas associated with imagined body movement. Precisely, the experimental tasks consisted of four different motor imagery tasks, namely the imagination of movement of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). Each subject participated in two sessions on different days. Each session is comprised of 6 runs separated by short breaks, and each run consists of 48 trials, yielding a total of 288 trials per session. For more details on the employed paradigm, see [10]). In this study, for simplicity, we considered only two classes (i.e., left hand and right hand), which are the most commonly used tasks in MI-BCI applications. Therefore, each subject has 72 trials per class for training and testing.

The activity of the brain was recorded using 22 EEG electrodes according to the 10-20 system (see Figure 5). It is



(a) Patient 02, Seizure 2.



(b) Patient 03, Seizure 4.

Fig. 4: Predicted soft vs. hard labels from the CHB-MIT dataset [9]. Class probability 0.0 indicates "Non-seizure", and 1.0 indicates "Seizure" classes. The transition between the two classes is gradual when soft labels are used for training.

bandpass filtered between 0.5 Hz and 100 Hz and sampled with 250 Hz. In addition to the 22 EEG channels, three Electrooculography (EOG) channels give information about eye movement. An expert marked the trials containing artifacts based on the EOG signal. This resulted in the removal of 9.41% of the trials from the dataset. However, the number of trials per class remains balanced [33].

2) Feature Extraction: Traditionally, Common Spatial Pattern (CSP) is employed as the feature extraction method for the BCI-MI dataset. CSP finds a set of spatial filters that transform the EEG data to be more discriminative in terms of variances. Since CSP suffers from the swelling effect in covariance matrix estimation [34], [35], Riemmannian geometry is utilized as it provides a more accurate approximation of the distance on smoothly curved spaces. Roughly speaking, Riemannian geometry studies smoothly curved spaces that locally behave like Euclidean spaces. Replacing Euclidean geometry with the Riemannian yields better performance in EEG-based computations, favoring a faster computational time for online, wearable applications [33], [35]. In this work, we utilized the Riemannian features introduced and validated in a previous study [33] as a fast and accurate inference algorithm for motor-imagery BCI classification.

3) Choice of Models: For the teacher model, we used the Linear Support Vector Machine Classifier (SVC) as used in the original implementation [33]. Instead of using all 22 electrodes, we used a subset of 12 electrodes (indexed between 7 and 18 in Figure 5) covering the Central, Temporal and Parietal regions of the brain, in which, the left-hand and right-hand motor imagery activations are most prominent [36], [37]. Fitting the time series data on Riemannian geometry space resulted in 3354 features.

For the student model, we chose the Linear Support Vector Machine Regressor (SVR) that supports training using continuous targets. Instead of using all 12 electrodes, here we reduced the number of electrodes to 6 (indexed as 7, 8, 14, 12, 13, 18 in Figure 5). This electrode combination resulted in 903 features after transforming data in Riemannian space.

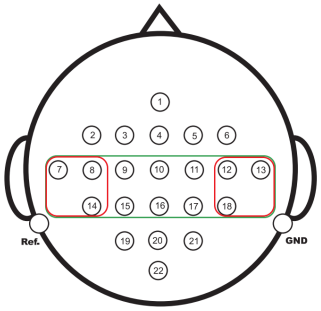


Fig. 5: Electrode montage configured in the 10-20 international system for the Motor Imagery BCI experiment [10]. In this study, we used 12 electrodes (in the green box) for training the teacher model and 6 electrodes (in red boxes) for training the student model. The choice of these electrodes comes from the domain knowledge [36].

