Recognizing motor imagery tasks from EEG oscillations through a novel ensemble-based neural network architecture

Antonio L. Alfeo, Vincenzo Catrambone, Mario G.C.A. Cimino, Gigliola Vaglini, Gaetano Valenza

```
@article{recognizing2022,
    author={Alfeo, Antonio L. and Catrambone, Vincenzo and Cimino, Mario G.C.A. and Vaglini, Gigliola and Valenza, Gaetano},
    title={Recognizing motor imagery tasks from EEG oscillations through a novel ensemble-based neural network architecture},
    journal={2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp; Biology Society (EMBC)},
    year={2022},
    volume={},
    pages={},
    publisher={IEEE},
    doi={10.1109/embc46164.2021.9629900},
    issn={},
}
```

Antonio L. Alfeo, Vincenzo Catrambone, Mario G.C.A. Cimino, Gigliola Vaglini, Gaetano Valenza. "Recognizing motor imagery tasks from EEG oscillations through a novel ensemble-based neural network architecture" 2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp; Biology Society (EMBC) (2022): .

Recognizing motor imagery tasks from EEG oscillations through a novel ensemble-based neural network architecture

Antonio L. Alfeo[1,*], Vincenzo Catrambone[1,2,*], Mario G.C.A. Cimino[1,2], Gigliola Vaglini[1,2], and Gaetano Valenza[1,2]

Abstract— **Brain-Computer Interfaces (BCI) provide effective tools aimed at recognizing different brain activities, translate them into actions, and enable humans to directly communicate through them. In this context, the need for strong recognition performances results in increasingly sophisticated machine learning (ML) techniques, which may result in poor performance in a real application (e.g., limiting a real-time implementation). Here, we propose an *ensemble* approach to effectively balance between ML performance and computational costs in a BCI framework. The proposed model builds a classifier by combining different ML models (base-models) that are specialized to different classification sub-problems. More specifically, we employ this strategy with an ensemble-based architecture consisting of multi-layer perceptrons, and test its performance on a publicly available electroencephalography-based BCI dataset with four-class motor imagery tasks. Compared to previously proposed models tested on the same dataset, the proposed approach provides greater average classification performances and lower inter-subject variability.**

I. Introduction

Brain-Computer Interfaces (BCI) are defined as neural interfaces allowing to communicate through neurophysiological signs [1], and several BCI systems have been developed and are daily used by patients worldwide [1]. Motor imagery (MI) electroencephalography (EEG)-based BCI are widely used BCI systems aimed to analyze and process the cortical activity generated during a MI, and convert it into an easy to use end-effector information, e.g. a categorical label associated with a predefined action [1]. Such a technological solution is possible because of the high time resolution and feasibility of EEG recordings, the knowledge that imagination activates areas of the brain that are also responsible for generating actual movement, and machine learning (ML) algorithms which are increasingly used for automatic knowledge extraction from physiological signals [2]–[4]. In order to identify optimal ML tools, BCI competitions have been organized [5], allowing the scientific community also to share tools and compare different algorithmic solutions on common experimental conditions.

Among the several ML architectures proposed in recent years to tackle BCI challenges, the *ensemble-based* approaches proved to be effective. Those models are defined as ensemble since they leverage on multiple learning algorithms defined as base-models and combine their outcome to determine the final prediction [6]. The benefits provided by ensemble-based approaches may include robustness towards over-fitting and local minima, as well as a generally improved coverage of the solution space [7], intended as the space of all potential solutions for a given algorithmic problem. It has been empirically and theoretically shown that the predictive performances of an ensemble-based approach strongly depends on the diversification of its base-models. According to the taxonomy presented in [7], this can be achieved by implementing several strategies that may be distinguished according to which of the three main blocks of a ML application they manipulate (i.e., input, learning algorithm, and output). First, differentiating the base models by manipulating their *inputs* consists of assigning different partitions of the data set to different base-models, e.g. using different instances of all features (horizontal partitioning), or using all instances of different features (vertical partitioning). Second, differentiating the base models by manipulating their *learning algorithm* consists of implementing different base-models by varying the hyper-parameter values or changing the learning algorithm. Third, differentiating the base models by manipulating their *output* consists of putting into place the *divide and conquer* strategy: a single complex problem is divided into multiple easy-to-handle sub-problems to be addressed through the different base-models. This can be implemented via the so-called multi-class problem decomposition, i.e. transforming a single multi-class problem into multiple binary problems. From a classification viewpoint, the two main decomposition strategies are *one-vs-one* and *one-vs-all* [8], which have successfully been applied in EEG-based BCI application [9], [10]. With a *one-vs-one* strategy, the C-multiclass problem is decomposed into $C(C-1)/2$ two-class classification problems, one problem for each pair of classes. With a *one-vs-all* strategy, a sub-problem is created for each class, and the classifier learns how to distinguish one class from all the other classes.

