ORIGINAL ARTICLE

EEG-based motor imagery recognition via novel explainable ensemble learning architecture

Antonio L. Alfeo[1,2] · Vincenzo Catrambone[1,2] · Mario G. C. A. Cimino[1,2] · Gaetano Valenza[1,2]

Received: 14 November 2024 / Accepted: 16 April 2025 / Published online: 17 May 2025

Abstract
Brain–computer interfaces (BCIs) are interactive machines using implicit neurophysiological signals, with applications ranging from medical rehabilitation to smart prostheses and entertainment. In this context, the need for high recognition performance demands increasingly complex machine learning (ML) architectures. Generally, the more complex the architecture, the less transparent its reasoning. This leads to difficulties in motivating their outputs and validating their internal model. Moreover, explainability is explicitly required by recent regulations on personal data processing, which advise against *black box* modeling. Here, a novel *ensemble learning* model is proposed aiming to effectively balance recognition performances and explainability. The proposed architecture employs different multilayer perceptrons, each one specialized to distinguish a single pair of classes and to provide *counterfactual explanations* and the minimal feature changes resulting in a classification shift. Subsequently, their outcomes are weighted to minimize the contribution of the non-competent classifiers and combined to address a multiclass classification problem. Results were gathered from two publicly available datasets on multiclass electroencephalography-based motor imagery and demonstrate that the proposed architecture overcomes state-of-the-art recognition performance while providing information on the most discriminant brain areas and power bands. For the sake of reproducibility, the implementation of the proposed approach is made publicly available.

Keywords Explainable artificial intelligence · Contrastive explanation · Multiclass decomposition · Motor imagery · Brain–computer interface · Ensemble learning

1 Introduction

Brain–computer interfaces (BCIs) are technological solutions aimed at interpreting neurophysiological signals to perform programmable actions, such as identifying movements [1, 16]. BCI systems find applications not only in medical devices but also in various fields like entertainment, education, and marketing [33, 52]. Among BCI systems, electroencephalography (EEG)-based BCIs are the most prevalent, designed to analyze and process cortical activity recorded from the scalp and translate it into interpretable end-effector information, such as categorical labels associated with predefined actions [1, 56]. This technological feat is enabled by factors like high time resolution and the feasibility of EEG recordings, alongside the understanding that imagination activates brain areas responsible for actual movements, and the utilization of machine learning (ML) algorithms for automatic knowledge extraction from physiological signals [13, 17, 34].

ML architectures are increasingly favored due to their recognition performance [5, 6, 38]. However, some drawbacks have been noted in this regard. Firstly, high-performing ML architectures often pose challenges for real-world applications due to their computational demands; for instance, deep neural networks require substantial computational power for training and operation [79]. To mitigate this, ensemble-based ML approaches, like OneVsAll and OneVsOne, have been found effective in balancing computational cost and recognition performance [10, 44]. Secondly, ML applications are often perceived as black boxes by the neuroscientific community, lacking explainability, and making it challenging for physicians and neuroscientists to understand and validate the algorithm's reasoning [70]. Explainability of ML models is crucial not only for supporting clinical decision systems but also for compliance with data processing regulations [8, 40]. Hence, there is a growing focus on designing explainable artificial intelligence (XAI) algorithms [35, 77].

However, the pursuit of explainability may compromise recognition performance. Complex and high-performance ML approaches tend to be less interpretable, while inherently explainable ML models are simpler but may struggle with real-world data intricacies [43, 64]. This trade-off between performance and explainability is well-recognized in the literature [43].

In this study, we aim to challenge this trade-off by introducing a novel XAI architecture for identifying different motor imagery (MI) activities using EEG data from healthy subjects. Our proposal utilizes a series of binary classifiers, implemented as multilayer perceptrons, arranged within a OneVsOne (OVO) multiclass decomposition framework. To maintain competitive recognition performance, we introduce a novel clustering-based mechanism to combine the outputs of the multilayer perceptrons, deriving the final prediction. Leveraging the OVO schema, each multilayer perceptron specializes in delineating a singular decision boundary between two specific classes (i.e., motor imagery). This specialization is leveraged when elucidating the motor imagery identified by the model with a single sample, targeting the minimal feature alteration on that sample necessary to induce a specific change in the recognized motor imagery. This approach does not aim to resolve the multi-class nature of the MI recognition task but rather lays the groundwork for generating explanations of the algorithm's decision logic. Indeed, our proposed approach delineates which features influenced its decision, elucidating whether the recognition outcome would shift based on the value of that specific feature.

To summarize, the proposed approach is characterized by:

- An improved ensemble architecture featuring a novel multiclass decomposition schema;
- A model-specific concise explanation generation protocol based on counterfactuals;
- A great trade-off between explainability and motor imagery recognition performance.

We show that the proposed approach outperforms previously evaluated models, which are tested on the same openly available benchmark datasets. Specifically, we applied the proposed algorithm to two distinct motor imagery EEG tasks. The first task involved imagining elementary movements of various peripheral systems, while the second task elicited higher-level motor processes associated with different types of object manipulation.

2 Background and related works

2.1 Ensemble learning

The ensemble-based approaches combine the outcomes of different ML algorithms, defined as base models, to obtain the final classification result [19]. Combining base classifier outcomes enables strong recognition performance because the error of a single base classifier is averaged with all the others. This allows using much simpler base classifiers, resulting in improved computational efficiency [63]. The taxonomy presented in [63] categorizes ensemble strategies according to the strategy to differentiate the base models. By *input manipulation*, the base models are differentiated by being assigned to different partitions of a dataset. With *learning algorithm manipulation*, the base models are differentiated by varying the hyperparameters of each base model or changing

the learning algorithm. With *output manipulation*, a multi-class classification problem is transformed into multiple binary recognition problems, and each base model deals with a single problem. This *divide and conquer* strategy is also called *multiclass decomposition schema*.

Using a multi-class decomposition schema can significantly enhance the efficacy of base classifier training [46].

The two primary decomposition schemas are *one-versus-all* and *one-versus-one* (OVO) [39]. In a C-multi-class recognition task with a *one-versus-all* decomposition scheme, each base classifier learns to distinguish one class from all others, resulting in C binary classification sub-problems. On the other hand, in a C-multi-class recognition task, an OVO schema yields $C \times (C-1)/2$ base classifiers, each trained to differentiate between two specific classes among all possible pairs of classes. Due to its superior recognition performance, the OVO decomposition schema is often preferred over a *one-versus-all* scheme [82]. Nevertheless, the robustness of the classification performance may be undermined by the increased number of base models in an OVO schema. This may cause different non-competent base classifiers to affect the recognition performance [37].

Specifically, a base classifier BC is non-competent for the classification of a sample s if the class to which s belongs does not match any of the classes used to train BC. Therefore, the outcome of BC is unreliable and unpredictable [36].

The exact method to find and isolate non-competent base models requires knowing *a priori* the class of the sample being classified, which, certainly, is unknown. To address this issue, several dynamic ensemble selection strategies have been proposed. For instance, a k-nearest neighbors (kNN) method can be employed to establish a region of competence around the sample being classified and aggregate the outcome of the base classifiers accordingly.

In more detail, dynamic ensemble selection strategies can be categorized as follows [21]: (i) *non–trainable*, having a low computational cost, but relying on strong assumptions about the classifier nature. An example is the majority voting strategy, which is effective under the assumption that all base classifiers are independent [28]; (ii) *trainable*, employing base model outcomes as the input for another ML model [20] that provides the final prediction; and (iii)*dynamic weighting*, where the outcomes of all base classifiers are weighted according to their expected local competence and subsequently aggregated to provide the final decision [81]. The latter offers the best trade-off between computational cost and recognition performance. For this reason, it will be employed in this research study.

2.2 Ensemble learning in BCIs

New and improved ensemble schemas are essential for optimal recognition performance, especially in complex recognition tasks like the ones with BCI data.

As an example, authors in [9] employ a multiclass decomposition scheme to select features for improved imagined speech recognition. In [55], the authors propose an ensemble approach for biometric recognition of individuals using EEG. The proposed solution provides great performance but uses a classic ensemble aggregation mechanism (i.e., a voting strategy). Ensemble-based approaches have been successfully applied in various BCI tasks [3], among which imaged movement recognition. For instance, in [59], authors employed a multiclass SVM based on the OVO schema to recognize ten different hand gestures from entropy features obtained from the EMG signal of 25 subjects, achieving an accuracy of 99.98%. In the context of imagined movements recognition, many recent works employ datasets such as the BCI dataset IV 2a, e.g., [78] and [47], leaving some room for improvement in terms of recognition performance (66.75 and 80.10 kappa score, respectively). To address this issue, other authors employ ensemble approaches to this recognition task. For example, the approach proposed in [75] provides better recognition performances (82.83 kappa score) using an ensemble representation of feature sequences. In [22], the authors propose an ensemble Flashlight-Net model aimed at processing different spectral components separately and then aggregating the results for final classification, achieving 81.23% accuracy. The authors in [24] demonstrate how the ensemble learning approach can provide good recognition performance on

the BCI dataset IV 2a, although the usage of a classical ensemble schema does not allow the final performance to be even comparable to other works in the literature. There is a need for improvements, e.g., a better aggregation mechanism for the ensemble schema, to ensure more accurate recognition. This research work addresses exactly this issue.

2.3 Explainable artificial intelligence (XAI)

The second problem related to the application of ML approaches in real-world scenarios is their black-box behavior. To address this issue, XAI approaches are conceived to provide classification labels supported by some explanations. The explanations can have different forms [51]: (i) *feature importance* or *feature attribution* is one of the most used explanation forms because many model-agnostic approaches generate a feature-ranking explanation [2]; (ii) *instance-based explanations* associate a labeled instance to some prototypes or counterexamples to trigger similarity-based reasoning in an end-user, thus providing human-friendly explanations [25]; (iii)*rule-based explanation* models summarize the decision process embedded in the algorithm, associating labels to the thresholds of the input features, and thereby compactly providing some insights about the behavior of the ML model using new instances [72]. Both rule- and instance-based explanations might be obtained using *counterfactual approaches* [72]. Counterfactual approaches aim to answer the question "why F rather than A?", where F is the fact to be explained and A is an alternative. From a classification viewpoint, this is equivalent to finding the minimal changes needed to shift the classification outcome of an instance to a specific alternative [67].

