A distributed neural network architecture for dynamic sensor selection with application to bandwidth-constrained body-sensor networks

Thomas Strypsteen and Alexander Bertrand, *Senior Member, IEEE*

***Abstract*—We propose a dynamic sensor selection approach for deep neural networks (DNNs), which is able to derive an optimal sensor subset selection for each specific input sample instead of a fixed selection for the entire dataset. This dynamic selection is jointly learned with the task model in an end-to-end way, using the Gumbel-Softmax trick to allow the discrete decisions to be learned through standard backpropagation. We then show how we can use this dynamic selection to increase the lifetime of a wireless sensor network (WSN) by imposing constraints on how often each node is allowed to transmit. We further improve performance by including a dynamic spatial filter that makes the task-DNN more robust against the fact that it now needs to be able to handle a multitude of possible node subsets. Finally, we explain how the selection of the optimal channels can be distributed across the different nodes in a WSN. We validate this method on a use case in the context of body-sensor networks, where we use real electroencephalography (EEG) sensor data to emulate an EEG sensor network. We analyze the resulting trade-offs between transmission load and task accuracy.**

***Index Terms*—Distributed deep neural networks, Sensor Selection, Wireless sensor networks, EEG channel selection**

I. Introduction

Wearable, physiological sensors are increasingly being used for ambulant health monitoring, thanks to technological advances like miniaturization of microprocessors and energy-efficient batteries. In many applications though, multiple modalities or multiple channels of one data type will have to be measured on different devices on different locations on the body. This has led to the creation of body-sensor networks (BSNs), where the sensor nodes communicate over a wireless network to share their data and solve a given task together. One example of such a BSN is a neuro-sensor network that monitors the brain with with electroencephalography (EEG) sensors, a so-called Wireless EEG Sensor Network (WESN) [1,2]. EEG is a widely used, noninvasive way to record electrical brain activity, measuring potentials on multiple locations on the scalp to yield multi-channel time signals. These signals contain useful information for a variety of tasks such as epileptic seizure detection [3], sleep stage analysis [4] and brain-computer interfacing (BCI) [5]. In a WESN, multiple lightweight mini-EEG devices record one or a few EEG channels from their respective scalp areas, process the data, and transmit it to other nodes or a central fusion center, rather than using a bulky cap as in traditional EEG recording methods.

Ensuring maximal battery lifetime is a crucial consideration in the design of these BSNs. The energy bottleneck will typically be found in the wireless transmission of the data between the sensors and/or a fusion center [1]. This not only motivates energy-efficient hardware design, but also a shift in the algorithmic design of the models running on these sensor nodes. Instead of optimizing the model only for accuracy, the amount of data that needs to be transmitted when the model is used in the context of a BSN becomes an important design factor.

In this paper, we focus on reducing this data transmission by teaching the nodes a policy where they only transmit their data when the contribution of this specific node towards the inference task would be very informative for the current input sample, while upholding a given computational constraint. Such a constraint could be, e.g., that each node can on average only transmit at most 50% of its collected sensor data. To this end, we propose a *dynamic channel selection* methodology. For each block of collected samples across the nodes of the BSN, a distributed dynamic channel selector computes an input-dependent, optimal subset of channels, represented by a binary mask across the channels. Inference is then performed on the masked input by a deep neural network (DNN) at a fusion center which collects the data transmitted by the sensors. The selector and the inference model are trained jointly in an end-to-end manner, with the discrete parameters involved in the selection process being made trainable through the Gumbel-Softmax trick [6,7]. How often each channel is selected is limited by a per-channel sparsity loss on the computed masks. The usage of this dynamic channel selection means that the inference model will be presented with different channel subsets for different inputs, as if channels were randomly missing. We show that applying dynamic spatial filtering (DSF) [8] to the masked input to re-weight the channels helps the inference model become more robust against the missing of channels and

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 802895). The authors also acknowledge the financial support of the FWO (Research Foundation Flanders) for project G.0A49.18N, and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

T. Strypsteen and A. Bertrand are with KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and with Leuven.AI - KU Leuven institute for AI, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (e-mail: thomas.strypsteen@kuleuven.be, alexander.bertrand@kuleuven.be).

improves performance.

To validate our proposed architecture, we focus on a specific use case in the area of brain-computer interfaces, where a WESN needs to solve a motor execution task. We note that, while our evaluation use case is focused on brain signals, our proposed methodology is generic and can be applied to other kinds of BSNs or more broadly, to any kind of wireless sensor network (WSN).

The main contributions of this paper are:

- We propose an end-to-end learnable dynamic sensor or channel selection method that selects, for each window of a multi-channel input, an optimal subset of channels to use for inference, given a certain selection budget. This dynamic selection is learned jointly with the task DNN model and the Gumbel-Softmax trick is used to enable backpropagation for the discrete decisions involved.
- We demonstrate how this methodology can be used to reduce the transmission load in a wireless sensor network, thus increasing its battery lifetime. We do this by moving from centralized to distributed channel selection and enforcing per-node constraints to ensure a proper balancing of the transmission load. In addition, we present a use case where the method can improve the robustness of the classifier to noise bursts.

The paper is organized as follows. In section II we go over previous work in static and dynamic feature selection. Section III formally presents our problem statement and dynamic channel selection methodology. In section IV we provide an overview of the used dataset and how it was used to emulate a WESN environment and provide more details on the used model architecture and training strategy for this specific experiment. Our experimental results are then presented in section V and we end with some conclusions in section VI.

Note on terminology: Throughout this paper, we will always use the term 'channel selection' to refer to a selection of channels from a multi-channel input signal. Sensor selection or node selection could be viewed as a special case of channel selection. In the case of single-channel sensors, sensor or channel selection refer to the same thing. However, in the case of multi-channel sensors, sensor selection refers to the problem of selecting pre-defined *groups* of channels rather than individual channels, where each group corresponds to a sensor. For the sake of an easy exposition, but without loss of generality, we will assume single-channel sensors throughout this paper. Sometimes, we will refer to sensors as 'nodes' for consistency in terminology with the WSN literature.