Regardless of the strategy chosen to differentiate the base-models of the ensemble, different strategies might be exploited to combine the base-models outcomes into a single prediction [7]. For instance, this can be obtained by *weighting* each base-model's output according to a given rule, e.g., assigning a weight that is proportional to each base-model's predictive performance, or using the most voted class as the final classification outcome [11]. Another

[1] Department of Information Engineering, University of Pisa, Pisa, Italy.
[2] Bioengineering & Robotics Research Center E. Piaggio, School of Engineering, University of Pisa, Pisa, Italy.

$*$ These authors contributed equally to this work.

Correspondence to: vincenzo.catrambone@ing.unipi.it

This research has received partial funding from the European Commission H2020 Framework Programme under Grant No. 101017727 of the project "EXPERIENCE", and the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

well-known strategy leverages on the so-called *stacking* (or stacked generalization), i.e. the base-models' output is employed as an input for an higher-level model, the meta-classifier, that generates the final classification outcome [7]. Specifically, for EEG-based classification, stacking-based approaches have been effectively combined with various base-models' diversification strategies [12]. As an example, in [13] a MI classification approach based on stacking was used to combine the contributions of temporal, spatial, and spectral information of the EEG signals. In [14] the authors differentiate each base-model via multiple manipulation strategies, i.e. by changing the machine learning algorithm, feeding them with a different subset of features, and training them to recognize different pairs of classes. Finally, their outcomes are processed by a meta-classifier that generates the final classification. Many classification approaches based on *stacking* leverage the labels obtained by using a set of classifiers as base-models, whereas only a few of them take advantage of the richer information that may be generated by using regressors, which provide continuous values as outcome. One example of a stacking scheme combining regressors and classifiers is shown in [15], in which the authors propose a three layers classification architecture: the first layer consists of (base) regressors with a one-vs-one decomposition scheme, which feed a layer of (meta) regressors with a one-vs-all decomposition scheme, whose outcome is used by a (super) classifier in charge of providing the final classification outcome. The effectiveness of this solution comes with an additional computational cost due to the "extra layer" of ML models in the stacking.

In this regard, most of the ML tools exploited for BCI applications are difficult to be used in a real-time and real-world context due to their limitations in terms of classification performances or computational cost, e.g. a deep neural network with many links between nodes and layers [10].

To overcomes these limitations, in this study we propose a (non-deep) multilayer perceptron (MLP) architecture consisting of a first level of MLP (base) regressors with a one-vs-one decomposition scheme, above which we stacked a MLP (meta) classifier aimed at recognizing four different activities of a MI task by analyzing subjects' EEG. The proposed approach is tested on a publicly available dataset from the BCI competition IV (dataset 2a) [5], which comprises a four-class classification task associated with different MI activities; the features employed to feed the ML architecture are extracted from EEG recordings on healthy subjects. We show that the proposed approach outperforms previously evaluated models that have been tested on the same dataset.

II. Materials and Methods

A. Experimental dataset

The experimental dataset comprised of EEG recordings gathered from nine healthy volunteers during 4 MI tasks and resting phase. Tasks consisted of imagery movement of the right hand, left hand, tongue, and feet. EEG series comprised data from a 22-channel EEG system sampled at 250 Hz, and included 288 trials equally distributed into the 4 tasks. During each trial subjects were watching a screen sitting in a comfortable armchair; a first $2s$ period of resting state with a fixation cross on the screen was performed, then an arrow was projected for $1.25s$. The arrow pointed either to right, left, up, or down. Each direction was related to one of the 4 MI classes: right hand (RH), left hand (LH), tongue (T), or feet (F), respectively. Subjects performed MI task in the following $6s$. A schematic representation of the experimental procedure is presented in Figure 1. A comprehensive description of data acquisition is reported in [5].

B. EEG processing and feature extraction

EEG signals were band-pass filtered in the range $[0.5Hz-100Hz]$ and notch filtered at $50Hz$; then, the power spectral density (PSD) was extracted through the well-known Welch's method, with a window length of $4s$ and 75% overlap to minimize the estimation variance. For each EEG channel, the PSD was than averaged only considering the $6s$ of MI task and then filtered in 6 frequency bands, namely: $\theta \in (4-8]Hz$,$\alpha \in (8-12]Hz$, $\mu \in (12-16]Hz$, $(\alpha + \mu) \in (8-16]Hz$, $\beta \in (16-30]Hz$, and $\gamma \in [30-45]Hz$. Each trial was collapsed in a vector of 132 features given by 22 channels $\times 6$ frequency bands.