The closest counterfactual to a given data instance can be found by setting up a multiobjective optimization problem or leveraging the local neighborhood from the training set [73]. The latter provide fastest computation, and for this reason, the former approach is employed in this research study. Despite its popularity, this explanation form is particularly challenging with ensemble-based approaches since the decision to be overturned is one of the model group (or at least its majority) and not of one single model. We tackle this challenge by building an explanation procedure based on our improved multiclass decomposition schema.

2.4 XAI in BCIs

When applying ML models to brain-derived data, the primary focus has been on maximizing accuracy. Recently, however, the need for explainable results has grown, driven by legal requirements to ensure trust, privacy, security, ethics, and compliance with GDPR standards [40].

Understanding how an algorithm discerns neural activities offers valuable insights for a scientific purpose. Additionally, explainable responses can provide feedback, such as identifying errors in BCI systems caused by low activity in specific brain regions. This feedback can guide quality checks or support subject training in human-in-the-loop systems [30].

In supported diagnosis, ML algorithms predict health conditions or diseases. Explanations help understand diagnoses and identify factors that could influence such outcome (i.e., support to prognosis decision process).

It has to be clarified that: (i) using XAI approaches cannot assume that statistical causality is equivalent to understanding pathological behavior and (ii) XAI tools are typically only support systems and should never be used completely autonomously in diagnosis/prognosis problems [30].

An algorithm providing such broad and informative explanations has still to be designed; however, some solutions have already been proposed in this context. For instance, the authors in [58] propose an adaptation of SHAP [50] to a state-of-the-art DL network for inter-subject MI classification. They utilized the computed importance of each EEG channel to select an 8-electrode configuration, achieving inter-subject accuracies of 86.5% on the Physionet dataset and 88.7% on the Carnegie Mellon University's dataset. Similarly, the authors in [45] employed an ensemble of decision trees for EEG-based human activity recognition and explained their model using LIME [71]. This integration enables the provision of clinical reasoning behind the EEG spectral features in human activity recognition. Other models require conducting *relevance analysis*, focusing on the

feature relevance associated with the classification of a given instance (e.g., a subject or an action) [27]. All these solutions focus on a subject/instance level; however, they still maintain a class-specific approach and do not provide information on the differences between classes, which is the primary goal of our proposal.

3 Proposed approach

Here, a novel approach is presented, namely a NEighborhood-based Weighted Aggregation for eXplainable OVO decomposition scheme (NEWAXOVO).

3.1 Ensemble approach based on decomposition schema

The proposed OVO ensemble classifies EEG-derived features into MI classes by weighting and aggregating outcomes of different base classifiers, implemented as multi-layer perceptron (MLP) neural networks. As shown in Fig. 1a, the OVO ensemble requires each base classifier (*bc*) to process the instance (i.e., the EEG features *s*) and provide the probabilities for the classes on which *bc* was trained. As anticipated in Sect. 2, these probabilities are weighted according to a strategy specific to each OVO schema. Once all the weighted

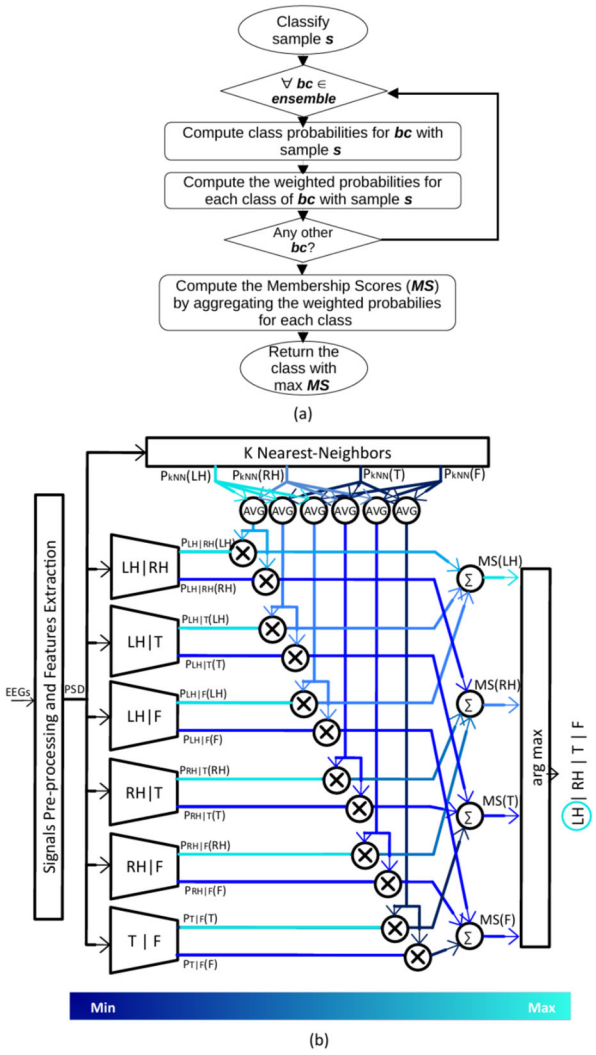

Fig. 1 **a** Overview of the classification procedure implemented by the OVO ensemble with a weighting schema. **b** NEWAXOVO architecture in the case of the BCI IV classification problem, i.e., a classification problem featuring four MI classes: right hand (RH), left hand (LH), tongue (T), and feet (F). The color scheme represents the values obtained by processing a sample of class LH of Subject 1. Each probability is bounded between 0 and 1, whereas MSs are positive unbounded, thus their rescaled values are considered.

probabilities are computed, those can be aggregated by class to obtain the class-wise membership scores (*MS*). The final classification of the ensemble is the class corresponding to the greatest *MS*.

More specifically, if the classification problem is constituted of n classes, the NEWAXOVO scheme consists of $n(n-1)/2$ MLP binary classifiers (Fig. 1b). For instance, with the BCI IV classification problem (i.e., 4 classes: right hand RH, left hand LH, tongue T, and feet F) NEWAXOVO will feature six MLPs as base classifiers. Each base model is trained with the instances of the PSD features belonging to a pair of classes; hence, it is specialized to distinguish them and provides as output the probability associated with these two specific classes.

A membership score $MS(C_i)$ is computed for each class, obtained by weighting the associated probabilities and subsequently aggregating them (see Eq. 1). Consequently, for a given sample s, the class predicted by NEWAXOVO corresponds to $C(s) = argmax(MS^s(C_i))$ among the $n = 4$ classes.

$$MS^s(C_i) = \sum_{j=1, j\neq i}^{n} \gamma_{i,j}^s * P^s_{[C_i \| C_j]}(C_i) \tag{1}$$

$$\gamma_{i,j}^s = \frac{P^s_{kNN}(C_i) + P^s_{kNN}(C_j)}{2} \tag{2}$$

The classification noise resulting from non-competent classifiers may result in similar *MS* values for different classes. In this case, small fluctuations in the *MS* values might lead to a classification error, due to the *argmax* operation. Consequently, the base model outputs are processed by a weighting operation affecting the base model contributions that are most likely to be non-competent, to minimize the *MS* of the wrong classes.

In Eq. 1, $P^s_{[C_i \| C_j]}(C_i)$ represents the probability that a given sample s belongs to class C_i, as the output of the base classifier trained on two classes C_i and C_j. The weighting factor, $\gamma_{i,j}^s$, (Eq. 2) exploits N_s, i.e., the close neighborhood of s from the training set. The competence of a specific base classifier trained on $[C_i \| C_j]$ is assessed as the average probability of classes C_i and C_j in N_s, obtained using the kNN classifier, as motivated in sec. 2.1. Specifically, the base models trained with the less frequent classes in the close neighborhood, N_s, are most likely to be non-competent, and thus, their outcomes are underweighted. This weighting operation propagates up to *MS*.

The presented architecture is a development of the preliminary proposal in [7]. In that study, a relatively simpler ensemble-learning architecture was designed and tested. Specifically, [7] leveraged a trainable ML approach for aggregating the base classifiers' outcomes. This resulted in a high computational cost and an impossibility of generating explanations using traditional methods because the approach consisted of two MLP layers. The present study aims both at reducing the computational cost and making the architecture explainable, without compromising recognition performance.

3.2 Explanation procedure

The explanation generation procedure aims to identify the minimum modification that a sample should endure to change the class it has been assigned to. In a classification problem, the set of features thresholds, and conditions that can lead to a different classification result is known as the *decision boundary*.

In NEWAXOVO, given a certain class (e.g., LH) and a counterfactual class (e.g., RH), the base classifier (*BC*) specialized in this decision boundary (LH vs RH) is the one most competent to analyze the minimal change needed for the classification outcome to shift. Indeed, trying to modify a sample starting from the most important features for that specific base classifier could be the fastest way to cross that specific decision boundary.

Intuitively, starting from a sample (e.g., a of class A) and a counterfactual class (e.g., class B), the proposed explanation procedure:

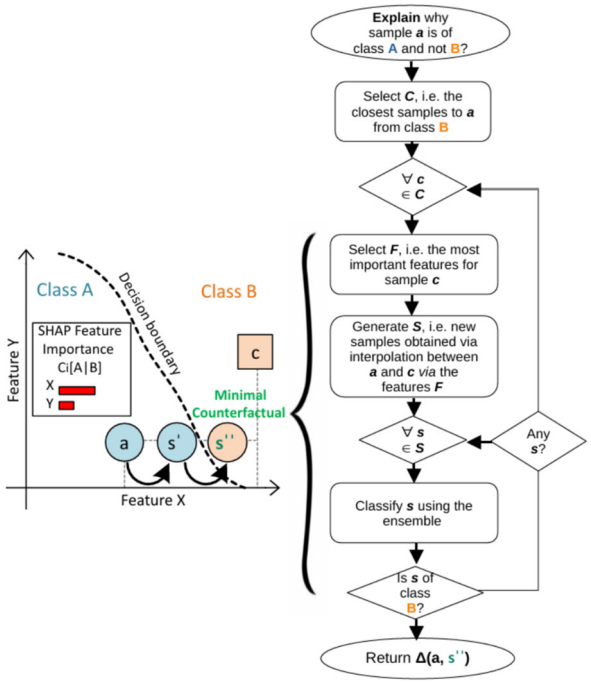

Fig. 2 Overview of the explanation procedure with a focus on the counterfactual search. The exemplified counterfactual search represents a recognition problem characterized by two features and just one of those is used for the interpolation

- Focuses on the base classifier trained to distinguish class A from class B, as it models the decision boundary between these two classes (dotted line in Fig. 2).
- Identifies the closest samples of class B to the sample a, since the proximity of two samples from different classes (e.g., samples a and c in Fig. 2) in the feature space suggests closeness to the decision boundary. The set of the closest class B samples to a is denoted as C in Fig. 2.
- Considers the most important features for the classification. To generate a counterfactual, the decision boundary must be crossed. Higher feature importance (obtained via SHAP) indicates that even small perturbations of that feature are more likely to cross the decision boundary.
- These perturbations are implemented using linear interpolation, applied to the most important features. In Fig. 2, the set of new samples (samples s' and s'') obtained via interpolation is denoted as S.
- Each new sample s generated through interpolation is progressively classified by the ensemble. When the classification changes (i.e., from class A to class B), the corresponding sample is considered a minimal counterfactual. This sample is used to build the explanation, i.e., the difference between the original sample a and the minimal counterfactual (Δ in Fig. 2).