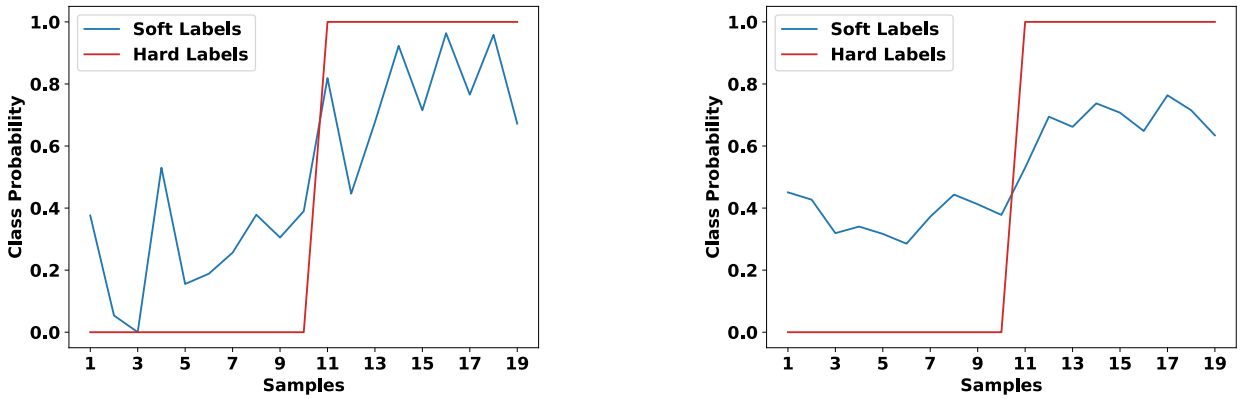
4) Cross-Validation Scheme: Our dataset contains 2 sessions for each subject as stated in Section IV-B1. Therefore, we performed a 2-fold cross-validation for each subject. In other words, for each subject, each fold contains different training and test files. The predicted probabilities (i.e. soft labels) of each test file were stored for training the student model later. The difference between soft and hard labels for a sample subject is shown in Figure 5.

5) Performance Evaluation: Since the number of samples in both classes (left-hand and right-hand) are same in all datasets, we considered the metric Accuracy, which measures the number of correctly detected positive classes over all samples and calculated as follows:

$$\text{Classification Accuracy} = \frac{N_{correct}}{N_{total}} \times 100\% \qquad (4)$$

V. Results

This section presents the results obtained in each of the considered datasets.



(a) Subject 01, Session 1.

(b) Subject 03, Session 1.

Fig. 6: Predicted soft vs target hard labels from the Motor-Imagery BCI Dataset [10]. Class probability 0.0 indicates "Left-hand", and 1.0 indicates "Right-hand" classes.

A. Seizure Detection

As mentioned in Section IV-A4, we performed the LOOCV for each patient and evaluated the detection performance in both episode and duration levels using measures defined in Section IV-A5. Figure 7 summarizes the overall performance at the Episode level and the False Alarm Rate (FAR/day) measured by combining both Episode and Duration levels. Instead, Table I presents the average performance achieved at the duration level.

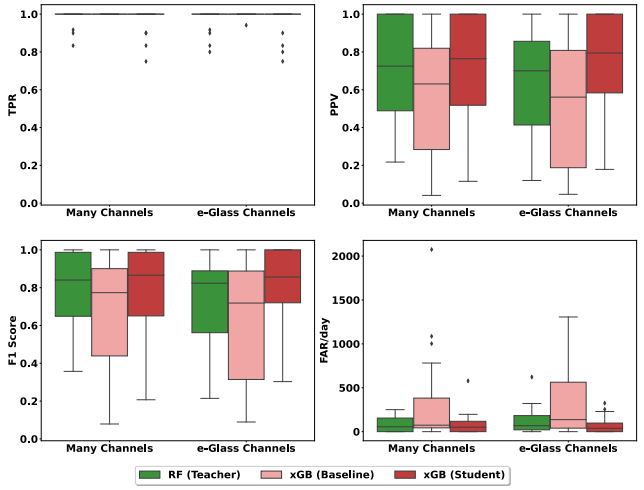


Fig. 7: Performance comparison of models using four metrics. The green, pastel red and red box plots represent the teacher, the uncalibrated naive, and the student models, respectively.

Channels\Models	RF (Teacher)			xGBoost (Baseline)			xGBoost (Student)		
	TPR	PPV	F1	TPR	PPV	F1	TPR	PPV	F1
$n = 18$	82.45	86.41	83.61	90.38	76.97	81.38	79.72	89.95	83.41
$n = 2$	79.76	83.27	80.51	87.73	76.01	79.49	74.42	90.37	80.22

TABLE I: Overall performance summary (duration level) of CHB-MIT database

When all 18 channels are considered and at the episode level (see Figure 7), it can be observed that KD reduces the True Positive Rate (TPR or Sensitivity). However, the average Positive Predictive Value (PPV) is improved from 70% to 72% resulting in an improvement in F1 Score by about 1% compared to the teacher model. This trend of decreased TPR but improved PPV is also observed in the "e-Glass Channels" scenario. As a result, the average FAR/day is drastically reduced from 125 to 79, considering all patients together using only two channels. In applications like epileptic seizure detection where precision (and FAR) is paramount, these results strongly favor our proposed KD-based channel reduction methodology.