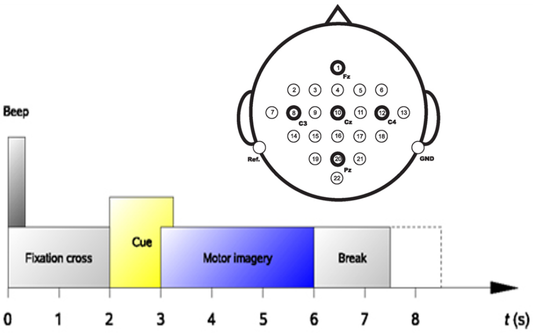

Fig. 1: A schematic representation of experimental setup including EEG channels and an MI task timeline.

C. Proposed Ensemble-based Neural Network Architecture

The proposed Ensemble-based Neural Network Architecture (ENNA) combines the stacking approach with a one-vs-one multi-class decomposition scheme. ENNA deploys the computation across two layers: the first one employs a set of regressors with a one-vs-one decomposition scheme, the second one consists of a meta-classifier. Both the regressors and the meta-classifier are implemented as multilayer artificial neural networks, so-called multi-layer perceptron (MLP). A schematic representation of the architecture is provided in Fig. 2.

The first layer of ENNA consists of six MLP regressors, one specialized to each pair of classes (e.g., RH vs LH) and trained by using (i) the instances belonging to the specific pair of MI classes (RH vs LH), and (ii) as desired output, 0 for the instances of the first class, 1 otherwise, as inputs. Specifically, each MLP regressor is trained with instances of PSD features belonging to two generic classes C_0 and C_1. At training instance j, the MLP regressor's outcome,

Hyper-parameter	MLP regressor	MLP classifier
Layers' size	[264, 32]	[264, 64, 16]
Activation	relu	relu
Solver	adam	adam
Max Epochs	8000	8000

TABLE I: Hyperparameters setting of the MLP regressors and the MLP classifier

$RO_j(C_0, C_1)$, is supposed to be closer to zero if $PSD(j)$ is more similar to the instances of C_0, or closer to one if $PSD(j)$ is more similar to the instances of C_1. As such, $RO_j(C_0, C_1)$ may be conceived as the probability that the sample $PSD(j)$ is more likely to belong to class C_0 than to class C_1 [15]. This property may not hold for instances of classes other than those used for MLP regressor's training, yet RO_j can be effectively employed as input for the meta-classification performed by the final MLP classifier [15].

The MLP classifier (MC) has to be trained with input including (i) the outcome of all the 6 MLP regressors (already trained) fed with PSD instances belonging to each of the 4 existing classes, and (ii) the desired output, i.e., the associated class. As an example, during MC training, the instance's feature PSD_j is given to all 6 MLP regressors. Their outcomes are eventually used as input by MC to predict the final $MI\ label$ associated with the instance j.

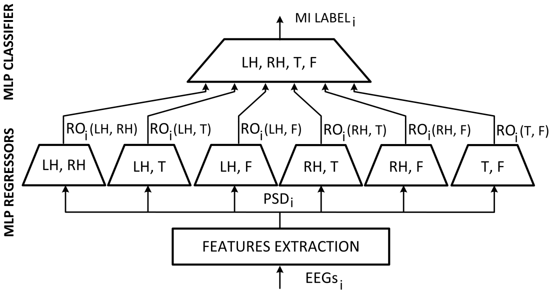

Fig. 2: Architecture of the proposed approach recognizing four MI classes: right hand (RH), left hand (LH), tongue (T), and feet (F).

Each architecture's module was built with *Python* by using well-known machine learning libraries, e.g. *sklearn*. Specifically, *sklearn* provides both MLP regressors and MLP classifier implementations. The employed hyper-parameters are detailed in Table I.

Performance evaluation: Classification results are evaluated in terms of Cohen's κ-coefficient [16], computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o measures the agreement among ground truth labels and predictions (i.e. the accuracy), and p_e measures the hypothetical probability of chance agreement. If $\kappa = 1$ ground truth labels and predictions are in complete agreement. If $\kappa = 0$ there is no agreement between ground truth labels and predictions other than what could be expected by chance.

III. Experimental Results

To ensure on the reliability of our results, we performed 100 repeated trials for each subject in the study. For each of these trials, the subject's data were randomly split: 90% of them are used to train the system, while the remaining 10% are used to test the performance of the proposed approach.