II. Related work

A. Static feature selection

The goal of feature selection is to find an optimal subset of an available set of features that maximizes the performance of a classification or regression model on a given task. A host of literature exists that solves this problem in a *static* way, i.e., the optimal subset is determined for a certain dataset as a whole and the same selection is then applied to all input samples. Filter-based approaches rank the available features by a criterion like mutual information (MI) with the target labels and select the K highest scoring features [9]. Wrapper-based approaches use methods like greedy backward selection to efficiently explore the space of possible feature subsets, train the model on these candidate subsets and finally select the one that performs the best [10]. Embedded approaches jointly learn the subset and the task model in an end-to-end way, by performing L_1 regularization on the input weights [11] or learning the discrete parameters of the feature selection using continuous relaxations [12,13]. In this paper, we employ this approach of continuous relaxations to perform *dynamic* channel selection instead.

B. Dynamic feature selection

In *dynamic*, or *instance-wise* feature selection, the aim is to find an optimal subset of features for each individual input sample. One area where this approach has been highly relevant is field of explainable machine learning, where the goal is to indicate which features contributed most to the model output. For instance, L2X [14] trains an explainer model that maximizes the mutual information between the feature subset of size K of a given sample and the class distribution yielded by a trained task model. This line of work however, is mainly interested in finding the most relevant features for an already trained model, not in optimizing the performance of the model on reduced feature sets.

Another relevant area is the field of active feature acquisition [15,16]. In this setting, obtaining features is associated with a certain cost. The goal is then to obtain maximal model performance with a minimal amount of features, without being able access all the features of a given input sample from the start. This typically results in an iterative procedure where, based on the current feature subset, the optimal feature to extend the set with is estimated, until sufficient confidence in the model prediction is reached or the budget is saturated. In our WSN setting in contrast, we do have access to all the features to perform subset selection, but we are not allowed to centralize all of them by transmitting them over the wireless link. We also aim to avoid iterative procedures that require multiple communication rounds between the sensors and the fusion center as these are prone to latency issues in real-time situations.

The most similar approach to ours in terms of methods is taken by Verelst et al. in the field of computer vision [17]. The aim of their work is to decrease the computation time and energy of a CNN by learning input-dependent binary masks that are applied to the feature maps of an image at each layer. That layer then only performs convolutions on the pixels that are not masked out. A sparsity loss on these masks then forces the network to adhere to a certain computational budget, with backpropagation for the discrete masks being enabled through the Gumbel-Softmax trick. We will employ a similar strategy

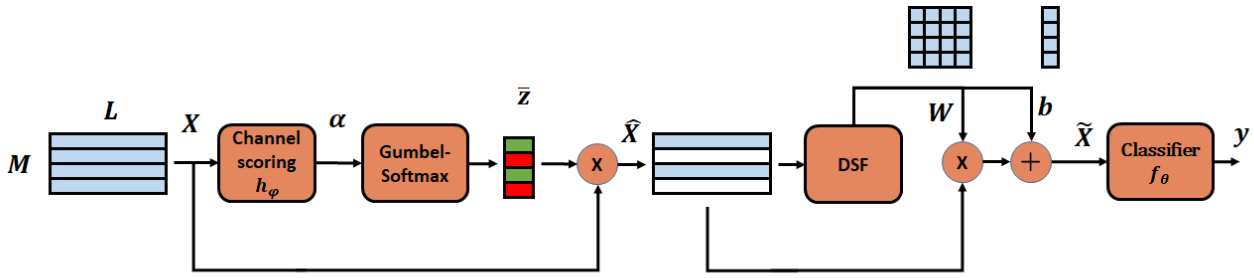


Figure 1: Overview of the dynamic channel selection. An L-sample window of an M-channel signal is passed to a channel scoring module, yielding unnormalized channel scores $\boldsymbol{\alpha}$. These are then converted in discrete selections by the Gumbel-Softmax module and applied to the input as a binary mask $\bar{\mathbf{z}}$, dropping a number of channels. This masked input $\hat{X}$ is then fed to the DSF module which re-weights the received channels with an attention mechanism, computing a weight matrix W and bias b that are multiplied with and added to the masked input: $\tilde{X} = W\hat{X} + b$ [8]. Finally, the original classifier is then applied to the resulting signal to obtain a prediction y. The entire pipeline is jointly learned in an end-to-end manner.

to learn binary masks for our dynamic selection, albeit in a distributed architecture and with the goal of reducing the data transmission over the wireless links, as detailed in the next section.

III. Proposed Method

In this section, we describe our dynamic channel selection methodology. Without loss of generality, we assume a classification task, although all methods can be easily extended to a regression task. For the sake of an easy exposition, we will initially assume a centralized architecture where all the channels are available to make a decision on the selection. In a WSN context, this implies that the channel selection is performed *after* transmitting all the channels to the fusion center, in which case there are no bandwidth savings. Nevertheless, this setting is still relevant to make the network robust against non-stationary noise influences and/or to reduce the computational complexity at the fusion center. Later on, in Section III-E, we will explain how the channel selection can be performed at the level of the sensor nodes, such that non-selected channels do not have to be transmitted at all.

A. Problem statement

Let $\mathcal{D} = \{(X^{(1)}, y^{(1)}), (X^{(2)}, y^{(2)}), \ldots, (X^{(N)}, y^{(N)})\}$ be a dataset of N samples of a multi-channel signal $X^{(i)}$ with class labels $y^{(i)}$. Each $X^{(i)} \in \mathbb{R}^{M \times L}$ contains M channels and a window of L consecutive time steps. We are also given a DNN model f_θ that is used to perform inference on these samples. Our goal is to learn a dynamic selection function that, for each separate input sample, determines this sample's optimal subset of channels to be presented to the inference model, while adhering to certain budget constraints on the amount of channels we are allowed to use on average. This selection of channels is based on a score vector $\boldsymbol{\alpha} \in \mathbb{R}^M$ that is computed by a channel scoring function $h_\varphi(X)$ for each input X. To go from a continuous score to a discrete selection in a way that still allows for end-to-end learning through backpropagation, we make use of a Gumbel-Softmax module G, which converts the score vector $\boldsymbol{\alpha}$ to a binary mask $\bar{\mathbf{z}} \in \{0,1\}^M$ to be applied to the input. Formally, this means learning parameters θ of the task model f_θ and the parameters φ of a selection model $s_\varphi = G \circ h_\varphi : \mathbb{R}^{M \times L} \mapsto \{0,1\}^{M \times 1}; X \mapsto \bar{\mathbf{z}}$ such that