In sum, we might arrive at two distinct conclusions: 1) Reducing the number of channels from 18 to 2 results in performance degradation (significantly, the PPV metric); 2) However, distilling knowledge from the teacher model (RF) consistently improves the naive model's baseline performance (highlighted in pastel red color in Figure 7) and achieves comparable (or better) performance as the teacher model, despite a significant channel reduction.

B. Motor Imagery BCI

As stated in Section IV-B4, we performed a two-fold cross-validation on the BCI competition IV-2a dataset. In each fold, the predicted probabilities (i.e., soft labels) were stored for training the student model later. Figure 8 shows the average performance summary for both "Many Channels" ($n = 12$) and "Reduced Channels" ($n = 6$). The green boxes in the figure represent the performance of the teacher model with 12 channels; the pastel red indicates the performance of the naive student model trained using six channels without knowledge distillation; the red indicates the performance of the student model after employing KD.

First, considering the "Many Channels" scenario, the Linear SVC model (teacher) achieves an average accuracy of 81%. Performance degradation of around 3% is observed when the same task is assigned to the uncalibrated, naive Linear SVR model (baseline). However, when the knowledge is distilled

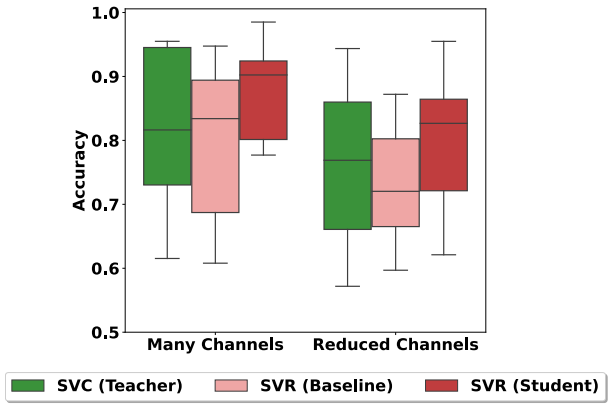


Fig. 8: Performance comparison of models. The green, pastel red and red box plots represent the teacher, the uncalibrated naive model, and the student models, respectively.

from the Linear SVC model, the same model improves accuracy by about 9% compared to the baseline SVR and 6% compared to the teacher model. As also observed in seizure detection dataset, even without channel reduction, KD already yields an improvement in accuracy. This is most likely because the soft labels provide richer information to discriminate the classes more precisely.

Channels\Models	Linear SVC (Teacher)	Linear SVR (Baseline)	Linear SVR (Student)
$n = 12$	81.38	78.75	87.04
$n = 6$	76.35	71.86	79.97

TABLE II: Overall performance summary (in terms of accuracy) of BCI Competition dataset.

After removing half of the channels ($n = 6$), the teacher SVC model demonstrates a reduction in performance accuracy of about 5% compared to the performance achieved in the "Many Channels" scenario. However, after knowledge distillation, the student SVR model achieves an average accuracy of 79.97% using only six channels, which is comparable to the accuracy obtained by the state-of-the-art teacher model (81.38%) using 12 channels (see Table II). This confirms the successful knowledge transfer via soft labels from the pretrained teacher to the student model.

VI. Discussion

Wearable EEG applications can provide real-time feedback on the significant pathological changes recorded in the brain's neural activity. Such applications usually operate in resource-constrained environments; therefore, an optimal trade-off between performance and energy consumption is desired. One of the ways to consume less energy is to develop computationally less-intensive yet reliable algorithms. Within the context of EEG, a straightforward solution is to reduce the dimension of input data (i.e., the number of channels). Fewer EEG channels reduce computational time, memory budget, system power consumption, and preparation time during electrode placement and equipment costs. Further, it can also reduce the overfitting risk that may occur when using irrelevant channels. However, selecting the optimal set of channels is crucial to avoid losing the task/pathology-specific information.