SUB. AVG±STD	1	2	3	4	5	6	7	8	9
[9] .66±.19	.77	.48	.83	.48	.60	.35	.86	.81	.79
[13] .70±.19	.83	.51	.88	.68	.56	.35	**.90**	.84	.75
[17] .66±.19	.80	.46	.82	.59	.38	.44	.81	.83	.81
[18] .73±.1	**.95**	.71	.78	.67	.63	.77	.69	.66	.74
[19] .68±.2	.69	.51	.87	**.85**	.78	.42	.54	**.97**	.45
[10] .81±.1	.92	.63	.86	.67	.81	.75	.86	.87	**.91**
[20] .71±.18	.87	.55	**.89**	.60	.58	.41	.88	.84	.80
ENNA **.85±.03**	.85	**.84**	.86	.81	**.84**	**.82**	.88	.88	.82

TABLE II: Comparison of the average κ score by subject

The ENNA approach results in an average $\kappa = 0.85$, which is 0.19 higher than the best result achieved during the competition [9], and 0.04 higher compared to the best of all methods proposed in more recent years. In fact, the BCI competition IV ended in 2012 with a maximum average accuracy across subjects of $\kappa = 0.66$ [9]. Several models have been proposed since then, outperforming such results; the proposed approaches widely differed in terms of ML architecture implemented, and the most performing ones, either in terms of subject-specific, as well as average across subject κ-values, are reported in Table II for further comparison.

In order to investigate how different MI tasks (i.e., left hand LH, right hand RH, tongue T, and feet F) and subjects affect the model performance, we consider the percentage of correctly classified instances. In Fig. 3 we show the 95% confidence interval of the correctly classified instances grouped by subject and MI task. The classification performance with LH class are consistently lower for all subjects, whereas classes RH and T show confidence intervals often including or even exceeding the 90% of correctly classified instances.

IV. Discussion and Conclusions

We presented a novel ensemble-based neural network architecture (ENNA), based on stacking and one-vs-one decomposition schema, to be exploited for BCI applications. ENNA was tested on real data from the publicly available BCI competition IV dataset 2a [5]; EEG data were gathered from 9 subjects performing four different MI tasks, namely, moving left hand LH, right hand RH, tongue T, and feet F. ENNA consists of a multi-layer perceptrons architecture featuring (i) a set of MLP regressors specialized to distinguish among each pair of MI classes, and (ii) a MLP classifier aimed at processing the output of the MLP regressors and recognizing the specific MI class. The ENNA architecture is fed with PSD features extracted from the 22-channel EEG recordings and filtered in the classical EEG frequency bands $\theta, \alpha, \mu, \beta, \gamma$, and a combination of $\alpha + \mu$ bands. Results were

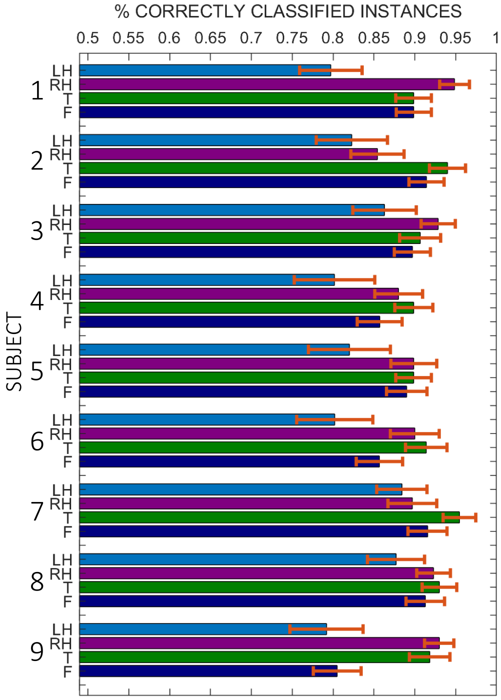

Fig. 3: ENNA approach. 95% confidence interval of the classification performance by subject and MI class: left hand (LH), right hand (RH), tongue (T), feet (F).

evaluated in terms of k-value [5].

ENNA performance clearly outperforms all approaches previously reported in literature for the BCI competition IV dataset 2a, considering average k-value [9], [10], [13], [18]–[20]. Of note, regarding subject-specific results, the ENNA approach reaches a higher κ in 3 out of 9 subjects, whereas none of the other methods has an equal or higher number of subjects excelling. Moreover, the performance of ENNA are characterized by the lowest standard deviation, thus meaning that the results are quite consistent over different subjects (no subjects has $k < 0.81$), in contrast with all other methods which perform poorly with at least one subject. Furthermore, other approaches lack performance consistency through subjects, i.e. always having some poorly performing subjects [9], [10], [13], [18]–[20]. This was generally explained as physiological subject-specific aspects for such MI classification task and was defined as ”BCI illiteracy” in literature [21]. Interestingly, our model was able to decrease the standard deviation of the performance across subjects by an order of magnitude.

