Automatic Feature Selection for Sensorimotor Rhythms Brain-Computer Interface Fusing Expert and Data-Driven Knowledge

Mushfika Sultana, *Member, IEEE*, and Serafeim Perdikis, *Member, IEEE*

***Abstract*— Early brain-computer interface (BCI) systems were mainly based on prior neurophysiological knowledge coupled with feedback training, while state-of-the-art interfaces rely on data-driven, machine learning (ML)-oriented methods. Despite the advances in BCI that ML can be credited with, the performance of BCI solutions is still not up to the mark, posing a major barrier to the widespread use of this technology. This paper proposes a novel, automatic feature selection method for BCI able to leverage both data-dependent and expert knowledge to suppress noisy features and highlight the most relevant ones thanks to a fuzzy logic (FL) system. Our approach exploits the capability of FL to increase the reliability of decision-making by fusing heterogeneous information channels while maintaining transparency and simplicity. We show that our method leads to significant improvement in classification accuracy, feature stability and class bias when applied to large motor imagery or attempt datasets including end-users with motor disabilities. We postulate that combining data-driven methods with knowledge derived from neuroscience literature through FL can enhance the performance, explainability, and learnability of BCIs.**

***Index Terms*— Brain-computer interface, expert knowledge, feature selection, fuzzy logic, motor imagery.**

I. Introduction

BCI is gaining ground as a promising control channel of assistive or rehabilitation solutions for people with devastating motor disabilities [1]. Non-invasive, electroencephalography (EEG)-based BCIs relying on the self-modulation of sensorimotor rhythms (SMR) have shown great applicability potential enabling assistive technology [2], [3] and rehabilitation [4] prototypes in a safe, minimally obtrusive manner.

Manuscript received 15 February 2024; revised 25 May 2024 and 28 August 2024; accepted 4 September 2024. Date of publication 9 September 2024; date of current version 18 September 2024. *(Corresponding author: Serafeim Perdikis.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committe of Canton de Vaud (CER-VD), Switzerland, and performed in line with the Declaration of Helsinki.

The authors are with the Brain-Computer Interface and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: ms17811@essex.ac.uk; serafeim.perdikis@essex.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2024.3456591, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2024.3456591

BCI systems have evolved from neurofeedback paradigms; these early BCI designs relied on expert neurophysiological knowledge to identify one or a few salient and learnable brain features a subject may be able to get control of, and on neurofeedback training [5]. Subsequently, the introduction of elaborate ML allowed to considerably minimize the user training requirements and boost the information transfer rates by facilitating the decoding of complex, highly multidimensional and distributed brain activity [6]. Despite great benefits brought forward by the "ML way to BCI", this approach has failed to deliver on its promise of zero-training interfaces, universal accessibility to BCI for all prospective users and stable performance. The BCI community gradually acknowledges these limitations leading to a recent revival of interest in issues pertaining to the "human in the loop" [3], [7].

Current ML-oriented BCI implementations tend to rely exclusively on purely data-driven methods; consequently, precious insights derived by research in neuroscience and neuropsychology are often ignored, potentially to the final system's detriment. Along these lines, the investment on ever more elaborate ML, including deep learning (DL) methods [8], renders the BCI a "black-box" that lacks interpretability and obscures the neurophysiological basis of a BCI [9]. This is particularly problematic as the BCI field mostly targets translational applications, where explainability can be as critical as performance. There is thus a need to combine and enrich data-driven methods with expert knowledge in order to enjoy the best of the two worlds in the form of both powerful and explainable, white-box models.

Towards this direction, we propose a novel feature selection method for BCI able to exploit both data-driven and expert information thanks to a fuzzy system [10]. Feature selection is essential for any pattern recognition (PR) system; it helps remove redundant and irrelevant features that can degrade decoding performance, and reduce the dimensionality of the feature space, which can attenuate or eliminate overfitting effects [11]. Despite the vast majority of published BCI work concerns ML methodology [12], research on feature selection tailored to BCI is scarce [13], [14], [15]. Furthermore, these works adopt popular data-driven feature selection techniques derived in general PR literature ignoring the particularities of brain signals. While the rise of DL lately shifts the

focus towards embedded feature extraction/selection, filter-based methods [11], where candidate features are ranked according to some fitness criterion reflecting their projected value for classification and the N-best are selected, are still the most popular in BCI [3], [15], [16], [17], [18], [19]. Wrapper methods have also been investigated [14], [20], but are not preferred in real-world settings because of their high computational complexity.