Algorithm 1 NEWAXOVO explanation procedure. It provides the features' changes needed for s to be classified as C_c

Require:
$s \Leftarrow$ sample to explain
$C_s \Leftarrow$ predicted class
$C_c \Leftarrow$ counterfactual class
$k \Leftarrow$ considered # neighbors of s
$e \Leftarrow$ # steps to explore the feature space
$F \Leftarrow$ max # features to explore
$BC(C_c, C_s) \Leftarrow$ base classifier trained to distinguish classes C_s and C_c
Procedure:
1: FI(s) $\Leftarrow$ SHAP vector obtained with sample s and BC(C_c, C_s)
2: countFacts(C_c, s) $\Leftarrow$ samples $\in C_c$ among the k closest samples to s
3: **for each** $c \in countFacts(C_c, s)$ **do**
4: FI(c) $\Leftarrow$ SHAP vectors obtained with sample c and BC(C_c, C_s)
5: distFI(c) $\Leftarrow$ cosineDistance(FI(s),FI(c))
6: **end for**
7: **for each** $c \in countFacts(C_c, s)$ sorted by $distFI(c)$ **do**
8: MIF(c) $\Leftarrow F$ Most Important Features for c
9: **for each** $n \in range(1, F)$ **do**
10: CF(c,n) $\Leftarrow$ first n features from MIF(c)
11: **for each** $f \in CF(c, n)$ **do**
12: featSpaceInit $\Leftarrow$ min(value of f in s, value of f in c)
13: featSpaceEnd $\Leftarrow$ max(value of f in s, value of f in c)
14: step $\Leftarrow$ (featSpaceEnd - featSpaceInit)/e
15: featSpaceVals(f) $\Leftarrow$ values **from** featSpaceInit **to** featSpaceEnd **each** step
16: **end for**
17: newSamples(*) $\Leftarrow$ generate samples by using featSpaceVals(*) to replace features' values in s
18: distNewSamples(*) $\Leftarrow$ cosineDistance(s,newSamples(*))
19: **for each** $ns \in newSamples(*)$ sorted by $distNewSamples(*)$ **do**
20: classify(ns) $\Leftarrow$ compute classification of ns
21: **if** classify(ns) == C_c **then**
22: $explanation \Leftarrow$ % change between s and ns **for each** f
23: **break**
24: **end if**
25: **end for**
26: **end for**
27: **end for**
28: **return** $explanation$

In the proposed approach, an array of Shapley values is computed using the SHAP method as the importance of the features (*FI*). SHAP [50] is a game theory approach for explaining the output of any ML model. It produces a vector of Shapley values quantifying the contribution of each feature in a classification: A high value implies a large contribution recognizing a specific class. With binary classifiers (like our base classifiers), Shapley values are symmetric: a feature that strongly contributes to predicting a class, does it as well to predict another class, but with an inverted sign. Thus, in the presented procedure, the amplitude's absolute values are considered at lines 1 and 4 of Algorithm 1.

More specifically, the proposed explanation procedure (see Algorithm 1) selects, among the k samples closest to s, the samples belonging to class C_c, with $C_c \neq C_s$. It sorts those samples (c) according to the cosine distance between their *FI* vectors and the same vector computed for s. The *FI* is computed by employing the base classifier (*BC*) trained to distinguish classes C_s and C_c. Finally, a whole set of new samples is generated by manipulating a part of F-most important features (i.e., the ones with higher *FI*) of sample s. Such manipulation consists of progressively changing their values toward the ones assumed by the same features in the counterfactual sample c. These new samples (*ns*) are ordered with respect to the distance to s and subsequently classified by the NEWAXOVO scheme. As soon as NEWAXOVO changes the classification outcome, it is assumed that the minimum change to flip the classification is found. The value change of the F-most important features between s and *ns* (as a percentage) is reported as an explanation.

For the sake of reproducibility, the implementation of the proposed approach is made publicly available at [60].

4 Experimental setup

4.1 Datasets

The proposed approach was tested on two publicly available datasets, namely the BCI competition IV (dataset 2a) [69] and the U-Limb dataset (group A) [12].

The first dataset (hereafter Dataset1) comprises EEG series from ten healthy volunteers undergoing a standard MI protocol. The paradigm included four different MI tasks: the imagination of the left hand (LH), right hand (RH), both feet (F), and tongue (T) movement. For each subject, two sessions were recorded on different days. Each session consisted of six runs with short pauses in between. A session consisted of 288 trials, with 48 trials in each run, 12 for each of the 4 potential classes. A recording of approximately $5min$ was made at the start of each session to estimate the electrooculogram influence. The participants were seated on a plush recliner in front of a computer screen. A fixation cross was displayed on a dark screen at the start of a trial ($t = 0s$), and a short acoustic warning tone was played. A cue in the form of an arrow pointing left, right, down, or up (corresponding to one of the four classes: LH, RH, F, or T) emerged after 2 s ($t = 2s$) and lasted on the screen for $1.25s$. The subjects were subsequently encouraged to complete the necessary MI task, and no feedback was provided to them. At $t = 6s$, the subjects were instructed to complete the MI task until the fixation cross vanished from the screen, which subsequently went black for a short time. The EEG was recorded using 22 Ag/AgCl electrodes (with $3.5cm$ inter-electrode intervals). All signals were monopolarly recorded, with reference to the left mastoid and ground to the right mastoid. The sampling frequency was set as $250Hz$, and data were bandpass filtered between $0.5Hz$ and $100Hz$. The sensitivity of the amplifier was set as $100V$. A notch filter at $50Hz$ was employed to reject line noise. An electrooculogram channel was also recorded and exploited for artifact processing. An expert performed a visual review of all datasets, marking and rejecting any trials containing artifacts. A subject was excluded owing to the poor quality of the recordings. A schematic representation of the experimental procedure is presented in Fig. 3. A comprehensive description of the data acquisition is reported in [69], and the entire dataset is freely available online at *https://www.bbci.de/competition/iv*.

The second dataset (hereafter Dataset2) comprises EEG series gathered from 34 healthy subjects (all right-handed), which after an initial 5-minute resting state were asked to perform 30 different right upper limb movements, each repeated three times, resulting in a grand total of 90 different movement executions per subject. The 30 tasks were separated into three main classes: 10 *intransitive* tasks, movement not involving the use of an object; 10 *transitive* actions, a movement involving the use of a single object; and finally, 10 *tool-mediated* movements, e.g., action requiring to use an object as a tool interacting with another object (e.g., poor water from a bottle to a glass). All movements were carefully explained and mimed to the volunteers by an experimenter at the

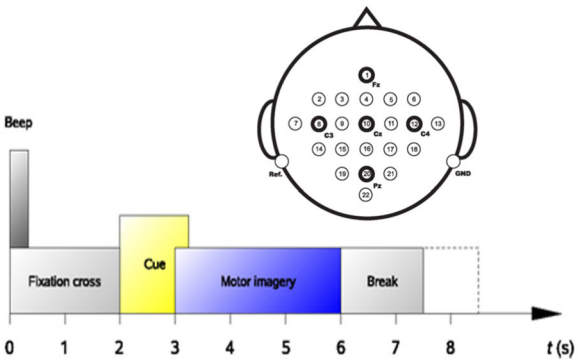

Fig. 3 Schematic representation of Dataset1 experimental setup including EEG channels and MI task timeline [69].

beginning of the experiment. The experimental procedure required participants to fulfill a 3-s motor imagery state and then to perform the actual imagined movement. In this study, to maintain the classification task in the motor imagery scenario, only the first 3-seconds imagery phase was analyzed. To discriminate the 3 classes of actions from the resting state, in which the subject was not moving or imagining a movement, the initial 5-minute resting state was divided into 30 non-overlapping segments. This allowed us to have 30 different repetitions of the resting state to be included in the classification process. Comprehensively, 30 trials for class and for each subject were recorded. For further details please refer to [12]. The EEG recording system used a $500Hz$ sampling rate and had 128 channels, but following results reported in [16], we kept a subset of 33 channels already used for classification. The preprocessing procedure employed for this dataset was analogous to the one implemented in [16]. Briefly, the EEG time series were subjected to a bandpass filter with a range of $0.5Hz$ to $100Hz$ to eliminate unwanted frequency components. Additionally, a notch filter at $50Hz$ was applied to remove line noise. To further ensure data quality, a semi-automated wavelet-enhanced independent component analysis (ICA)-based algorithm [32] was used for the rejection of physiological artifacts. Subsequently, an expert visually inspected all datasets, rejecting any trials that exhibited undetected artifacts.

4.2 EEG signal processing

4.2.1 Power spectral density

In the first dataset, EEG signals were segmented into time windows lasting $4s$ with $0.5s$ steps to minimize estimation variance. Similarly, in the second dataset, EEG signals were divided into time windows of $1.5s$ with $0.5s$ steps for the same purpose.

For each EEG channel, we extracted the power spectral density (PSD) using Welch's method and then applied filtering in six frequency bands: $\theta \in (<spanclass =' convertEndash' > 4 - 8 </ span >]Hz$, $\alpha \in (<spanclass =' convertEndash' > 8 - 12 </ span >]Hz$, $\mu \in (<spanclass =' convertEndash' > 12 - 16 </ span >]Hz$, $(\alpha + \mu) \in (<spanclass =' convertEndash' > 8 - 16 </ span >]Hz$, $\beta \in (<spanclass =' convertEndash' > 16 - 30 </ span >]Hz$, and $\gamma \in [<spanclass =' convertEndash' > 30 - 45 </ span >]Hz$.

Consequently, each trial produced four vectors containing 198 features, derived from the combination of 33 channels and 6 frequency bands.

4.2.2 Other EEG features

In this study, our aim is to provide an ensemble-based solution that balances explainability and recognition performance, rather than solely maximizing the latter. However, we acknowledge that the recognition performance of any model can be influenced by the choice of features used.