$$\begin{aligned}\varphi^*, \theta^* =& \underset{\varphi,\theta}{\text{argmin}}\ \mathcal{L}_{CE}(f_\theta(X \odot \bar{\mathbf{z}}\mathbf{1}_L^\top), y) + \lambda \mathcal{L}_S(\bar{\mathbf{z}}) \\ =& \underset{\varphi,\theta}{\text{argmin}}\ \mathcal{L}_{CE}(f_\theta(X \odot s_\varphi(X)\mathbf{1}_L^\top), y) + \lambda \mathcal{L}_S(s_\varphi(X))\end{aligned} \tag{1}$$

with $\mathcal{L}_{CE}(p, y)$ the cross-entropy loss between the predicted label p and the ground truth y, $\odot$ an element-wise product, $\mathbf{1}_L^\top$ the row vector of dimension L containing only ones, $\mathcal{L}_S$ a cost function that enforces sparsity in the learned masks and λ a hyperparameter to balance the two losses. A schematic overview of our method is presented in Fig. 1. We will now delve deeper into the design of each of the modules involved.

B. Learning discrete decisions with Gumbel-Softmax

To enable the network to learn discrete decisions while still keeping the entire network end-to-end learnable we make use of the Gumbel-Softmax trick [6,7]. Take a discrete random variable, drawn from a categorical distribution with K classes and class probabilities $\pi_1, \ldots \pi_K$, represented as a one-hot vector $\bar{\mathbf{y}} \in \{0,1\}^K$, with the index of the one indicating the class $\bar{\mathbf{y}}$ belongs to. Discrete samples from this distribution can then be drawn with the Gumbel-Max trick:

$$\bar{\mathbf{y}} = \text{one_hot}(\underset{k}{\text{argmax}}\ (\log \pi_k + g_k)) \tag{2}$$

with g_k independent and identically distributed (i.i.d.) samples from the Gumbel distribution [18] and one_hot(i) the operator that generates a one-hot $K \times 1$ vector where the one is placed at position i. The Gumbel-Softmax is then a continuous, differentiable relaxation of this discrete sampling procedure, approximating the discrete one-hot vectors $\bar{\mathbf{y}}$ with continuous vectors $\mathbf{y}$ whose elements sum to one instead by replacing the argmax with a softmax. For the k-th element y_k, this results in:

$$y_k = \frac{\exp((\log\ \pi_k + g_k)/\tau)}{\sum_{j=1}^K \exp((\log\ \pi_j + g_j)/\tau)} \tag{3}$$

with τ the temperature of this continuous relaxation. Lowering the temperature causes the softmax to more closely resemble an argmax, thus causing the continuous $\mathbf{y}$ to be a closer approximation of the discrete $\bar{\mathbf{y}}$. It will however, also cause the relaxation to become less smooth and increase the variance of the gradients.

Our goal is to model a learnable, binary random variable $\bar{z}_m$ for each channel m, which is 1 when the channel is selected and 0 otherwise. In the case of such a binary random variable $\bar{z}_m$, with $P(\bar{z}_m = 1) = \pi_1$, it can be shown [17] that by setting $K = 2$ in Eq. 3, the Gumbel-Softmax trick can be simplified to

$$\begin{aligned} y_1 &= \sigma\left(\frac{\log \pi_1 + g_1 - g_2}{\tau}\right) \\ y_2 &= 1 - y_1 \end{aligned} \tag{4}$$

with $\sigma(\cdot)$ the sigmoid function. A continuous relaxation z_m of the binary random variable $\bar{z}_m$ can then be obtained by taking $z_m = y_1$ We can use this binary Gumbel-Softmax trick to transform unnormalized, learnable channel scores $\boldsymbol{\alpha} \in \mathbb{R}^M$ yielded by a network $h_\varphi(X)$ into continuous, differentiable approximations $\mathbf{z} = [z_1, ..., z_M]^\top$ of the discrete $\bar{\mathbf{z}} \in \{0, 1\}^M$. There are a number of ways this continuous relaxation can be used to obtain approximating gradients for the discrete $\bar{\mathbf{z}}$, but we will follow the Straight-Through estimator approach [6,17]. This means that we will sample discrete decisions from our binary distribution in the forward pass:

$$\begin{aligned} \bar{z}_m &= \left\lfloor \sigma\left(\frac{\alpha_m + g_1 - g_2}{\tau}\right) \right\rceil \\ &= \begin{cases} 1, & \text{if } z_m = \sigma\left(\frac{\alpha_m + g_1 - g_2}{\tau}\right) > 0.5 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

with $\lfloor \cdot \rceil$ the rounding operator, resulting in a binary distribution where $P(\bar{z}_m = 1) = \sigma(\alpha_m)$ (replacing π_1 in Eq. 4). To enable backpropagation through the discrete rounding operator, we use gradients from the continuous relaxation in the backward pass, which implies the approximation

$$\nabla_\varphi \bar{\mathbf{z}} \approx \nabla_\varphi \mathbf{z}. \tag{6}$$

This scheme allows for hard decisions to be used during training and learned through end-to-end backpropagation. This process is schematically illustrated in Fig. 2. At inference time, Gumbel noise is no longer added to the score vector, resulting in the network no longer sampling from binary distributions, but behaving in a deterministic manner instead, i.e. $\bar{z}_m = 1$ if $\sigma(\alpha_m) > 0.5$.