On the other hand, BCI systems that do take into account the outputs of basic neuroscience research in the design of the PR modules tend to ignore or only naively fuse data evidence. One example is the category of adaptive interfaces that either fix [21], [22] or narrow down [23] the brain features fed to the classifier to the EEG channels, time points and/or frequency bands that previous neurophysiological studies show will be modulated by an "average" user. In other works, data-driven and prior knowledge are combined through manual intervention [2], [3], which requires the presence of a BCI expert, a pre-requisite that cannot be acceptable in clinical and home applications of BCI. There is thus a gap in the BCI literature with respect to feature selection methods capable of integrating neurophysiological priors with data-based knowledge in a fully-fledged, transparent and fully automatic manner.

Effective embedding of expert information into ML models has been mainly demonstrated with Bayesian techniques [24], FL [10] and Dempster-Shafer theory [25]. Here, we choose to work within the FL framework for its great intuitiveness and flexibility. Our proposed feature selection module is formalized as a Mamdani fuzzy inference system that produces a modified feature fitness (FF) ranking of all candidate features, taking into account both their separability and anticipated suitability according to the relevant neuroscience literature. FL-inspired feature selection has been previously attempted [26], [27], [28], [29]; however, the methods presented in these articles are fully data-driven, with no effort to exploit expert information. FL has been used in BCI also for feature extraction and classification with EEG signals [30], [31], [32].

The main finding of this work is that fusing neurophysiological priors with data separability information for feature selection through fuzzy inference results in up to 7% classification accuracy increase on average in SMR BCI. We additionally show that these gains are larger the more noisy the underlying data are. Importantly, we reveal that benefits mainly arise because purely data-driven methods are unable to distinguish between physiological and noise-induced feature separability; embedding expert knowledge addresses to a great extent this shortcoming by attenuating the influence of discriminant power (DP) emerging in physiologically unlikely locations and/or frequency bands, thus delivering a more consistent, informed and parsimonious feature selection outcome. Our study is further distinguished for assessing the proposed method on large SMR BCI datasets comprising one or more sessions of 92 users, including 46 end-users with disability, thus solidly grounding the reliability of our findings.

II. Materials and Methods

A. Dataset

The data reported here consist in SMR BCI EEG recordings from a total of 92 participants and three distinct datasets. The first dataset regards 46 able-bodied volunteers (41±9 years, 10 female), the second includes 21 end-users with motor disabilities undergoing motor imagery (MI) BCI training to operate Assistive Technology (AT) BCI solutions [2] (56±26 years, 3 female), and the third dataset comprises 25 acute/sub-acute stroke patients with severe hemiplegia (62±11 years, 9 female) undergoing a BCI-based motor rehabilitation therapy [33]. The AT patients are affected by various pathologies including myopathy, spinal cord injury, amputation, spinocerebellar ataxia or multiple sclerosis. All participants are without cognitive deficits. All reported experiments were approved by the local ethical committees [2], [33]. Informed consent was received from all human subjects.

B. EEG and System Configuration

For all datasets, the brain activity was acquired via 16 active EEG channels over the sensorimotor cortex: Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, and CP4 according to the international 10-20 system with reference on the right ear and ground on AFz (see supplementary Fig. S1(a)). The EEG was recorded with a g.USBamp system (g.Tec, Austria) at 512 Hz. Raw signals were bandpassed between 0.1 Hz and 100 Hz, and notch-filtered at 50 Hz.

C. Experimental Protocols and Feature Extraction

The able-bodied and AT patient datasets are composed of 1-10 MI BCI sessions per subject, each comprising 3-4 calibration (open-loop) and/or "online" (closed-loop) runs. Each run contains 15 trials for each of the MI tasks included. The MI tasks present in the database are right hand, left hand, both hands, both feet MI and "resting" (idling). The experimental protocol is detailed in [2]. In the rehabilitation dataset, each participant executed 15 BCI therapy sessions of 3-7 closed-loop runs each, where the BCI detects the patients' motor attempts with the affected hand (15 trials) against "rest/no-movement" (15 trials). The corresponding protocol is elaborated in [33]. Pre-processing and Power Spectral Density (PSD) feature extraction using the Welch method [34] followed the methodology reported in [2]. The extracted PSD features reflect the power of 16 channels × 23 frequency bands yielding a total of 368 candidate features fed to the feature selection module.