To evaluate the effect of substituting the PSD features, we maintained the same number of features (6 features $\times$ 22 electrodes). Specifically, the following EEG-based features were extracted:

- Hjorth's Parameters: activity, mobility, complexity [15, 76].
- Sample Entropy [62, 74].
- Signal's standard deviation.
- Total power as the sum of the squared values of the EEG within the considered time window [11].

4.3 Experimental setting

To prevent a suboptimal set of hyperparameters impacting the performance, an automatic hyper-parametrization of the NEWAXOVO scheme and its competing approaches, the distance-based relative competence weighting

combination OVO (DRCWOVO) and the unweighted OVO, was performed using OPTUNA [4]. OPTUNA employed 500 iterations, and the hyperparameters ranges are reported in Table 1.

Since the datasets analyzed here consisted of a 132-dimension feature space and the cosine distance is more robust while computing distances among samples with numerous features [41], any distance (e.g., between samples and between features importances) was implemented as a cosine distance. Furthermore, the kNN method weights each neighbor by its inverse distance with respect to the sample being classified; thus, close neighbors have a large influence on determining class probability. This choice has demonstrated working well with non-homogeneous feature spaces [21]. As per [80], in this study, the number k of instances considered by the kNN method was 3, 5, and 10.

In our experimental framework, we employed a stratified Monte-Carlo cross-validation scheme (Fig. 4). Specifically, when dealing with the data of one subject, 90% of it is randomly selected to build the training set, leaving the remaining 10% for the testing set. Its stratified composition ensures that the proportion of instances for each class (i.e., imagery movements) mirrors that of the original dataset. The training process for NEWAXOVO involves using the entire training set to train the KNN algorithm and portions of it (specifically, those pertaining to only one unique pair of imagery movements) to train each base classifier. The outputs of these individual components are then amalgamated to generate predictions for each instance within the testing data presented to NEWAXOVO. By comparing these predictions with the ground truth, we can quantify the recognition performance for the ongoing iteration of cross-fold validation. This iterative process is repeated ten times for each subject, and the results are subsequently aggregated and presented in terms of both mean and variance.

4.4 Performance evaluation

The proposed NEWAXOVO methodology was evaluated considering the recognition performance, architecture complexity (e.g., number of trainable parameters), and readability of the proposed explanations. A key component introduced by NEWAXOVO is a novel OVO aggregation mechanism based on KNN clustering. In our experiments, we assess the impact of this specific component on the model by utilizing the exact same MLP ensemble and substituting the NEWAXOVO aggregation mechanism with the"classic"OVO voting mechanism (i.e.,"Unweighted OVO"), or with other aggregation mechanisms proposed in the literature, such as the one proposed in DRCWOVO [37]. The only distinction among these three architectures is the aggregation of base classifier outputs. Consequently, their comparison enables us to evaluate the specific contribution of our proposal to the model's effectiveness.

In short, the DRCWOVO model weights the base classifier outputs classifying a sample s according to its distance to the k nearest neighbors taken from each class [37]. The effectiveness of the DRCWOVO method has been extensively tested using different datasets and in conjunction with other schemas, outperforming other approaches [80].

As mentioned in Sect. 2.1, a limitation in the application of the OVO decomposition schemas is that the classification noise generated by the non-competent classifiers may propagate up to the final prediction, i.e., increasing the *MS* value for the wrong classes (see also sec. 3.1). To assess how the weighting operation

Table 1 Hyperparameters ranges of MLP base classifier

Hyperparameters	*Values*
First Layers size	[64, 128]
Second Layers size	[32, 64]
Activation Function	identity, ReLU, tanh
Optimizer	lbfgs, adam
Batch size	32, 64, 128, 256
Optimization tolerance	loguniform(1e-4, 1e-6)

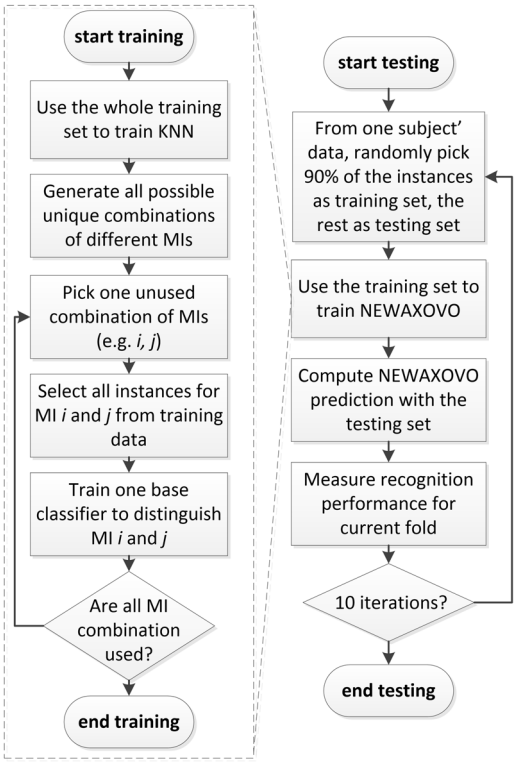

Fig. 4 Flowchart of the training and testing procedures for NEWAXOVO

purposely affects the contribution of the non-competent classifiers, minimizing the classification noise, *competence reinforcement* (CR) is defined in Eq. 3.

$$CRs = \frac{MS^s(C_s)}{\sum_{j=1}^{\#classes} MS^s(C_j)} \tag{3}$$

CR for sample s is defined as the ratio of the MS^s of the true class C_s (i.e., $MS^s(C_s)$), and the sum of all membership score values associated with s (i.e., belonging to all four classes). *CR* is considered to be equal to one in the ideal case in which all erroneous classes *MS* are null, i.e., *CR* equal to the ratio of $MS^s(C_s)$ and itself. In the worst-case scenario, the correct class has a null *MS* and null *CR*. A high *CR* value, close to one, implies a low membership score associated with the wrong classes and a low classification noise due to non-competent classifiers. To illustrate, in Fig. 5 we schematically depict how a OVO ensemble processes a real-world instance (from Dataset 1, Subject 1) to recognize the corresponding imagined movement (i.e., RH). We report the actual values of all probabilities generated by the base classifiers and the membership scores as per the schema presented in Fig. 1b. Comparing the OVO ensemble with no weighting (a) and NEWAXOVO (b), the latter ensures a clear reduction in the classification noise, i.e., the membership scores of the classes others than the correct one. Those are indeed all reduced to 0.01, and this results in greater *CR*.

Finally, the recognition performance is evaluated in terms of Cohen's κ coefficient [14], as reported in Eq. 4.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

where p_o measures the accuracy in terms of the agreement between the ground truth and the labels given by the classification, and p_e measures the hypothetical probability of chance agreement. $\kappa = 1$ implies that the classification outcomes equal the ground truth labels. $\kappa = 0$ is associated with a classification by chance.

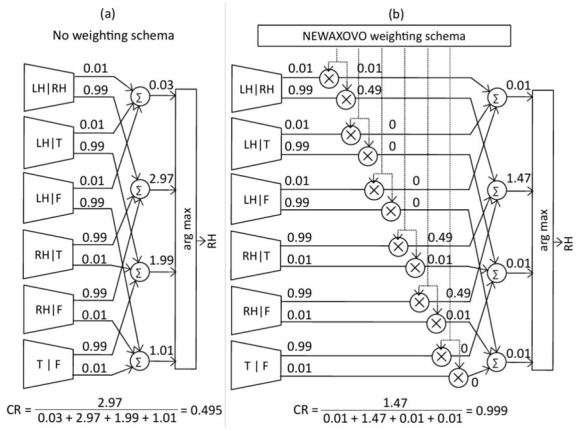

Fig. 5 Effect of the weighting schema exemplified with an instance of class RH from Dataset1, Subject 1. Following the schema presented in Fig. 1b, we report all the values of the probabilities generated by the base classifiers and the membership scores. If compared to the OVO ensemble with no weighting schema (a), the proposed approach (b) decreases the membership score of classes different from the correct one, thus resulting in greater *CR*.

To assess the impact of the informativeness of the different feature groups presented in sections 4.2.1 and 4.2.2 we employed four shallow classifiers provided by *scikit-learn* [57] (with default hyperparameter values) for each of these approaches:

- Gaussian Process Classifier (GPC) [18], a kernel-based ML approach similar to Support Vector Machine, aimed at predicting highly calibrated class membership probabilities.
- K-Nearest Neighbors Classifier (KNN) [68], a ML algorithm based on the idea that similar objects should be close in the feature space. It outputs the class label that is most frequently represented among the K neighbors of the analyzed sample.
- Multi-layer Perceptron (MLP) [61], a ML approach based on fully connected neural networks.
- Tree Classifier (TREE) [31], a ML approach that constructs a tree-like model of decisions and predicts the class of an input based on a series of binary decisions, such as whether a given feature value is above or below a certain threshold.

5 Results

Experimental results are here presented first in terms of classification accuracy, where the robustness of the proposed approach is validated on both Dataset1 and Dataset2, and then as explanations readability, where exemplary explanations are provided to clarify how they can be useful in practical applications.

5.1 Classification performance

5.1.1 Dataset 1: BCI Competition IV

Table 2 reports the mean κ score values obtained from Dataset1. The κ scores are reported for a single-subject and group-wise (average $\pm$ standard deviation across subjects). The highest κ values are highlighted in bold. NEWAXOVO is compared to state-of-the-art ML approaches that were applied to the same dataset. In addition, alternative OVO approaches, including unweighted OVO method (i.e., formally MS^s calculated as in Eq. 1, with $\gamma^s_{i,j}$ always equal to 1) and a weighted OVO approach, i.e., the DRCWOVO method [37], were compared as well.