C. Enforcing sparsity

We assume each channel is measured on a different node of a wireless sensor network, whose nodes are able to communicate with each other over bandwidth-constrained links. To reduce the communication load of these nodes, we want each node m to only transmit their data (i.e. yield a 1 in their binary mask) for a maximal target percentage $T \in [0, 1]$ of the input samples. We thus want the expected value of each element of the binary mask over the distribution of our input samples X to be below this target T:

$$\begin{aligned} &\mathbb{E}_X[\bar{z}_m] \leq T \\ &m = 1, .., M \end{aligned} \tag{7}$$

These per-node constraints ensure that the masks are not only sparse, but balanced across the different nodes as well, meaning that there is no single node that is transmitting significantly more than the others. Secondly, by applying the constraints to the expected value of the masks instead of on each separate masks, we allow for a variable amount of nodes to be used for different input samples. To enforce these constraints, we use a mini-max optimization in which we impose a per-node sparsity loss on the decisions made during training and aggregate these by penalizing the node that currently violates this constraint the most:

$$\begin{aligned} \mathcal{L}_{S,m} &= \max\left(\frac{1}{B}\sum_{b=1}^{B} \sigma\left(\frac{\alpha_m^{(b)}}{\tau_0}\right) - T, 0\right)^2 \\ \mathcal{L}_S &= \max_m \mathcal{L}_{S,m} \end{aligned} \tag{8}$$

with B the batch size, $\alpha_m^{(b)}$ the score for node m for the b'th input sample of the batch and τ_0 a temperature constant we set at 0.1. This sparsity loss replaces the expected value in the constraints of Eq. 7 with a batch average and the discrete node decisions $\bar{z}_m$ with the continuous approximation $\sigma\left(\frac{\alpha_m^{(b)}}{\tau_0}\right)$. The fact that this approximation is computed through a sigmoid with a low temperature, without the addition of Gumbel noise means we more closely approximate the behaviour of the selection layer at inference time than we would if we directly penalized the hard decisions $\bar{\mathbf{z}}$. This is important to ensure that if the sparsity constraints are met at training time, they will also be met at inference time. For instance, if the network ensures that for all X, $\sigma(\alpha_m) = 0.51$, then due to the addition of Gumbel noise in the computation of $\bar{z}_m$, the network will sample $\bar{z}_m = 1$ and $\bar{z}_m = 0$ an about equal amount of times during training. At inference time however, when no noise is added, the network will always yield $\bar{z}_m = 1$ since $\sigma(\alpha_m) > 0.5$, surely violating the constraints. Also important to note is that using Eq. 8 requires training with a large enough batch size such that the batch average of Eq. 8 is a meaningful estimate of the expected value used in Eq. 7.

D. Dealing with different channel subsets

Performing dynamic channel selection means the classification network f_θ will see a different subset of active channels depending on the current input sample. Stated in another way, our network needs to be able to deal with missing inputs. This can cause problems in the learning of the network weights, as the network needs to be able to extract relevant information when a channel is selected, but not cause interference when the channel is not selected and the corresponding input only contains zeros. Ideally, we would employ a number of separate classification networks, each optimized for a specific channel

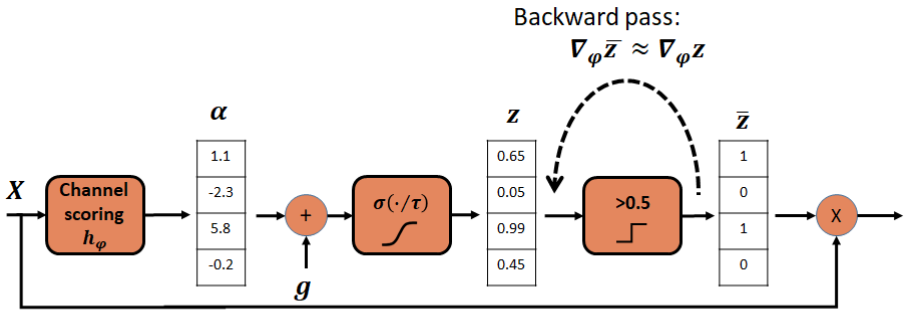


Figure 2: Illustration of the Gumbel-Softmax trick. During training, hard decisions are sampled by perturbing the channel scores $\boldsymbol{\alpha}$ with Gumbel noise, passing this through a sigmoid to obtain soft probabilities and applying a thresholding operator. Backpropagation through the thresholding operation is enabled by using a Straigth-Through estimator, treating the threshold in the backward pass as an identity function, i.e. $\frac{\partial \bar{\mathbf{z}}}{\partial \mathbf{z}} \approx 1$ and thus $\nabla_\varphi \bar{\mathbf{z}} \approx \nabla_\varphi \mathbf{z}$

subset. In practice however, this would require training and storage of 2^M networks, which quickly becomes infeasible. Thus, the question arises how we can make a single network be able to cope as efficiently as possible when multiple input sets are possible. We tackled this issue by extending our network with the Dynamic Spatial Filtering (DSF) proposed by Banville et al. [8]. The idea of DSF is to re-weight the M input channels using an attention layer. In this setting, new (virtual) channels are formed by applying a spatial filter to all input channels, i.e., making linear combinations of the channels, with the weights being computed from the spatial covariance matrix of the current input window. This re-weighting decreases the impact of missing channels on the network activations and has been shown to make a network more robust against noisy or missing channels.

E. From centralized to distributed

The channel scoring function h_φ in Eq. 1 currently still uses all M input channels to make a decision. However, an important aspect to be taken into account is the distributed nature of WSN platforms, where different channels are recorded on different physical devices. In this setting, we want to reduce the transmission load of these devices by only selecting and thus transmitting the signal of a node when its information is relevant for the current sample. However this will only actually be beneficial when we are able to perform the selection *without centralizing the data of the different sensors*. We will consider three different cases corresponding to different constraints on our dynamic channel scoring function $h_\varphi(X)$:

- *Centralized*: The selection is derived from the joint information of all channels, i.e., $\boldsymbol{\alpha} = h_\varphi(X)$. This setting serves as a theoretical upper bound for the following two practical settings.
- *Distributed*: Each node has to decide whether to transmit solely on its own data, i.e., $\alpha_m = h_{\varphi,m}(\mathbf{x}_m)$ where x_m denotes the m-th row of X.
- *Distributed-Feedback*: Each node computes a short vector $\boldsymbol{\beta}_m = h_{\varphi,m}(\mathbf{x}_m) \in \mathbb{R}^C$ with $C << L$, that is transmitted to the fusion center. At the fusion center, the $\boldsymbol{\beta}_m$ of all nodes are combined into the stacked vector $\boldsymbol{\beta}$ to determine a final scoring vector $\boldsymbol{\alpha} = g_\phi(\boldsymbol{\beta})$. The discrete selection $\bar{\mathbf{z}}$ resulting from this scoring vector is then returned to the nodes to inform them which of them should transmit. The size of these vectors $\boldsymbol{\beta}_m$ should be small compared to the length L of the window to be transmitted to minimize the overhead cost of the selection. In order to reduce the trainable parameters, one can decide to make the different $h_{\varphi,m}$ models copies of each other, with shared weights for all layers, except the final layer having its own set of parameters for each channel m.