In this work, we proposed a novel channel reduction methodology based on the Knowledge Distillation (KD) framework without potential performance degradation. We showed that it is possible to achieve a similar (or even better) performance by training the lightweight models (i.e., with reduced input data dimension) using predicted probabilities obtained from the pretrained larger models (i.e., using features from many EEG channels). The predicted probabilities replaced hard (binary) labels and thus transferred granular knowledge from larger to simpler models. We validated the approach in two clinical EEG datasets to prove the generalization of the proposed method. Our results showed that for the CHB-MIT database, it is possible to reduce the overall false positives up to 37% after channel reduction compared to baseline models trained without distilling the knowledge. Likewise, around 3.5% improvement in accuracy using KD is observed in the BCI competition IV-2a dataset.

A primary drawback of the existing EEG channel reduction/selection methods is that they lack universality as they are typically developed and validated for a particular EEG application. For example, in [15], the authors selected channels with maximum variance. Since variance is one of the commonly used features for seizure detection tasks [28], the proposed strategy is successful. However, for motor-imagery BCI classification tasks, where CSP-based channel selection is widely employed [21], selecting channels based on variance might be ineffective. Thus, this work introduces a generic EEG channel reduction methodology using the KD framework, which is reliable for any EEG application domain as the underlying mechanism is based on a well-established probabilistic framework [6], [7]. However, one of the limitations of our approach is that domain-specific knowledge is required to choose significant channels. As done in this work, for the seizure detection task, we used the 2 frontal-temporal electrodes relevant for seizure detection [12], and we chose the most significant 6 central-parietal electrodes (3 on the left; 3 on the right) for the motor-imagery classification problem [36], [37]. As a future work, it might be possible to rank the channels using the probabilities for the positive class predicted by the larger (i.e., teacher) model for an automated channel selection for the student. Also, it is noteworthy that the performance of student models is always dependent on teacher models. The objective of this study is to demonstrate that, through the knowledge distillation process, the student models achieve a similar or slightly better performance compared to the teacher models, using limited data.

One of the major hindrances in machine learning-based solutions for EEG pathology detection or task classification is the limited amount of labeled data. This requires human experts to visually look at the data to label significant events, which can be overwhelmingly time-consuming. At the same time, proper care must be taken as the performance of intelligent models heavily relies on labeled datasets. However, the

experts within/across labs often find it challenging to have a common consensus on labeling the data. For example, agreeing on the start/end of a seizure event is difficult. Moreover, the traditional labeling of data (0 for negative and 1 for positive class) does not reflect the natural transition between positive and negative classes. Our proposed KD-based channel reduction approach addresses both of these challenges. First of all, a pretrained model can be used to label entirely unseen data (acquired using a similar EEG setup) as also done in other works [38], [39]. Secondly, we provide richer labeling for training new models by utilizing the predicted probabilities, which tend to show the realistic transition between positive and negative classes (Figure 4). Indeed, this justifies why naive models achieve better performance compared to the state-of-the-art models trained using hard targets despite significant channel reduction.

VII. Conclusion

Artificial Intelligence (AI) has been accelerating rapidly in the EEG domain in recent years. However, high-performing models demand a large computational and memory budget. This hinders the progress in bringing wearable EEG solutions to monitor patients continuously outside the clinical environment. A possible solution to overcome this limitation is to minimize the number of EEG channels without a significant drop in performance. In this work, we presented an EEG channel reduction methodology by leveraging the Knowledge Distillation (KD) framework. To the best of our knowledge, this is the first time KD is employed for reducing the number of EEG channels. To this end, we first trained the high-performing model on features extracted from all EEG channels (teacher). Then, we trained a naive model on features extracted from a few task-specific channels using the soft labels estimated by the pre-trained teacher model. The knowledge via softened labels helps the student to mimic the teacher, but with a reduced set of features. We considered two publicly available datasets for validation: i) CHB-MIT database with annotated epileptic seizures and ii) Motor-Imagery BCI Competition IV 2a dataset. In both datasets, KD-based channel reduction resulted in improved performance compared to the baseline models. Precisely, an improvement in precision of about 8% and accuracy of around 3% was observed in both datasets, respectively. Thus, this work showed that the performance is not compromised by distilling the knowledge from pretrained models, despite channel reduction. As EEG channel reduction improves portability and reduces computational complexity, the proposed approach is a promising strategy for future high-performance wearable EEG systems.

References

[1] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, 2005.