Table 2 Recognition performance for ML approaches on Dataset1. Performance is expressed as number of neighbors (#nn), group-wise κ score (mean $\pm$ std), and κ per subject (column from 1 to 9)

Approach	#nn	Group κ	1	2	3	4	5	6	7	8	9
ANN1 [29]	–	0.554 ± 0.186	0.62	0.57	0.89	0.27	0.59	0.34	0.75	0.56	0.38
Shallow ConvNet [53]	–	0.567 ± 0.194	0.73	0.42	0.85	0.75	0.43	0.38	0.89	0.75	0.72
She et al. [66]	–	0.66 ± 0.08	0.797	0.460	0.819	0.593	0.382	0.438	0.811	0.828	0.811
He et al. [42]	–	0.66 ± 0.13	0.69	0.51	0.87	0.85	0.78	0.42	0.54	**0.97**	0.45
Ang et al. [10]	–	0.663 ± 0.065	0.769	0.475	0.834	0.484	0.601	0.347	0.862	0.807	0.788
Yu et al. [78]	–	0.667 ± 0.134	0.767	0.562	0.853	0.53	0.656	0.454	0.822	0.689	0.672
Lu et al. [49]	–	0.69 ± 0.079	0.84	0.65	0.78	0.64	0.65	0.56	0.64	0.73	0.70
Nicolas et al. [54]	–	0.70 ± 0.19	0.83	0.51	0.88	0.68	0.56	0.35	0.90	0.84	0.75
Selim et al. [65]	–	0.714 ± 0.067	0.874	0.551	0.893	0.604	0.580	0.410	0.875	0.837	0.801
EEGNet [53]	–	0.718 ± 0.164	0.82	0.49	0.91	0.64	0.69	0.45	0.87	0.83	0.77
Das et al. [23]	–	0.73 ± 0.10	**0.95**	0.71	0.78	0.67	0.63	0.77	0.69	0.66	0.74
TCNet–Fusion [53]	–	0.777 ± 0.131	0.87	0.60	0.93	0.68	0.76	0.58	**0.92**	0.85	0.81
EEG–TCNet [53]	–	0.784 ± 0.123	0.86	0.63	**0.97**	0.68	0.78	0.61	0.91	0.82	0.80
Lian et al0. [47]	–	0.80 ±.08	0.87	0.62	0.80	0.83	0.80	0.71	0.91	0.81	0.86
Zhang et al0. [79]	–	0.81 ± 0.10	0.92	0.63	0.86	0.67	0.81	0.75	0.86	0.87	**0.91**
Unweighted OVO	–	0.835 ± 0.042	0.859	0.848	0.900	0.799	0.823	0.839	0.831	0.867	0.751
	3	0.866 ± 0.040	0.885	0.860	0.919	0.855	0.855	0.847	0.849	0.891	0.831
DRCWOVO [37]	5	0.858 ± 0.042	0.890	0.853	0.914	0.860	0.844	0.838	0.828	0.889	0.803
	10	0.832 ± 0.049	0.845	0.842	0.891	0.835	0.805	0.824	0.807	0.852	0.786
	3	**0.879 ± 0.035**	0.891	0.848	0.911	**0.901**	**0.869**	**0.901**	0.822	0.907	0.856
NEWAXOVO	5	0.875 ± 0.037	0.913	**0.862**	0.917	0.888	0.849	0.863	0.842	0.889	0.855
	10	0.857 ± 0.039	0.868	0.841	0.920	0.850	0.836	0.874	0.831	0.868	0.827

The NEWAXOVO model showed an average $\kappa = 0.879 \pm 0.035$, which is 0.13 higher than the DRCWOVO implementation and 0.216 higher than the best result achieved during the BCI competition [10]. Both NEWAXOVO's and DRCWOVO's performance reach a maximum with $K = 3$ and tend to decrease as K increases. In terms of intra-subject κ score, NEWAXOVO shows the best results for four out of nine subjects.

As evident from Table 2, approaches [66] and [29] provide the lowest average recognition performance and wide variability among different subjects. Also, [29] provides the worst performance among all the studied included in Table 2 with subject 4. Inter-subject variability is also observed in [42] and [10]. Among the approaches considered, [42] obtains the worst performance for subjects 7 and 9, whereas [10] yields the worst performances for subject 6. Conversely, [42] achieves the best recognition performance for subject 8, and [66] provides average k-scores above 0.8 for subjects 3, 7, and 8, outperforming all OVO-based ensemble approaches in the case of subject 7. The average recognition performance provided in [78] falls in the middle of the overall group despite how recent this approach is. Comparatively, [47, 54, 65], and [23] exhibit reduced performance fluctuation between subjects. Among them, [54] demonstrates the best performance for subject 7 and [23] for subject 1. In [79], the authors employ a deep neural network approach combined with a multi-class decomposition approach. Although the strategy differs from ours, this approach shows the closest performance to the OVO-based ensemble approaches, achieving recognition performances above 0.8 for almost all subjects analyzed. Subject 5 demonstrates the worst average k-scores among all the considered approaches, while subjects 7, 8, and 9 consistently exhibit higher average k-scores above 0.8.

To perform a comprehensive comparison relative to the state-of-the-art, Table 2 also showcases the recognition performance achieved with four recent deep learning solutions applied to MI recognition [53]. Among these approaches, Shallow ConvNet yields the lowest average performance, exhibits the highest variance, and produces

the poorest result for subject 2. EEGNet, on the other hand, demonstrates a notably superior average performance compared to Shallow ConvNet, positioning itself in the middle among all average performances listed in Table 2. Both TCNet-Fusion and EEG-TCNet lead to reduced performance fluctuations across subjects, resulting in the best recognition performance for subjects 7 and 3, respectively.

Considering the performances shown in Table 2 on a subject-by-subject level, subjects 3 and 8 consistently exhibit average recognition performances above 0.8 with all the considered approaches, except for [23]. In contrast, subjects 2 and 6 yield average performances below 0.6 when considering only the approaches from the literature. With these subjects, the adoption of OVO-based ensemble approaches consistently provides a significant improvement in terms of recognition performance.

Finally, although not included in Table 2 due to the absence of subject-wise performance measurements, we highlight that the proposed approach outperforms three other recent approaches that performed the same recognition task on this dataset, i.e., [47, 78] and [75] providing 0.667, 0.801, and 0.828 group-wise kappa score, respectively.

Figure 6 shows the confusion matrix of recognition performance obtained using NEWAXOVO with a parametrization of K equals to 3. On the main diagonal, the percentage of correctly classified instances is displayed for each class, whereas the off-diagonal boxes report the percentage of instances of class i (i-th column) mistakenly recognized as instances of class j (j-th row). Overall, the classes labeling the imagined movements of hands (i.e., RH and LH) are the most frequently misclassified, with false detection ranging between 4.55 and 5.03%, whereas the best-recognized class is associated with the feet imagined movement. According to the results in Fig. 6, the worst misclassification occurs between LH and RH imaged movements. This result may be due to an inherent difficulty of the analyzed dataset. Indeed, this pair of classes results to be hard to distinguish also in other recent works analyzing the same dataset, as evident from the confusion matrices reported in [29]. Similarly, in [26], the authors report the average recognition performance of an ensemble of OVO binary classifiers, exactly showing that distinguishing LH from RH results in the worst performance. This can also be explained physiologically considering that LH and RH refer to movements of the left and right hand, which differ in terms of lateralization, but are equal (symmetrical) in terms of actual movements and associable meaning [48]. Indeed, the other classes, i.e., class F and class T which refer to imagined movements of the feet and tongue, respectively, recruit different populations of neurons for the completely different motor units needed and for the abruptly different movements' communicative and intrinsic meaning. For these reasons, it is authors' opinion that LH and RH are the most similar classes in terms of actual underlying neural activity, and this is reflected in a slightly higher misclassification rate.

A well-known issue with OVO schemas is that each base classifier is trained only on a pair of classes. Thus, if an instance of a different class is presented, the classifier can be considered non-competent, as its prediction is essentially random. This increases the likelihood of a wrong class being considered correct by the ensemble, as the ensemble's decision aggregates outputs from both competent and non-competent classifiers. Since it is impossible to know in advance which classifier is non-competent for a new instance, some OVO schemas try to estimate the most likely non-competent classifiers and minimize their influence on the ensemble's final decision. If this occurs, the membership score of the correct class (*MS* in Fig. 1) should be maximized, while those of

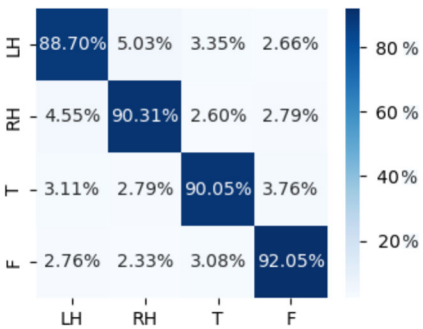

Fig. 6 Confusion matrix obtained using NEWAX-OVO (with $K = 3$) and all subjects of Dataset1. The motor imagery (MI) is the right hand (RH), left hand (LH), tongue (T), and feet (F).

incorrect classes should be minimized. To measure this occurrence, we use the Competence Reinforcement (CR, defined in Section 4.4), i.e., the ratio of membership scores coming from the correct class (the higher, the better).

Table 3 presents the average CR values obtained with all subjects using three OVO approaches. Unweighted OVO does not apply any strategy to reduce the impact of non-competent classifiers and serves as the performance baseline. DRCWOVO and NEWAXOVO, instead, do apply weighting strategies to mitigate the impact of non-competent classifiers, resulting in greater CR values. Specifically, the CR achieved by Unweighted OVO is 0.467, which is slightly lower than the one obtained with DRCWOVO (0.544) and significantly lower than the CR of NEWAXOVO (0.794). These results prove the importance of using a weighting strategy to minimize the impact of non-competent classifiers and ensure more robust recognition performance. More importantly, the weighting strategy proposed in this study outperforms DRCWOVO.

We evaluated the computational complexity of the proposed approach in terms of the number of trainable parameters. With the chosen hyper-parameterizations for NEWAXOVO, the total number of trainable parameters is approximately 63700. To contextualize this computational complexity within the state of the art, we offer a comparison of the number of trainable parameters with respect to the state-of-the art.

Compared with NEWAXOVO, the approaches analyzed in [53] offer reduced computational complexity (3 to 4 times less), albeit with a performance loss ranging between 9% and 17%. On the other hand, comparing the proposed approach with the most recent approaches included in Table 4, a considerable increase in the number of trainable parameters can be observed, i.e., 183% for [29], 570% for [47], and 6518% for [22]. Also considering all the other approaches, those exhibit much higher computational complexities (from 439% to 2326% more), despite a recognition performance up to 19% smaller. DRCWOVO and Unweighted OVO share the same number of trainable parameters as NEWAXOVO. The only difference between them and NEWAXOVO lies in the way the outputs of the base classifiers are aggregated. Although this does not impact the number of trainable parameters, it can affect training times since the aggregation mechanism is initialized during the training phase.