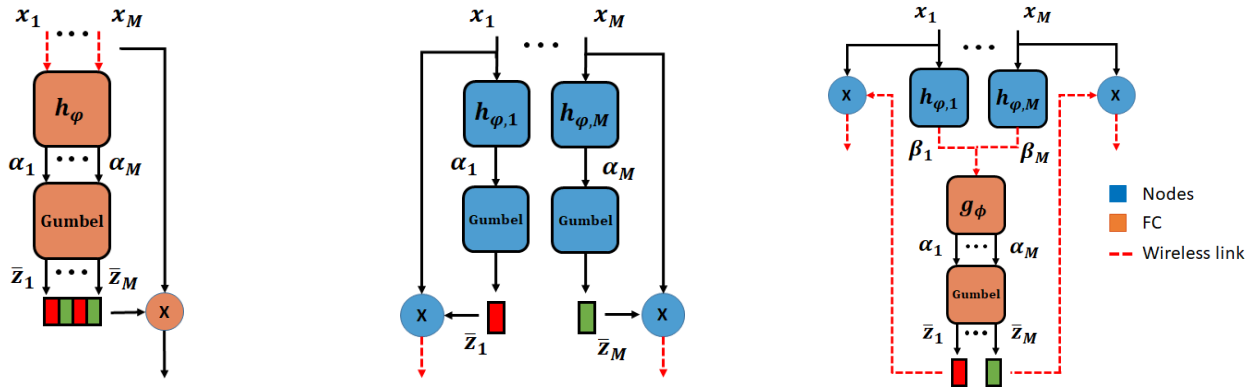
These three settings are illustrated in Fig. 3.

F. Training strategy

Successfully training a model with masking units typically hinges on a good initialization. Since a sparsity loss is much easier to minimize than the training objective - by simply driving the weights of the binary masks to zero - the network can quickly collapse into a state where barely any units are executed [17,19]. Once this has happened, it is very hard for the network to learn task-relevant information that could pull it out of this state. To avoid this, we adopt a step-wise training strategy that learns one module at a time.

1) Initialize the weights of the classifier f_θ with the weights of the original M-channel network trained without any dynamic selection.
2) Add the *centralized* dynamic selection layer and train it while fine-tuning (i.e., training at a lower learning rate) the classifier.
3) Add the DSF module and train it while fine-tuning the the dynamic selection and classifier.
4) Transform the centralized dynamic selection layer in a distributed dynamic selection layer and fine-tune the whole model (see below).

To go from a centralized to a distributed architecture, we employ a 2-step transfer learning approach. First, we employ the centralized channel scoring function h_φ as a teacher model and try to ensure that the outputs $\boldsymbol{\alpha}_{distr}$ of the student model



(a) Centralized (b) Distributed (c) Distributed-Feedback

Figure 3: Different settings for the channel scoring module in a sensor network. Blue blocks indicate modules on the sensor nodes, orange modules on the fusion center and red dotted lines communications between the two. (a) The centralized upper bound employs all joint information of all channels to compute a score, but would require all data to be centralized. (b) The distributed setting does not allow for communication between the nodes (via the fusion center) and makes each score only dependent on the local data. (c) The distributed-feedback setting allows for a small amount of communication between the nodes to make a better, joint decision.

- the distributed channel scoring function - match the outputs $\boldsymbol{\alpha}_{centr}$ of the teacher by minimizing the following loss:

$$\mathcal{L}(h_{\varphi,distr}) = \mathcal{L}_{BCE}(\sigma(\boldsymbol{\alpha}_{distr}), \lfloor \sigma(\boldsymbol{\alpha}_{centr}) \rceil) \qquad (9)$$

with $\mathcal{L}_{BCE}$ the binary cross-entropy loss. By minimizing this loss, we do not necessarily ensure that the channel scores $\boldsymbol{\alpha}$ are exactly alike, but rather that the discrete outputs at inference time will be similar, which is what we ultimately want. In the final step, we use the newly learned distributed channel selection layer, initialize the DSF module and the classifier with the corresponding weights of the centralized model and fine-tune the whole network in an end-to-end fashion.

IV. Application to Wireless EEG Sensor Networks

A. Data set

In the field of BCI, the motor execution paradigm is used to decode body movement from the corresponding neural signals in the motorsensory areas of the brain. The High Gamma Dataset [20] contains EEG data from about 1000 trials of executed movements for each of the 14 subjects, as well as a separate test set of about 180 trials per subject, all following a visual cue. The dataset includes 4 classes of movements: left hand, right hand, feet, and rest. As in [20] we only use the 44 channels that cover the motor cortex, which are preprocessed by resampling at 250 Hz, highpass filtering above 4 Hz, standardizing the per-channel mean and variance to 0 and 1 respectively, and extracting a window of 4.5 seconds for each trial. This pre-processing is adopted from [20] and described in full detail there.

B. WESN node emulation and selection

In mini-EEG devices, we cannot measure the potential between a given electrode and a distant reference (e.g. the mastoid or Cz electrode) , as we would in traditional EEG caps [10]. Instead, we can only record the local potential between two nearby electrodes belonging to the same sensor device. To emulate this setting using a standard cap-EEG recording, we follow the method proposed in [10], which considers each pair of electrodes within a certain maximum distance as a candidate electrode pair or node. By subtracting one channel from the other, we can remove the common far-distance reference and obtain a signal that emulates the local potential of the node. Applying this method with a distance threshold of 3 cm to our dataset, we obtain a set of 286 candidate electrode pairs or nodes, which have an average inter-electrode distance of 1.98 cm and a standard deviation of 0.59 cm.