[2] F. Paissan, V. P. Kumaravel, and E. Farella, "Interpretable CNN for single-channel artifacts detection in raw EEG signals," in *2022 IEEE Sensors Applications Symposium (SAS)*. IEEE, Aug. 2022. [Online]. Available: https://doi.org/10.1109/sas54819.2022.9881381

[3] S. Baghersalimi, A. Amirshahi, F. Forooghifar, T. Teijeiro, A. Aminifar, and D. Atienza, "Many-to-one knowledge distillation of real-time epileptic seizure detection for low-power wearable internet of things systems," 2022. [Online]. Available: https://arxiv.org/abs/2208.00885

[4] E. D. Giovanni, F. Forooghifar, G. Surrel, T. Teijeiro, M. Peon, A. Aminifar, and D. A. Alonso, "Intelligent edge biomedical sensors in the internet of things (IoT) era," in *Emerging Computing: From Devices to Systems*. Springer Nature Singapore, Jul. 2022, pp. 407–433. [Online]. Available: https://doi.org/10.1007/978-981-16-7487-7_13

[5] D. Sopic, T. Teijeiro, D. Atienza, A. Aminifar, and P. Ryvlin, "Personalized seizure signature: An interpretable approach to false alarm reduction for long-term epileptic seizure detection," *Epilepsia*, Feb. 2022. [Online]. Available: https://doi.org/10.1111/epi.17176

[6] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Aug. 2006. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01453-z

[9] A. Shoeb and J. Guttag, "Application of machine learning to epileptic seizure detection," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 975–982.

[10] C. Brunner, R. Leeb, G. Muller-Putz, A. Schogl, and G. Pfurtscheller, "BCI competition 2008 - Graz data set IVa," 2008. [Online]. Available: http://bnci-horizon-2020.eu/database/data-sets

[11] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, Aug. 2015. [Online]. Available: https://doi.org/10.1186/s13634-015-0251-9

[12] D. Sopic, A. Aminifar, and D. Atienza, "e-glass: A wearable system for real-time detection of epileptic seizures," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5.

[13] H. Varsehi and S. M. P. Firoozabadi, "An EEG channel selection method for motor imagery based brain–computer interface and neurofeedback using granger causality," *Neural Networks*, vol. 133, pp. 193–206, Jan. 2021. [Online]. Available: https://doi.org/10.1016/j.neunet.2020.11.002

[14] M. R. Karimi and H. Kassiri, "A multi-feature nonlinear-SVM seizure detection algorithm with patient-specific channel selection and feature customization," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, Oct. 2020. [Online]. Available: https://doi.org/10.1109/iscas45731.2020.9180729

[15] A. A. E. Shoka, M. H. Alkinani, A. S. El-Sherbeny, A. El-Sayed, and M. M. Dessouky, "Automated seizure diagnosis system based on feature extraction and channel selection using EEG signals," *Brain Informatics*, vol. 8, no. 1, Feb. 2021. [Online]. Available: https://doi.org/10.1186/s40708-021-00123-7

[16] V. Shah, M. Golmohammadi, S. Ziyabari, E. V. Weltin, I. Obeid, and J. Picone, "Optimizing channel selection for seizure detection," in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, Dec. 2017. [Online]. Available: https://doi.org/10.1109/spmb.2017.8257019

[17] G. C. Jana, A. Tripathi, and A. Agrawal, "EEG channel selection approach for seizure detection based on integrated BPSO and ELM," in *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, Feb. 2020. [Online]. Available: https://doi.org/10.1109/spin48934.2020.9071106

[18] R. Jana and I. Mukherjee, "Deep learning based efficient epileptic seizure prediction with EEG channel optimization," *Biomedical Signal Processing and Control*, vol. 68, p. 102767, Jul. 2021. [Online]. Available: https://doi.org/10.1016/j.bspc.2021.102767

[19] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, Dec. 2000. [Online]. Available: https://doi.org/10.1109/86.895946

[20] J. K. Feng, J. Jin, I. Daly, J. Zhou, Y. Niu, X. Wang, and A. Cichocki, "An optimized channel selection method based on multifrequency CSP-rank for motor imagery-based BCI system," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–10, May 2019. [Online]. Available: https://doi.org/10.1155/2019/8068357

[21] S. Chen, Y. Sun, H. Wang, and Z. Pang, "Channel selection based similarity measurement for motor imagery classification,"

in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Dec. 2020. [Online]. Available: https://doi.org/10.1109/bibm49941.2020.9313336