To quantitatively assess the impact of the proposed weighting schema on training time, we provide subject-wise statistics (expressed as mean and standard deviation) for training time on a machine featuring an Intel Core $i5-10210U$ CPU @ 1.60 GHz−2.11 GHz with 8 GB RAM. The training time for Unweighted OVO was 26.29 ± 3.43 seconds, while for NEWAXOVO it was 26.16 ± 3.20 seconds. The only difference between them is the single usage of KNN. Since KNN is a"lazy learner,"these approaches result in comparable training times. DRCWOVO, however, requires slightly longer training time, approximately 30.09 ± 2.65 seconds. This difference can be attributed to a key distinction between NEWAXOVO and DRCWOVO which lies in their respective procedures for searching nearest neighbor samples to weight the outcomes of the base classifiers. NEWAXOVO needs K nearest neighbor samples in total, while DRCWOVO requires K nearest neighbor samples for each class in the recognition problem. To accomplish this class-wise search, a KNN instance is employed for each class. At prediction time, these individual KNN instances are queried to find the K nearest neighbor samples for their corresponding classes. It is important to note that this class-wise search does impact the number of operations required by DRCWOVO.

To assess the impact of using features other than PSDs, we extracted additional EEG-based features from Dataset1 and processed the resulting samples using different ML approaches, including NEWAXOVO. Table 5 presents the results obtained using EEG features listed in subsection 4.2.2. Notably, all tested approaches

Table 3 Dataset1. Average ± standard deviation of competence reinforcement. NEWAXOVO and DRCWOVO feature k = 3 in kNN procedure

Architecture	Competence reinforcement
Unweighted OVO	.467 ± .077
DRCWOVO [37]	.544 ± .108
NEWAXOVO	**.794 ± .235**

Table 4 Number of trainable parameters and comparison with respect to the state-of-the art

Approach	# Trainable parameters
EEGNet [53]	15.6 k
TCNet-Fusion [53]	17.5 k
EEG-TCnet [53]	20.5 k
NEWAXOVO	63.7 k
DRCWOVO [37]	63.7 k
Unweighted OVO	63.7 k
ANN1 [29]	116.6 k
Alfeo et al. [7]	280 k
Zhang et al. [79]	363 k
Lian et al. [47]	392.5 k
Lu et al. [49]	1.482 M
Dang et al. [22]	4.152 M

Table 5 Group-wise κ score (mean $\pm$ std) obtained via different machine learning algorithms and NEWAXOVO (with k = 3). Dataset1 is processed by extracting the PSD features or other 6 different EEG-based features

Features	PSDs	Other EEG features
GPC	0.529 $\pm$ 0.075	0.200 $\pm$ 0.047
KNN	0.588 $\pm$ 0.073	0.432 $\pm$ 0.137
MLP	0.535 $\pm$ 0.079	0.177 $\pm$ 0.045
TREE	0.428 $\pm$ 0.071	0.307 $\pm$ 0.060
NEWAXOVO	0.879 $\pm$ 0.350	0.746 $\pm$.097

experienced performance improvements by incorporating PSD features, with gains ranging from.133 (NEWAXOVO) to.358 (MLP). Remarkably, our proposed approach consistently outperformed the others across various feature sets, especially when compared to MLP. The fact that our architecture outperforms NEWAXOVO, which is an ensemble of MLPs, further validates the significance of our approach in enhancing recognition performance.

As per the prevailing scientific literature and to maintain consistency, next we utilize PSD features for the analysis of Dataset2.

5.1.2 Dataset 2: U-Limb

The average κ score obtained subject-wise, as well as the CR, are listed in Table 6. In contrast with Tables 2, 6 does not display a measure for subject-wise recognition performances. Dataset2 comprises data from 34 subjects, and the recognition performance consistently approaches or equals 1 for all subjects, leaving little room for further subject-wise considerations. In terms of recognition performance, the simple OVO approach already achieves good performance (.996 average κ score), leaving little room for significant improvement. Indeed, both

Table 6 Performance metrics on Dataset2 reported as average $\pm$ standard deviation among subjects. NEWAXOVO and DRCWOVO use k = 3 features in the kNN procedure

Architecture	κ score	*Competence Reinf.*
Unweighted OVO	.996 $\pm$ .007	.499 $\pm$ .001
DRCWOVO [37]	.996 $\pm$ .009	.584 $\pm$ .027
NEWAXOVO	.995 $\pm$ .013	**.979 $\pm$ .040**

DRCWOVO and NEWAXOVO offer comparable performance, resulting in minor differences that may be due to pseudorandom fluctuations in the training procedures on repeated trials. Note that high recognition performances were already reported in previous studies [16, 17].

On the other hand, the average CR value for Unweighted OVO (i.e., the baseline performance) is 0.499, which is slightly lower than the one obtained with DRCWOVO (0.584) and significantly lower than the one obtained with the proposed approach (0.979). These results confirm the effectiveness of the proposed OVO scheme and its weighting strategy in reducing the impact of non-competent classifiers. Indeed, the proposed approach not only results in an average CR value nearly double the one of Unweighted OVO, but also close to 1, i.e., the maximum possible CR value.

The significance of the recognition obtained through NEWAXOVO (with K=3) on Dataset2 is confirmed by the corresponding confusion matrix shown in Fig. 7. Each class is correctly recognized over 99% of the time. Overall, misclassifications occur with a frequency lower than 0.1% with the exception of the instances of class IN misclassified as class RE and instances of class TM misclassified as class TR.

5.2 Explainability

Here, we show the explainability of the proposed approach by using as a case study the analysis of Dataset1. Specifically, Fig. 8 illustrates counterfactual explanations provided by the NEWAXOVO. The explanation details the minimum change in features (i.e., EEG channel and frequency band) of a given sample to change the classification result. Such changes are represented as percentage variation positive (green) or negative (red).

More in detail, the explanation shown in Fig. 8.a provides insights on why sample #84 of subject 1 of Dataset1 was recognized as LH instead of RH, i.e., it identifies the minimum variation resulting in the former classification outcome. On the other hand, the explanation shown in Fig. 8.b is obtained with sample #8901 of subject 9 of the Dataset1. Such an instance was misclassified as LH, instead of F class. The minimum variation identified for this sample to be correctly recognized as F is shown. This information might be useful to understand the reason for the misclassification, localizing the issue.

Complementary to the counterfactual instance-specific explanations in Fig. 8, the proposed algorithm can provide aggregated information at class and subject levels. Figure 9 is a graphical counterfactual representation of the most important features that need to change in order to switch the classification outcome from class LH to RH with subject 7 of Dataset1. The 132 input features are represented as electrodes on a scalp and divided by the frequency band (i.e., θ, μ, α, $\mu + \alpha$, β, and γ). To trigger the class shift, a variation in the PSD obtained from each electrode-band combination must occur (see positive variations in green and negative variations in red). When colored, the radius of each electrode is directly proportional to the percentage of instances associated with the subject that would switch the classified label, if that feature shows a variation.

The explanations generated by NEWAXOVO can be used to analyze the classification uncertainty since the percentage variations can be used as a proxy to estimate the distance between the decision boundary and the border instances (i.e., those that require minor changes to switch the classification results). The closer a sample is to the decision boundary, the easier the misclassification to occur. In Table 7, the absolute value of the median

Fig. 7 Confusion matrix obtained using NEWAX-OVO (with K=3) and all subjects of Dataset2. The motor imagery (MI) categories are rest (RE), intransitive (IT), transitive (TR), and tool-mediated (TM)

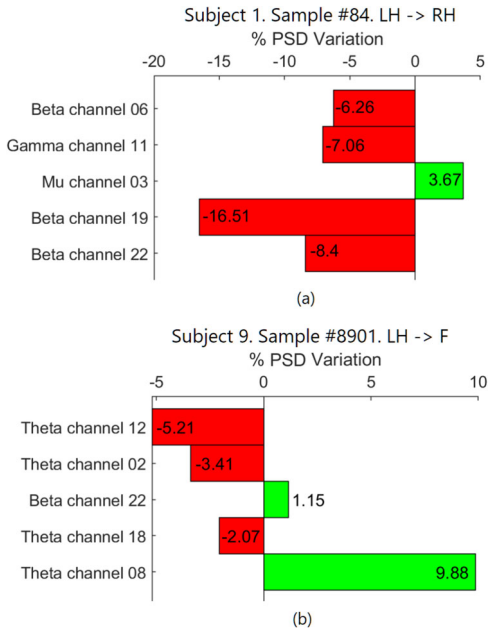

Fig. 8 Exemplary counterfactual explanations provided by NEWAXOVO. The explanation details the minimum change in the features (EEG channel and frequency band) of a given sample to shift the resulting classification. The relative change can be positive (green) or negative (red). (a) Sample #84, belonging to subject 1, was correctly recognized as LH. The minimum variation identified for this sample to be recognized as RH is shown. (b) Sample #8901, belonging to subject 9, was misclassified as LH, instead of F. The minimum variation identified for this sample to be correctly recognized as F is shown. Both subjects are from Dataset1.

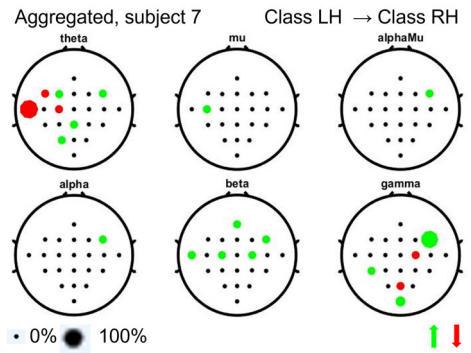

Fig. 9 Exemplary explanation of the switch from class LH to class RH for subject 7 of Dataset1. To trigger the class shift, a variation in the PSD obtained from each electrode-band combination must occur (see positive variations in green and negative variations in red). When colored, the radius of each electrode is directly proportional to the percentage of instances associated with the subject that would switch the classified label, should that feature show a variation

Table 7 Aggregated PSD % variation needed to switch classification result with subject 7 of the BCI competition IV dataset

From \ To	LH	RH	T	F
LH	-	10.86	17.35	12.03
RH	9.64	-	13.37	18.22
T	18.31	13.43	-	15.94
F	11.36	16.58	18.05	-

percentage variations with subject 7 of Dataset1 is reported. Those are obtained via the explanation method of NEWAXOVO (detailed in algorithm 1). Note that LH and RH result as the classes whose samples are generally closer to the decision boundary. In addition, LH and RH are also the classes that are more often misclassified. Specifically, with subject 7, 8.6% of LH class instances were misclassified as RH, while 11.7% of RH class instances were misclassified as LH, confirming the trend highlighted at an aggregated level (i.e., considering all the subjects) in Fig. 6.