Given that our WESN will consist of a limited number of mini-EEG devices, we first need to select the M most informative sensor nodes from the pool of 286 candidate nodes. To achieve this, we adopt the static channel selection method described in [13], which enables us to learn the M optimal nodes for a given task and neural network by jointly training the network and a selection layer. Note that this is a fixed selection for the entire data set, not for each sample separately. We train this selection layer, along with the centralized network (f_θ in Fig. 1 we will use for classification (see Section IV-C), using data from all subjects in the dataset, which results in a subject-independent set of M mini-EEG nodes that are best suited for solving the motor execution task. We do this for 3 different values of M, corresponding to a small WESN ($M = 4$ nodes), a medium-size WESN ($M = 8$ nodes), and a high-density WESN ($M = 16$ nodes).

C. Model architecture

As mentioned above, the neural network architecture we employ for classification (f_θ in Fig. 1) is the MSFBCNN proposed in [21], which was designed specifically for a motor execution task. Inspired by the Filterbank-CSP approach of [22], this model computes log-power features by applying

a number of temporal filters in parallel, aggregating these with spatial filters, and then applying squaring and average pooling over time. These features are then classified by a single linear layer. While the details of this network are not relevant for this study, we provide a summary of this network in table format in Appendix A for completeness.

For our channel scoring module h_φ in the centralized setting, we will employ the same architecture, with the final linear layer being adapted to output a vector of dimension $M \times 1$. In the two distributed settings, the different node scoring models $h_{\varphi,m}$ are copies of each other, with shared weights for all layers, except the final layer having its own set of parameters for each node/channel m. The network architecture of these $h_{\varphi,m}$ is simply a single-input version of the M-input network define by f_θ, but where the last fully-connected layer outputs the scalar α_m in the distributed setting and the node summary $\boldsymbol{\beta}_m \in \mathbb{R}^{C\times 1}$ in the distributed-feedback setting. The dimension of these node summaries $\boldsymbol{\beta}_m$ is chosen to be $C = 10$, ensuring the overhead of its transmission is negligible compared to the transmission of the full window of $L = 1125$ time samples.The module $g_\phi(\boldsymbol{\beta})$ aggregating the node summaries in the fusion center is a simple 2-layer multilayer perceptron (MLP) with a hidden dimension of 50 and ReLU nonlinearity. The DSF module also consists of a 2-layer MLP with hidden dimension 50 and ReLU nonlinearity, which is applied to the vectorized sample covariance matrix $\frac{1}{L}\hat{X}\hat{X}^\top$ of the masked input sample and which produces a weight matrix $W \in \mathbb{R}^{M\times M}$ and a bias $b \in \mathbb{R}^{M\times 1}$ which are used to compute a re-weighted output $\tilde{X} = W\hat{X} + b$.

Finally, for training, we follow the procedure described in section III-F, using the Adam optimizer [23] with a learning rate of 10^{-3} when a module is trained for the first time and 10^{-4} when it is fine-tuned during subsequent steps. A batchsize of 64 is employed and training lasts for 100 epochs with early stopping activated when the validation loss does not decrease for 10 epochs. The hyperparameter λ, controlling the penalization of the sparsity loss was set to 10 for this application and a fixed temperature $\tau = 1$ was used for the Gumbel-Softmax module.

V. Experimental Results

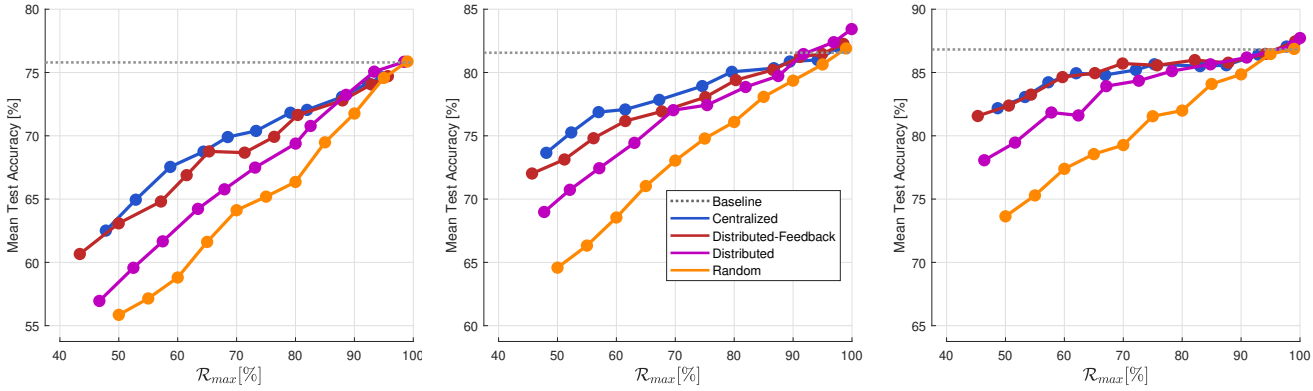
A. Centralized versus distributed performance

We first analyze the rate-accuracy trade-off obtained by our proposed dynamic channel selection method and investigate the impact of going from a theoretical, centralized approach to a practical, distributed one. To do this, we train our model for a given target rate T, indicating the maximal percentage of input samples for which the data of each node should be used and transmitted. Since the lifetime of the WSN as a whole will ultimately be determined by the node with the highest transmission rate, we will always report the *maximal* rate $\mathcal{R}_{max}$ among the nodes, rather than the *average*. As a proof-of-concept benchmark, we will employ a naive system that for each sample and for each node randomly determines whether the data should be transmitted, doing so with a probability equal to the relative target rate. This allows us to investigate whether the dynamic selection is truly able to make intelligent decisions that go beyond simply making sure the constraints are met. We train our model for a range of target rates and for networks consisting of 4, 8 and 16 nodes, each time averaging the results over 5 runs.

Fig. 4 shows the resulting rate-accuracy tradeoffs. Firstly, we can observe that the dynamic selection indeed consistently outperforms the random selection, with the gap widening as the target rate decreases. Secondly, while the distributed network shows a small performance gap with the centralized upper bound, allowing for even a small amount of communication between the nodes by employing the distributed-feedback setting compensates for this as it performs very similarly to the centralized setting. When comparing the networks with a different amount of nodes, it can be observed that the more nodes are being used, the smaller the relative performance losses are when moving from the starting rate of 100% to a rate of 50% (for instance, the 16-node network only loses 5% accuracy, while the 4-node network loses 10%). Since there will be a higher amount of redundancy between the 16 nodes than between the 4 nodes, it makes sense that dropping channels in the former has less of an impact on the accuracy than in the latter.