[22] H. Zhang, X. Zhao, Z. Wu, B. Sun, and T. Li, "Motor imagery recognition with automatic EEG channel selection and deep learning," *Journal of Neural Engineering*, Nov. 2020. [Online]. Available: https://doi.org/10.1088/1741-2552/abca16

[23] M. Radman, A. Chaibakhsh, N. Nariman-zadeh, and H. He, "Generalized sequential forward selection method for channel selection in EEG signals for classification of left or right hand movement in BCI," in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, Oct. 2019. [Online]. Available: https://doi.org/10.1109/iccke48569.2019.8965159

[24] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[25] G. I. Webb, C. Sammut, C. Perlich, T. Horváth, S. Wrobel, K. B. Korb, W. S. Noble, C. Leslie, M. G. Lagoudakis, N. Quadrianto, W. L. Buntine, N. Quadrianto, W. L. Buntine, L. Getoor, G. Namata, L. Getoor, J. H. Xin Jin, J.-A. Ting, S. Vijayakumar, S. Schaal, and L. D. Raedt, "Leave-one-out cross-validation," in *Encyclopedia of Machine Learning*. Springer US, 2011, pp. 600–601. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_469

[26] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, "Epileptic seizures and epilepsy: Definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, Apr. 2005. [Online]. Available: https://doi.org/10.1111/j.0013-9580.2005.66104.x

[27] U. Pale, T. Teijeiro, and D. Atienza, "Systematic assessment of hyperdimensional computing for epileptic seizure detection," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 6361–6367.

[28] U. Pale, T. Teijeiro, and D. Atienza, "Exploration of hyperdimensional computing strategies for enhanced learning on epileptic seizure detection," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Jul. 2022. [Online]. Available: https://doi.org/10.1109/embc48229.2022.9870919

[29] U. Pale, T. Teijeiro, and D. Atienza, "Hyperdimensional computing encoding for feature selection on the use case of epileptic seizure detection," 2022. [Online]. Available: https://arxiv.org/abs/2205.07654

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[31] G. Pfurtscheller, "The hybrid BCI," *Frontiers in Neuroscience*, 2010. [Online]. Available: https://doi.org/10.3389/fnpro.2010.00003

[32] V. P. Kumaravel, V. Kartsch, S. Benatti, G. Vallortigara, E. Farella, and M. Buiatti, "Efficient artifact removal from low-density wearable EEG using artifacts subspace reconstruction," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 333–336.

[33] M. Hersche, T. Rellstab, P. D. Schiavone, L. Cavigelli, L. Benini, and A. Rahimi, "Fast and accurate multiclass inference for mi-bcis using large multiscale temporal and spectral features," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1690–1694.

[34] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, Jan. 2007. [Online]. Available: https://doi.org/10.1137/050637996

[35] S. Blum, N. S. J. Jacobsen, M. G. Bleichner, and S. Debener, "A riemannian modification of artifact subspace reconstruction for EEG artifact handling," *Frontiers in Human Neuroscience*, vol. 13, Apr. 2019. [Online]. Available: https://doi.org/10.3389/fnhum.2019.00141

[36] X. Pei and C. Zheng, "Classification of left and right hand motor imagery tasks based on eeg frequency component selection," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, 2008, pp. 1888–1891.

[37] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain–computer interface," *GigaScience*, vol. 6, no. 7, May 2017. [Online]. Available: https://doi.org/10.1093/gigascience/gix034

[38] K. min Su, W. D. Hairston, and K. Robbins, "EEG-annotate: Automated identification and labeling of events in continuous signals with applications to EEG," *Journal of Neuroscience Methods*, vol. 293, pp. 359–374, Jan. 2018. [Online]. Available: https://doi.org/10.1016/j.jneumeth.2017.10.011

[39] G. Soghoyan, A. Ledovsky, M. Nekrashevich, O. Martynova, I. Polikanova, G. Portnova, A. Rebreikina, O. Sysoeva, and M. Sharaev, "A toolbox and crowdsourcing platform for automatic labeling of independent components in electroencephalography," *Frontiers in Neuroinformatics*, vol. 15, Dec. 2021. [Online]. Available: https://doi.org/10.3389/fninf.2021.720229