6 Discussion

We proposed a novel ML architecture, the NEWAXOVO, based on ensemble learning and XAI. NEWAXOVO comprises different MLPs specialized in discerning specific pair of classes and an aggregation scheme aimed at processing the outputs of the base models and recognizing a specific class. The binary base models are specifically employed to generate counterfactual explanations since those are specialized on the decision boundary between each couple of classes. The NEWAXOVO scheme was tested on publicly available EEG data, namely the BCI competition IV dataset 2a [69], which is a benchmark dataset, including 22-channel EEG from nine subjects performing four-class motor imagery, and the U-Limb dataset [12], including 128-channel EEG data collected from 34 subjects performing three-class motor imagery plus resting state. Remarkably, the two datasets comprised very different motor imagery classes, engaging distinct neural circuits and cortical regions. This diversity supports the wide applicability range of the proposed approach within the motor imagery framework. The proposed approach was compared with state-of-the-art ML models in terms of the achieved classification performance, training time, and capability to address multiclass OVO decomposition.

6.1 Classification performance

The OVO ensemble showed a valid trade-off between recognition performance and model complexity, i.e., the number of trainable parameters. OVO models resulted in the best recognition performance, as reported in Table 2 and Table 6, even when combined with a probability aggregation scheme. Note that a small number of nearest neighbors leads to high recognition performance for both the NEWAXOVO and DRCWOVO schemes. This supports the hypothesis that the base classifiers' competence can be approximated by the local sample population and that widening the competence region leads to performance degradation. There is a direct proportionality between model complexity and its training time; while NEWAXOVO and unweighted OVO showed the shortest training times, comparable performance has previously been achieved using a significantly higher number of trainable parameters [7]. NEWAXOVO requires fewer operations both for training and classification phases than DRCWOVO since it searches for the closest K instances class-wise; DRCWOVO, instead, performs this search for the closest K instances of each class, thus resulting in 4 kNN instances and $4 \times K$ searches. The recognition performance of OVO approaches is affected by non-competent base models; NEWAXOVO employs a weighted aggregation of the probabilities provided by each MLP base classifier, preventing random outcomes.

We introduced *CR* in order to measure the ability to reduce this effect by a weighting scheme, and the results show that the NEWAXOVO scheme outperforms DRCWOVO in terms of recognition robustness. Such a *CR* difference is even more evident when compared to the effect of the unweighted OVO approach. Nonetheless, the difference in *CR* results in a minor improvement in recognition performance. This may be due to a saturation effect linked to the effectiveness of the unweighted OVO scheme in providing high membership scores to correct classes. Especially in the case of Dataset2, the classification performances are so high with the unweighted OVO approach that there is no actual room for significant improvement.

In addition to the performance differences observed between Dataset 1 and Dataset 2, it is important to highlight the explainability aspect of our proposed method, NEWAXOVO. While the state-of-the-art methods, including the ones evaluated in this study, achieve high performance on Dataset 2, our focus goes beyond mere

recognition accuracy. NEWAXOVO incorporates explainable artificial intelligence techniques, allowing for enhanced interpretability of the classification decisions. The ensemble learning architecture of NEWAXOVO combines MLP classifiers, providing not only accurate predictions but also transparent and interpretable explanations for the classification outcomes. This attribute is crucial in domains where model transparency and understandability are of utmost importance, such as in clinical applications or real-time decision-making systems. By integrating explainability into our approach, we not only achieve significant improvements in classification performance on Dataset 1, but also offer insights into the underlying mechanisms driving the classification decisions, promoting trust and understanding of the system's output. The incorporation of explainability techniques further enhances the value and applicability of our proposed method in practical scenarios.

Moreover, it is worth mentioning that the proposed algorithm has been evaluated on two different multi-class motor imagery EEG tasks, thus demonstrating the generalizability of the approach.

6.2 Explainability

NEWAXOVO is an XAI architecture specifically designed for EEG data and BCI applications, providing explainable classification outcomes in the form of counterfactual explanations.

Explanation results at a single-instance level (see Fig. 8) allow to understand why a specific instance has received a given classification. This may help implement a so-called *human-in-the-loop* experimental procedure [30], in which subjects using a BCI system continuously learn and adapt their activity. Alternatively, it can be used for feedback, error detection, and quality check, being embedded in a problem resolution procedure. Indeed, the feature variation needed to counteract misclassifications represents a relevant flag for system corrections, e.g., sampling or other technical issues on an EEG sensor.

Explanation aggregated at subject level (see Fig. 9) allows for subject-specific evaluations, which might be useful in BCI applications involving, e.g., neurological patients [1].

We emphasize that the NEWAXOVO explanation can be further evaluated at a population level, allowing for a piece of more general knowledge about the differences, and therefore, of the decision boundary, between classes.

7 Conclusion

In this study, a novel explainable ensemble-based approach was proposed and successfully tested in publicly available data. The explanation generation approach utilized by NEWAXOVO relies on SHAP, which, however, inherits SHAP's primary limitations. While SHAP is widely recognized as one of the most established XAI approaches, it suffers from computational costs that escalate as the number of features and samples in the dataset increases. Furthermore, existing literature has highlighted SHAP's sensitivity to the presence of correlations and interdependencies among features. Given that both limitations are inherent to the explanation mechanism employed in NEWAXOVO, future developments will explore explanation strategies that are not reliant on SHAP. Furthermore, while the algorithm has been tested on two distinct EEG tasks within the motor imagery domain, future studies will be directed toward a more comprehensive evaluation of its performance across various EEG-based applications. The current study employs an ensemble of neural networks, but the same ensemble schema can be readily used with base classifiers of a different nature (e.g., decision trees). This would broaden the solution search space and provide an additional degree of freedom to further reduce misclassification errors and improve recognition performance. Future works will explore this direction.

Nevertheless, the proposed NEWAXOVO approach outperforms state-of-the-art methods including the models using ensemble learning, multiclass decomposition, and weighted aggregation (e.g., DRCWOVO). NEWAXOVO allows for a minimization of non-competent base classifier contribution in an OVO scheme to provide robust recognition performance while explaining its classification results. We conclude that NEWAXOVO

represents a convenient compromise between high recognition performance and explainability, especially in case of motor imagery BCI tasks based on EEG analysis.

Acknowledgements Work partly funded by the European Commission under: the NextGeneration EU program, PNRR - M4 C2, Investment 1.5"Creating and strengthening of"innovation ecosystems,"building"territorial R&D leaders,"and"THE - Tuscany Health Ecosystem,"Spoke 6"Precision Medicine and Personalized Healthcare"; and Italian Ministry of University and Research (MUR) in the frameworks:"FAIR"PE00000013 Spoke1"Human-centered AI"; National Center for Sustainable Mobility MOST/Spoke10; FoReLab project (Departments of Excellence)."Reasoning"project, PRIN 2020 LS Programme, Project number 2493 04-11-2021. The authors thank Mirco Quintavalla for his work on the subject during his master's thesis.

Author contributions Antonio Luca Alfeo is a correspondence and is responsible for conceptualization, methodology, study design, literature review, writing—original draft preparation. Vincenzo Catrambone contributed to methodology, study design, validation, data curation, data analysis, literature review, and writing—original draft preparation. Mario G. C. A. Cimino and Gaetano Valenza contributed to methodology, study design, validation, funding acquisition, project management, and writing—review and editing.

Data availability All data analyzed in this study are publicly available and included in the published articles [69] and [12].

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Abiri R, Borhani S, Sellers EW et al (2019) A comprehensive review of EEG-based brain-computer interface paradigms. J Neural Eng 16(1):011001
2. Afchar D, Guigue V, Hennequin R (2021) Towards rigorous interpretations: a formalisation of feature attribution. In: International Conference on Machine Learning, PMLR, pp 76–86
3. Aggarwal S, Chugh N (2022) Review of machine learning techniques for EEG based brain computer interface. Arch Comput Methods Eng 29(5):3001–3020
4. Akiba T, Sano S, Yanase T, et al (2019) Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2623–2631
5. Alfeo AL, Cimino MGC, Egidi S, et al (2017) Stigmergy-based modeling to discover urban activity patterns from positioning data. In: Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10, Springer, pp 292–301
6. Alfeo AL, Cimino MG, Lepri B et al (2019) Assessing refugees' integration via spatio-temporal similarities of mobility and calling behaviors. IEEE Trans Comput Soc Syst 6(4):726–738
7. Alfeo AL, Catrambone V, Cimino MG, et al (2021) Recognizing motor imagery tasks from eeg oscillations through a novel ensemble-based neural network architecture. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, pp 5983–5986
8. Alfeo AL, Zippo AG, Catrambone V et al (2023) From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. Comput Methods Programs Biomed 236:107550
9. Alizadeh D, Omranpour H (2023) Em-csp: an efficient multiclass common spatial pattern feature method for speech imagery EEG signals recognition. Biomed Signal Process Control 84:104933
10. Ang KK, Chin ZY, Wang C et al (2012) Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. Front Neurosci 6:39
11. Arpaia P, Esposito A, Natalizio A et al (2022) How to successfully classify EEG in motor imagery BCI: a metrological analysis of the state of the art. J Neural Eng 19(3):031002
12. Averta G, Barontini F, Catrambone V, et al (2021) U-limb: A multi-modal, multi-center database on arm motion control in healthy and post-stroke conditions. GigaScience 10(6):giab043
13. Azevedo T, Passamonti L, Liò P, et al (2020) A deep spatiotemporal graph learning architecture for brain connectivity analysis. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, pp 1120–1123