D. Adding Noise to Extracted PSD

To highlight the added value of the proposed method in the presence of EEG artifacts, various levels of noise are artificially superimposed on the PSD features. We have used an independent dataset containing artifact-contaminated EEG described in [35]. The PSD of each artifact trial for each EEG channel was extracted with the same parameters as for the SMR trials. Each SMR trial may be contaminated

by artifacts based on a random decision with probability p_{nl} (i.e., p_{nl} determines the proportion of trials inflicted with noise, against $1 - p_{nl}$ that are left intact). Hence, the parameter p_{nl} defines the "noise level" added and is varied across repetitions of our analysis in [0, 1] (extremities express no, and full contamination, respectively) with a step of 0.1. Contamination is carried out through a weighted average $\mathbf{x'^{f}_{t}} = 0.5\mathbf{x^{f}_{t}} + 0.5\hat{\mathbf{x}}^{f}_{T}$, where $\mathbf{x'^{f}_{t}}$ the artificially contaminated PSD value for feature f at time t, $\mathbf{x^{f}_{t}}$ the corresponding feature value extracted naturally from each subject's MI data and $\hat{\mathbf{x}}^{f}_{T}$ the PSD value of the randomly selected artifact trial T. Only the "frowning" and "teeth clenching" types of artifacts are used from the data in [35], selected with a probability of 0.5 (uniform distribution). Each channel is contaminated with noise from the same (or, if missing, a commonly available) channel of the artifact trial. To better simulate the vulnerability of peripheral channels to artifacts under real-world conditions, channel Fz of our narrow channel layout is contaminated with the most affected peripheral channel of the broader artifact dataset layout, instead; this is a channel among those in the following channel set: Fz, F2, AF4, FP2, FP1, P2, PO4, P6, O2, P5, O1, PO3, P1, AF3, F1, Oz, POz, and Pz. This special treatment of Fz is meant to compensate for the fact that, in the low-density EEG channel configuration used in our SMR studies, Fz may be the most peripheral channel, but is still not "peripheral enough" to reflect well the amount of noise interference that can occur in the periphery of denser and more widely spread EEG layouts during real BCI application. Please, refer to the supplementary material for further details.

E. Feature Selection Through FL

To automatically select the N-best features compromising separability with neurophysiological relevance, we propose Fuzzy Logic-based Automatic Feature Selection (FL-AFS), a method that combines neuroscientific knowledge concerning the cortical areas and frequency bands that are expected to be activated by the employed MI or motor attempt tasks [36], [37] with the features' data-driven DP ranking.

F. Fuzzy System Design

1) Fuzzy sets and membership functions: FL-AFS operates independently for each feature to output a modified FF value as opposed to the FF encoded by a conventional, data-driven separability/DP metric, hereafter termed Data-driven Automatic Feature Selection (DD-AFS). The r^2 measure is selected to implement DD-AFS [38] (other feature ranking algorithms are also tested). This FF value is used to rank the candidate features and allow the selection of the N best ones. The Mamdani-type fuzzy inference system receives three inputs for each candidate feature: Location (channel), Frequency Band and DP indexed again by r^2 feature separability, i.e., the coefficient of determination between the feature values and the MI class label. The coefficient of determination is defined as $r^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$ where SS_{res} is the sum of squares of residuals of the linear fit and SS_{tot} is the total sum of squares. We argue that the choice of separability metric is of no particular importance (i.e., Fisher Score [21], Canonical Variate Analysis [2], mutual information [17] or other feature ranking and filter-based feature selection approaches can be used without detriment to implement DD-AFS and the DP input of FL-AFS) because it is normalized by dividing the raw DP of each feature with the total DP of the candidate feature set, thus merely quantifying the DP input with any one among many different alternatives and without affecting its numerical bounds. The supplementary material corroborates this by replicating the results after replacing r^2 with Fisher Score, Gradient Boosting and Extreme Gradient Boosting (XGBoost) for DD-AFS and the DP input of FL-AFS.

A single fuzzy set is defined for each of the three inputs whose membership function (MB) specifies what truth values can be considered "good" for the input in question. Hence, the respective fuzzy sets convey the information about which channels and which bands can be considered good (and how much) for a particular taskset; and which r^2 values are good, in general. In this binary (good/bad) modelling for the fuzzy system's input variables the explicit definition of a fuzzy set for "bad" is not necessary, as the fuzzy operator NOT is later used in the fuzzy rules, thus implicitly defining the "bad" fuzzy set with the complementary MB $f_{bad} = 1.0 - f_{good}$.