The ENNA architecture was specifically designed for real BCI applications, in which non-computational-heavy models can enable quasi-real-time recognition [1]. Indeed, ENNA has lower model complexity compared to other approaches in Table II. The ENNA model features approximately 280k parameters (i.e. links between nodes of the neural networks). Compared to ENNA, the first runner up in Table II, [10], employs a deep learning architecture characterized by more than 30% additional parameters in the model.

Future developments will be directed to apply the presented model to finer MI tasks, comprising activity of daily living [22], and to perform quasi-real-time predictions in a large cohort of subjects.

References

[1] R. Abiri *et al.*, “A comprehensive review of eeg-based brain–computer interface paradigms,” *Journal of neural engineering*, vol. 16, no. 1, p. 011001, 2019.

[2] A. Alfeo *et al.*, “Measuring physical activity of older adults via smartwatch and stigmergic receptive fields,” in *The 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pp. 724–730, PRT, 2017.

[3] A. Alfeo *et al.*, “Sleep behavior assessment via smartwatch and stigmergic receptive fields,” *Personal and ubiquitous computing*, vol. 22, no. 2, pp. 227–243, 2018.

[4] V. Catrambone *et al.*, “Toward brain–heart computer interfaces: a study on the classification of upper limb movements using multisystem directional estimates,” *Journal of Neural Engineering*, vol. 18, no. 4, p. 046002, 2021.

[5] M. Tangermann *et al.*, “Review of the bci competition iv,” *Frontiers in neuroscience*, vol. 6, p. 55, 2012.

[6] M. Hosseini *et al.*, “A review on machine learning for eeg signal processing in bioengineering,” *IEEE reviews in biomedical engineering*, 2020.

[7] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[8] I. Goienetxea *et al.*, “Problems selection under dynamic selection of the best base classifier in one versus one: Pseudovo,” *International Journal of Machine Learning and Cybernetics*, pp. 1–15, 2021.

[9] K. Ang *et al.*, “Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b,” *Frontiers in neuroscience*, vol. 6, p. 39, 2012.

[10] R. Zhang *et al.*, “Hybrid deep neural network using transfer learning for eeg motor imagery decoding,” *Biomedical Signal Processing and Control*, vol. 63, p. 102144, 2021.

[11] M. Galar *et al.*, “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes,” *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

[12] A. Datta and R. Chatterjee, “Comparative study of different ensemble compositions in eeg signal classification problem,” in *Emerging Technologies in Data Mining and Information Security*, pp. 145–154, Springer, 2019.

[13] L. Nicolas-Alonso *et al.*, “Adaptive semi-supervised classification to reduce intersession non-stationarity in multiclass motor imagery-based brain–computer interfaces,” *Neurocomputing*, vol. 159, pp. 186–196, 2015.

[14] E. Abdulhay *et al.*, “Automated diagnosis of epilepsy from eeg signals using ensemble learning approach,” *Pattern Recognition Letters*, 2017.

[15] E. Menahem *et al.*, “Troika–an improved stacking schema for classification tasks,” *Information Sciences*, vol. 179, no. 24, pp. 4097–4122, 2009.

[16] M. Banerjee *et al.*, “Beyond kappa: A review of interrater agreement measures,” *Canadian journal of statistics*, vol. 27, no. 1, pp. 3–23, 1999.

[17] Q. She *et al.*, “Balanced graph-based regularized semi-supervised extreme learning machine for eeg classification,” *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2020.

[18] R. Das *et al.*, “Fbcsp and adaptive boosting for multiclass motor imagery bci data classification: A machine learning approach,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1275–1279, IEEE, 2020.

[19] L. He *et al.*, “Common bayesian network for classification of eeg-based multiclass motor imagery bci,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 6, pp. 843–854, 2015.

[20] S. Selim *et al.*, “A csp\am-ba-svm approach for motor imagery bci system,” *IEEE Access*, vol. 6, pp. 49192–49208, 2018.

[21] C. Vidaurre and B. Blankertz, “Towards a cure for bci illiteracy,” *Brain topography*, vol. 23, no. 2, pp. 194–198, 2010.

[22] V. Catrambone *et al.*, “Predicting object-mediated gestures from brain activity: an eeg study on gender differences,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 411–418, 2019.