14. Banerjee M, Capozzoli M, McSweeney L et al (1999) Beyond kappa: a review of interrater agreement measures. Can J Stat 27(1):3–23
15. Boostani R, Moradi MH (2004) A new approach in the bci research based on fractal dimension as feature and adaboost as classifier. J Neural Eng 1(4):212
16. Catrambone V, Greco A, Averta G et al (2019) Predicting object-mediated gestures from brain activity: an EEG study on gender differences. IEEE Trans Neural Syst Rehabil Eng 27(3):411–418
17. Catrambone V, Averta G, Bianchi M et al (2021) Toward brain-heart computer interfaces: a study on the classification of upper limb movements using multisystem directional estimates. J Neural Eng 18(4):046002
18. Challis E, Hurley P, Serra L et al (2015) Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fmri. Neuroimage 112:232–243
19. Cimino MG, Pedrycz W, Lazzerini B et al (2009) Using multilayer perceptrons as receptive fields in the design of neural networks. Neurocomputing 72(10–12):2536–2548
20. Cruz RM, Cavalcanti GD, Ren TI (2010) An ensemble classifier for offline cursive character recognition using multiple feature extraction techniques. In: The 2010 International Joint Conference on Neural Networks (IJCNN), Ieee, pp 1–8
21. Cruz RM, Sabourin R, Cavalcanti GD (2018) Dynamic classifier selection: Recent advances and perspectives. Information Fusion 41:195–216
22. Dang W, Lv D, Tang M, et al (2024) Flashlight-net: a modular convolutional neural network for motor imagery EEG classification. IEEE Transactions on Systems, Man, and Cybernetics: Systems
23. Das R, Lopez PS, Khan MA, et al (2020) 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), IEEE, pp 1275–1279
24. Degirmenci M, Yuce YK, Perc M et al (2024) EEG channel and feature investigation in binary and multiple motor imagery task predictions. Front Hum Neurosci 18:1525139
25. Delaney E, Greene D, Keane MT (2021) Instance-based counterfactual explanations for time series classification. In: International Conference on Case-Based Reasoning, Springer, pp 32–47
26. Dip MSS, Hasan MA, Kabir S, et al (2023) Deep learning based motor imagery intention classification using electroencephalogram signal. In: 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), IEEE, pp 143–147
27. Dong S, Jin Y, Bak S et al (2021) Explainable convolutional neural network to investigate age-related changes in multi-order functional connectivity. Electronics 10(23):3020
28. Duin RP (2002) The combining classifier: to train or not to train? In: Object recognition supported by user interaction for service robots, IEEE, pp 765–770
29. Echtioui A, Zouch W, Ghorbel M et al (2024) Classification of bci multiclass motor imagery task based on artificial neural network. Clin EEG Neurosci 55(4):455–464
30. Fellous JM, Sapiro G, Rossi A et al (2019) Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Front Neurosci 13:1346
31. Ferracuti F, Iarlori S, Mansour Z et al (2021) Comparing between different sets of preprocessing, classifiers, and channels selection techniques to optimise motor imagery pattern classification system from EEG pattern recognition. Brain Sci 12(1):57
32. Gabard-Durnam LJ, Mendez Leal AS, Wilkinson CL et al (2018) The harvard automated processing pipeline for electroencephalography (happe): standardized processing software for developmental and high-artifact data. Front Neurosci 12:97
33. Gagliardi G, Alfeo AL, Catrambone V, et al (2023a) Improving emotion recognition systems by exploiting the spatial information of EEG sensors. IEEE Access
34. Gagliardi G, Alfeo AL, Catrambone V, et al (2023b) Using contrastive learning to inject domain-knowledge into neural networks for recognizing emotions. In: 2023 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, pp 1587–1592
35. Gagliardi G, Alfeo AL, Catrambone V, et al (2023c) Fine-grained emotion recognition using brain-heart interplay measurements and explainable convolutional neural networks. In: 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, pp 1–6
36. Galar M, Fernández A, Barrenechea E et al (2011) An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn 44(8):1761–1776
37. Galar M, Fernández A, Barrenechea E et al (2015) Drcw-ovo: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. Pattern Recogn 48(1):28–42
38. Galatolo FA, Cimino MG, Marincioni A, et al (2021) Noise boosted neural receptive fields. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, pp 1–6
39. Goienetxea I, Mendialdua I, Rodríguez I et al (2021) Problems selection under dynamic selection of the best base classifier in one versus one: Pseudovo. Int J Mach Learn Cybern 12(6):1721–1735
40. Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation''. AI Mag 38(3):50–57

41. Han J, Pei J, Tong H (2022) Data mining: concepts and techniques. Morgan kaufmann
42. He L, Hu D, Wan M et al (2015) Common bayesian network for classification of EEG-based multiclass motor imagery BCI. IEEE Trans Syst, Man, Cybernet: Syst 46(6):843–854
43. Herm LV, Heinrich K, Wanner J et al (2023) Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. Int J Inf Manage 69:102538
44. Hosseini MP, Hosseini A, Ahi K (2020) A review on machine learning for EEG signal processing in bioengineering. IEEE Rev Biomed Eng 14:204–218
45. Hussain I, Jany R, Boyer R et al (2023) An explainable EEG-based human activity recognition model using machine-learning approach and lime. Sensors 23(17):7452
46. Li P, Liu H (2023) Binary decomposition for multi-class classification problems: Development and applications. In: 2023 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, pp 452–457
47. Lian S, Li Z (2024) An end-to-end multi-task motor imagery EEG classification neural network based on dynamic fusion of spectral-temporal features. Computers in Biology and Medicine p 108727
48. Lotze M, Montoya P, Erb M et al (1999) Activation of cortical and cerebellar motor areas during executed and imagined hand movements: an fmri study. J Cogn Neurosci 11(5):491–501
49. Lu P, Gao N, Lu Z et al (2019) Combined CNN and LSTM for motor imagery classification. 2019 12th International Congress on Image and Signal Processing. IEEE, BioMedical Engineering and Informatics (CISP-BMEI), pp 1–6
50. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems 30
51. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
52. Mudgal SK, Sharma SK, Chaturvedi J et al (2020) Brain computer interface advancement in neurosciences: applications and issues. Interdiscip Neurosurg 20:100694
53. Musallam YK, AlFassam NI, Muhammad G et al (2021) Electroencephalography-based motor imagery classification using temporal convolutional network fusion. Biomed Signal Process Control 69:102826
54. Nicolas-Alonso LF, Corralejo R, Gomez-Pilar J et al (2015) Adaptive semi-supervised classification to reduce inters-ession non-stationarity in multiclass motor imagery-based brain-computer interfaces. Neurocomputing 159:186–196
55. Nisar H, Cheong JY, Yap VV (2024) EEG-based biometrics for user identification using deep learning method. In: 2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA), IEEE, pp 1–6
56. Olcay BO, Karaçalı B (2023) Time-resolved EEG signal analysis for motor imagery activity recognition. Biomed Signal Process Control 86:105179
57. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
58. Pérez-Velasco S, Marcos-Martínez D, Santamaría-Vázquez E et al (2024) Unraveling motor imagery brain patterns using explainable artificial intelligence based on shapley values. Comput Methods Programs Biomed 246:108048
59. Prabhavathy T, Elumalai VK, Balaji E (2024) Hand gesture classification framework leveraging the entropy features from SEMG signals and VMD augmented multi-class SVM. Expert Syst Appl 238:121972
60. Quintavalla AALCMGCAM., Vaglini G (2024) NEWAXOVO. https://github.com/AntonioLucaAlfeo/NEWAXOVO, online; accessed 18-April-2024
61. Ranjan A, Singh VP, Singh AK et al (2020) Classifying brain state in sentence polarity exposure: An ann model for fmri data. Revue d'Intelligence Artificielle 34(3):361–368
62. Richman JS, Moorman JR (2000) Classifying brain state in sentence polarity exposure: an ANN model for FMRI data. Am J Physiol-Heart Circul Physiol 278(6):H2039–H2049
63. Sagi O, Rokach L (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249
64. Salmeron JL, Correia MB, Palos-Sanchez PR (2019) Complexity in forecasting and predictive models
65. Selim S, Tantawi MM, Shedeed HA et al (2018) A csp\am-ba-svm approach for motor imagery bci system. Ieee Access 6:49192–49208
66. She Q, Zou J, Meng M et al (2021) Balanced graph-based regularized semi-supervised extreme learning machine for EEG classification. Int J Mach Learn Cybern 12:903–916
67. Stepin I, Alonso JM, Catala A et al (2021) A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 9:11974–12001
68. Tallapragada SKC, Turlapaty AC, Gokaraju B, et al (2021) Optimal features for cross subject classification of imagined left and right fist movements using EEG signals. In: 2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, pp 1–5
69. Tangermann M, Müller KR, Aertsen A, et al (2012) Review of the bci competition iv. Frontiers in neuroscience p 55
70. Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Trans Neural Netw Learn Syst 32(11):4793–4813
71. Vimbi V, Shaffi N, Mahmud M (2024) Interpreting artificial intelligence models: a systematic review on the application of lime and Shap in Alzheimer's disease detection. Brain Inf 11(1):10

72. van der Waa J, Nieuwburg E, Cremers A et al (2021) Evaluating xai: A comparison of rule-based and example-based explanations. Artif Intell 291:103404
73. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL & Tech 31:841
74. Wang L, Xu G, Yang S, et al (2012) Motor imagery bci research based on sample entropy and svm. In: 2012 Sixth International Conference on Electromagnetic Field Problems and Applications, IEEE, pp 1–4
75. Wang L, Li M, Xu D, et al (2024) Cortical roi importance improves mi decoding from EEG using fused light neural network. IEEE Transactions on Neural Systems and Rehabilitation Engineering
76. Xu C, Sun C, Jiang G et al (2020) Two-level multi-domain feature extraction on sparse representation for motor imagery classification. Biomed Signal Process Control 62:102160
77. Xu Y, Liu X, Pan L et al (2021) Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. IEEE J Biomed Health Inform 26(3):1362–1373
78. Yu S, Wang Z, Wang F, et al (2024) Multiclass classification of motor imagery tasks based on multi-branch convolutional neural network and temporal convolutional network model. Cerebral Cortex 34(2):bhad511
79. Zhang R, Zong Q, Dou L et al (2021) Hybrid deep neural network using transfer learning for EEG motor imagery decoding. Biomed Signal Process Control 63:102144
80. Zhang ZL, Luo XG, Garcia S et al (2017) Exploring the effectiveness of dynamic ensemble selection in the one-versus-one scheme. Knowl-Based Syst 125:53–63
81. Zhang ZL, Chen YY, Li J et al (2019) A distance-based weighting framework for boosting the performance of dynamic ensemble selection. Inf Process Manag 56(4):1300–1316
82. Zhang ZL, Zhang CY, Luo XG et al (2023) A multiple classifiers system with roulette-based feature subspace selection for one-vs-one scheme. Pattern Anal Appl 26(1):73–90

Authors and Affiliations

Antonio L. Alfeo[1,2] · Vincenzo Catrambone[1,2] · Mario G. C. A. Cimino[1,2] · Gaetano Valenza[1,2]

✉ Antonio L. Alfeo
luca.alfeo@unipi.it

Vincenzo Catrambone
vincenzo.catrambone@unipi.it

Mario G. C. A. Cimino
mario.cimino@unipi.it

Gaetano Valenza
gaetano.valenza@unipi.it

[1] Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56126 Pisa, Italy

[2] Bioengineering and Robotics Research Center E. Piaggio, University of Pisa, Largo Lucio Lazzarino 1, 56126 Pisa, Italy