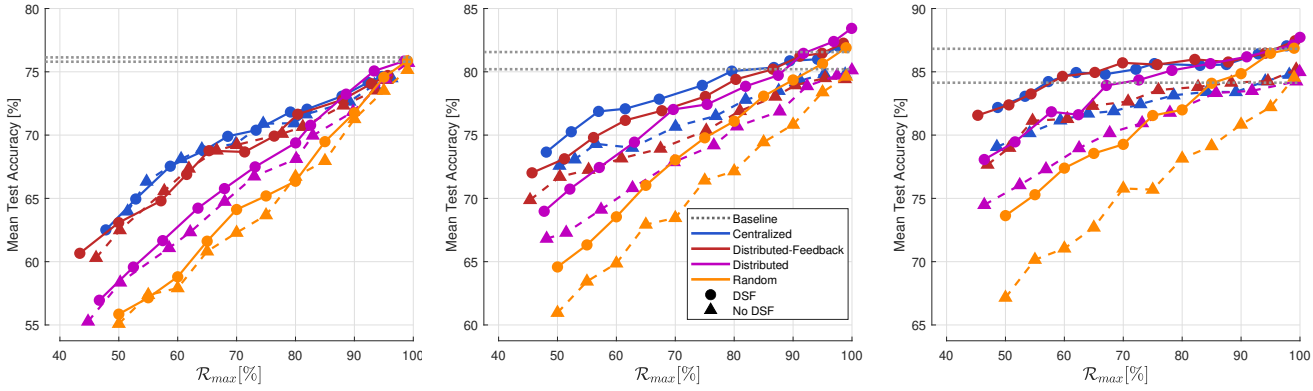
B. Impact of dynamic spatial filtering

Next, we analyze the importance of the presence of the DSF module by comparing it with the networks where it has not been added. The results are illustrated in Fig. 5. For the 4-node network, it can be observed that the inclusion of DSF only results in a small improvement for the random selection network and no improvement at all on the dynamic selection network. As mentioned in section III-D, the main purpose of the DSF module is to increase the capability of the classifier to cope with the different channel subsets it is presented with. In this small, 4-node network, the amount of channel subsets 2^M is still manageable. Furthermore, in contrast to the random selection, the dynamic selection will not randomly sample from *all* possible channel subsets, but will focus its sampling on a more select number of these. For instance, it will almost always make sure that both a node from the left and right hemisphere is included, since the difference between both is highly informative for motor execution. This explains why the inclusion of DSF is slightly more necessary - and thus beneficial - in the case of random selection than in the case of dynamic selection. When moving to 8- and 16-node networks, the amount of subsets quickly grow and the performance gains afforded by DSF become more and more salient, with these gains once again being slightly higher for the random selection than for the dynamic selection.



(a) $M = 4$ nodes (b) $M = 8$ nodes (c) $M = 16$ nodes

Figure 4: Rate-accuracy trade-off for the proposed dynamic channel selection method for networks of 4,8 and 16 nodes. Mean test accuracies are plotted against the percentage of samples for which the critical node of the network needs to transmit, i.e. the node with the highest percentage of transmission. Each data point is an average of 5 runs for a given maximal target rate T (see Eqs. 7 and 8). Baseline performance indicates accuracy without dynamic selection involved, i.e. each node transmits at a rate $\mathcal{R}$ of 100%. Dynamic selection consistently outperforms random channel selection. While a gap exists between the distributed implementation and the centralized upper bound, sharing a small amount of information between the nodes in the distributed-feedback setting largely overcomes this and performs about as well as the centralized upper bound.



(a) $M = 4$ nodes (b) $M = 8$ nodes (c) $M = 16$ nodes

Figure 5: Rate-accuracy trade-off for the proposed dynamic channel selection method for networks of 4,8 and 16 nodes, with and without inclusion of the DSF module. Baseline performance indicates accuracy without dynamic selection involved, i.e. each node transmits at a rate $\mathcal{R}$ of 100%. While DSF is not necessary for the 4-node network, it becomes more important as the size of the network increases, delivering a consistent performance gain across all settings.

C. Impact of noisy environments

Up until now, we have discussed situations where we perform a trade-off between the amount of channels we reject and the accuracy of the classifier. In some cases however, working with only a subset of the channels can actually be beneficial for the accuracy as well. In environments where sudden noise bursts can occur, these unexpected inputs, even when limited to a single channel, can heavily disturb the activations of the entire neural network network and lead to misclassifications. To make it easier for the network weights to be robust against these noise bursts, it can be beneficial to detect when these happen and zero the corresponding input instead. To test this hypothesis, we repeated our previous experiments with the dynamic selection method, but in this case, each channel of each input window had a 25% chance to be replaced by Gaussian noise with zero mean and standard deviation uniformly sampled between 0 and 3 instead, leaving no more relevant information on this channel. Fig. 6 compares the performance of the distributed-feedback dynamic selection with a baseline network, which directly takes the perturbed data as input. The noise is added during both training and testing to enable a fair comparison, i.e., the network without dynamic channel selection can in principle learn how to cope with these noise bursts. Firstly, it can be observed that the dynamic selection never transmits more information than absolutely necessary: only 75% of the channels actually contain information, so the resulting rate is automatically capped around 75%. Secondly, the automatic rejection of noisy channels does indeed lead to an increased accuracy compared to the baseline accepting this noise as input. A probable reason is that it will be easier for the classifier to find weights that process normal inputs normally and minimize the impact on the activations of disturbances when these disturbances are zero inputs rather than noise bursts.