Custom composite MBs are created for all three inputs to accurately reflect the relevant neurophysiological knowledge from the literature [36], [37], [39] and from our own SMR BCI expertise. In summary, it is known that subjects performing right or left hand MI exhibit strong contralateral SMR in the μ (8-14 Hz) band and/or the β band (18-24 Hz). If feet imagery is part of the mental exercises, the topographic SMR distribution maps typically reveal central cortical activation in the β band. During hand imagery tasks, ipsilateral activation may also be observed, although most often less pronounced than the contralateral one. The contralateral activations tend to be more pronounced and stronger in the μ band [37].

Based on such "rules of thumb" that can be derived from the literature of MI BCI, regarding the Location input variable, a different MB is created for each MI taskset (i.e., pair of MI tasks used for closed-loop BCI control) included in our analysis: right hand and left hand (RHLH), right hand and feet (RHBF), left hand and feet (LHBF), right hand and rest (RHRST), or left hand and rest (LHRST). These MBs are constructed as follows: the highest membership degree is assigned to contralateral (with respect to the right/left hand) channels (i.e., the maximum weight of 1.0 for C1 and C2, Fig. 1(a)) when the respective MI is part of the taskset employed. Slightly lower membership values are assigned to ipsilateral channels and electrodes that are more peripheral or vertically displaced with respect to the central zone (i.e., centro-parietal and fronto-central channels). Medial channels are assigned no membership whatsoever in the "good" fuzzy set for the RHLH taskset (Fig. 1(a)). For tasksets containing feet MI, a high membership value is assigned to the medial channels (1.0 for Cz, 0.80 for CPz and FCz) and less weight is given to the lateral ones (0.40 for FC1, CP1 concerning LHBF; 0.40 for FC2, CP2 with respect to RHBF). When a certain channel falls in two of these categories (e.g., for the RHLH taskset, channel C1 is both contralateral to the right hand

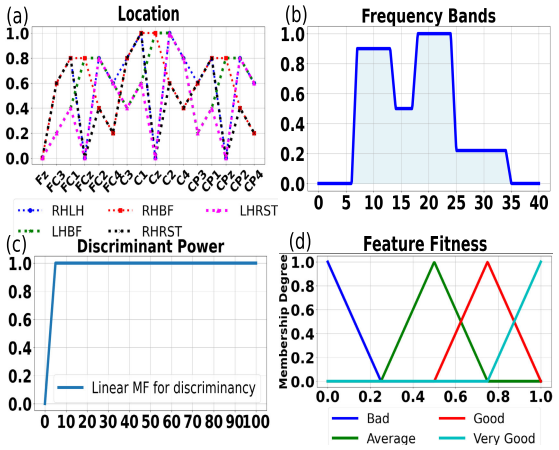

Fig. 1. Membership functions of all input/output fuzzy sets defined.

and ipsilateral to the left hand) the maximum of the possible membership values is taken. Fig. 1(a) illustrates the MBs designed in this manner for each of the aforementioned five tasksets examined in our analysis. In the interest of clarity, the individual membership functions for the Location input (one for each of the five different tasksets considered) are provided in the supplementary Fig. S3.

The MB for the "good" fuzzy set of the frequency band input (Fig. 1(b)) is built by superimposing suitable trapezoidal function templates in such a way so that high membership is given to the μ and β bands, in accordance with the literature. The MB for separability, which serves to fuse also the data-derived information into feature selection, is based on a linear function where the membership value starts at 0.0 and increases to 1.0 linearly. Since, as underlined above, the DP is normalized, the truth value of separability for each feature in fact corresponds to the percentage (%) of the total separability accounted for by the feature in question. Based on experience, and aided by the visual inspection of the DP % values that high- and low-performing subjects in the dataset of [21] tend to achieve, we concluded that 5% of the total DP or above signifies a very separable feature. Accordingly, the "good" separability MB of our fuzzy system will reach the maximum membership of 1.0 at 5% of DP (Fig. 1(c)).

Finally, 4 fuzzy sets are defined for the output FF variable, associated with "bad", "average", "good" and "very good" FF. The respective MBs, based solely on triangular function templates, are depicted in Fig. 1(d).