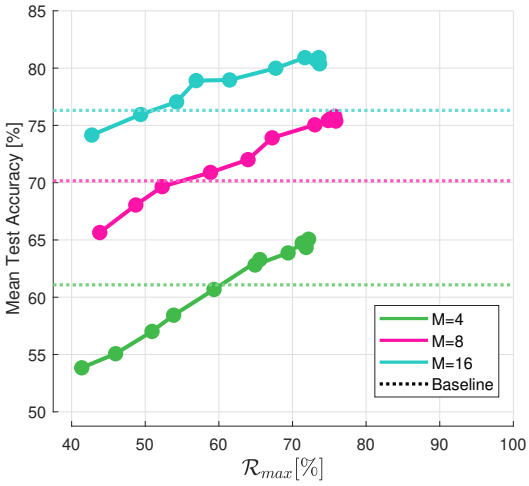


Figure 6: Rate-accuracy trade-off for the proposed dynamic channel selection method in the distributed-feedback setting for networks of 4, 8 and 16 nodes in a simulation of a noisy environment where each channel has a 25% probability to be replaced by Gaussian noise. When the model is trained with a target rate above this threshold, it automatically rejects the transmission of the noisy and the resulting rate is capped at 75%. Rejecting these noisy channels yields higher accuracies than the baseline network which accepts the noisy channels as input, demonstrating the network is now more robust against this noise.

VI. Conclusion and Future Outlook

We have proposed a dynamic channel or sensor selection method in order to reduce the communication cost and improve the battery lifetime of WSNs. For each input window, the method selects the optimal subset of sensors to be used by a neural network classifier, while optimizing a trade-off between the amount of channels selected and the accuracy of the given task. The dynamic selection and the classifier are jointly trained in an end-to-end way through backpropagation. The dynamic selection module consists of three major parts: a channel scoring function assigning a relevance score to each channel, a binary Gumbel-Softmax trick converting these scores to discrete decisions and the dynamic spatial filtering module of Banville et al. [8] to make the classifier more robust against the resulting absence of channels. A crucial aspect of this dynamic selection is that it can computed in a *distributed* way, requiring minimal communication overhead between the nodes.

We have demonstrated the use of this method to perform a trade-off between the transmission rate of the nodes in an emulated wireless EEG sensor network and the accuracy of a motor execution task. Additionally, we have presented a use case where the dynamic selection can even improve the accuracy of the model, by automatically rejecting inputs that might harm performance, such as heavy bursts of noise. Though we have focused on the application use case of wireless EEG sensor networks, our methodology is generic and can be applied to sensor networks with any kind of modalities. In future work, we will explore applications of this method in other distributed platforms than WESNs.

References

[1] A. Bertrand, "Distributed signal processing for wireless EEG sensor networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 6, pp. 923–935, 2015.

[2] T. Strypsteen and A. Bertrand, "Bandwidth-efficient distributed neural network architectures with application to neuro-sensor networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 933–943, 2022.

[3] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International journal of neural systems*, vol. 29, no. 04, p. 1850011, 2019.

[4] O. De Wel, M. Lavanga, A. C. Dorado, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Complexity analysis of neonatal EEG using multiscale entropy: applications in brain maturation and sleep stage classification," *Entropy*, vol. 19, no. 10, p. 516, 2017.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[6] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[7] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[8] H. Banville, S. U. Wood, C. Aimone, D.-A. Engemann, and A. Gramfort, "Robust learning from corrupted EEG with dynamic spatial filtering," *NeuroImage*, vol. 251, p. 118994, 2022.

[9] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG channel selection in brain computer interfaces by mutual information maximization," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 7064–7067.

[10] A. M. Narayanan and A. Bertrand, "Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 234–244, 2019.

[11] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, 2017.

[12] A. Abid, M. F. Balin, and J. Zou, "Concrete autoencoders for differentiable feature selection and reconstruction," *arXiv preprint arXiv:1901.09346*, 2019.

[13] T. Strypsteen and A. Bertrand, "End-to-end learnable EEG channel selection for deep neural networks with gumbel-softmax," *Journal of Neural Engineering*, vol. 18, no. 4, p. 0460a9, 2021.

[14] J. Chen, L. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 883–892.

[15] Y. Li and J. Oliva, "Active feature acquisition with generative surrogate models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6450–6459.

[16] I. Covert, W. Qiu, M. Lu, N. Kim, N. White, and S.-I. Lee, "Learning to maximize mutual information for dynamic feature selection," *arXiv preprint arXiv:2301.00557*, 2023.

[17] T. Verelst and T. Tuytelaars, "Dynamic convolutions: Exploiting spatial sparsity for faster inference," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2320–2329.

[18] E. J. Gumbel, *Statistical theory of extreme values and some practical applications: a series of lectures*. US Government Printing Office, 1948, vol. 33.

[19] M. Figurnov, M. D. Collins, Y. Zhu, L. Zhang, J. Huang, D. Vetrov, and R. Salakhutdinov, "Spatially adaptive computation time for residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1039–1048.

[20] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[21] H. Wu, F. Li, Y. Li, B. Fu, G. Shi, M. Dong, and Y. Niu, "A parallel multiscale filter bank convolutional neural networks for motor imagery EEG classification," *Frontiers in Neuroscience*, vol. 13, p. 1275, 2019.

[22] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 2390–2397.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

APPENDIX A
MSFBCNN ARCHITECTURE

Layer	# Filters	Kernel	Stride	# Params	Output	Activation	Padding
Input					(C,T)		
Reshape					$(1, T, C)$		
Timeconv1	F_T	$(64, 1)$	$(1, 1)$	$64F_T$	(F_T, T, C)	Linear	Same
Timeconv2	F_T	$(40, 1)$	$(1, 1)$	$40F_T$	(F_T, T, C)	Linear	Same
Timeconv3	F_T	$(26, 1)$	$(1, 1)$	$26F_T$	(F_T, T, C)	Linear	Same
Timeconv4	F_T	$(16, 1)$	$(1, 1)$	$16F_T$	(F_T, T, C)	Linear	Same
Concatenate					$(4F_T, T, C)$		
BatchNorm				$2F_T$	$(4F_T, T, C)$		
Spatialconv	F_S	$(1, C)$	$(1, 1)$	$4CF_TF_S$	$(F_S, T, 1)$	Linear	Valid
BatchNorm				$2F_S$	$(F_S, T, 1)$		
Non-linear					$(F_S, T, 1)$	Square	
AveragePool		$(75, 1)$	$(15, 1)$		$(F_S, T/15, 1)$		Valid
Non-linear					$(F_S, T/15, 1)$	Log	
Dropout					$(F_S, T/15, 1)$		
Dense	N_C	$(T/15, 1)$	$(1, 1)$	$F_S(T/15)N_C$	N_C	Linear	Valid

Table I: Architecture of the MSFBCNN used for motor execution classification. This table is cited from [21].