2) Fuzzy rules and fuzzification: We defined 8 IF-THEN-statement rules shown in Table I. Each rule regards one of the $8 = 2^3$ possible combinations of our 3 binary (bad/good) inputs. For example, the first rule is described by the statement:

> R1: IF Location IS good AND Frequency Band IS good AND Discriminant Power IS good THEN Feature Fitness IS very good

Fuzzification in our system proceeds as follows for each candidate feature: First, for each rule, the membership values corresponding to the three truth values (channel, band and DP of the said feature) are extracted. These are then combined with the AND operator to form the rule's antecedent value. Operator AND is given its most common interpretation as the "min" membership value among the 3 inputs. Subsequently, in the rule's implication process, the fuzzy set of the consequent for this rule is computed with the "min" operator; effectively, this truncates the MB of the output variable's fuzzy set that appears in this rule's consequent, as typically done in Mamdani-type fuzzy systems. Fuzzification is concluded by applying this procedure with all 8 rules present in our system and aggregating the outputs of all rules through the "max" operator. All rules exercise the same effect on the output (i.e., all rule weights are set to 1.0). The final outcome of the fuzzification process is the MB of a final output FF fuzzy set.

TABLE I
SUMMARY OF FUZZY RULES OF FL-AFS

Rule	Location	Frequency Band	DP	Feature Fitness
R1	good	good	good	very good
R2	not good	good	good	bad
R3	good	not good	good	bad
R4	good	good	not good	bad
R5	not good	not good	good	bad
R6	not good	good	not good	bad
R7	good	not good	not good	bad
R8	not good	not good	not good	bad

3) Defuzzification and final output: The defuzzification process extracts the system's final, "crisp" FF value. Among the different defuzzification methods proposed in the literature [40], [41], we have selected the most popular one, namely, the "centroid" method (i.e., the output MB's centre of mass).

G. Fuzzy System Design Rationale and Effects

The FL inference system design described above targets two main interventions on a regular, solely data-driven FF output (DD-AFS). Primarily, we intend to suppress "false positive" DP that often arises as a result of the presence of EEG artifacts. Some typical, but by no means unique, cases are shown in Fig. 2 and 3, and supplementary Fig. S2. Ocular and muscular artifacts may occur arbitrarily during the execution of one or more trials of some MI task. Given that the amplitude of these signals is at least one order of magnitude greater than pure EEG activity, the EEG signature these signals generate by propagating on the scalp is mistaken by the feature ranking/separability monitoring method as a strong EEG correlate of the underlying MI task. Much greater DP than what the actual MI correlates yield is therefore inferred for at least some of the candidate features. There is no easy and efficient data-driven way to discriminate between "fake" and "authentic" DP. There do exist artifact detection and removal methods [35]; however, no method guarantees absolute success in removing artifacts and all methods are still computationally expensive. Therefore, the best, and probably simplest, way to identify "fake good" features is to challenge the origin of their high fitness on the grounds of prior neurophysiological knowledge. Indeed, artifact-induced separability will very often be found on EEG features that cannot be justified by neuroscience studies. In the context of MI BCI, where spatio-spectral features are commonly used due to their ability to capture all aspects of the Event-Related Desynchronization (ERD)/Event-Related Synchronization (ERS) phenomenon,

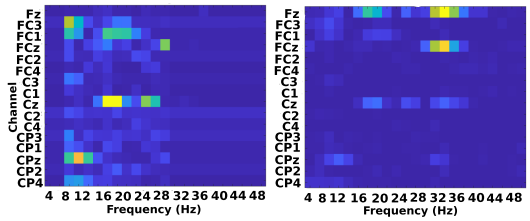

Fig. 2. FF heatmaps (FL-AFS left, DD-AFS right). Blue indicates low and yellow good FF. FL-AFS suppresses noise on Fz.

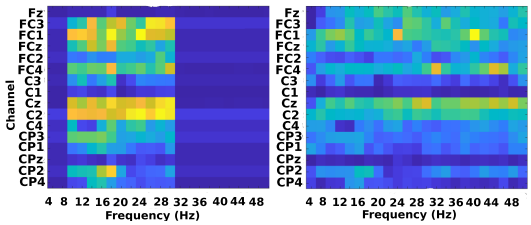

Fig. 3. FL-AFS suppresses abnormal DP in high frequency bands.

this practically means that fake separability will most likely be found in channels and frequency bands not justified by MI studies.

Fig. 2 shows an example of fake DP on channel Fz, the most frontal electrode of the EEG layout employed here, which makes it vulnerable to the eye and facial muscle artifacts (eye movements, blinking, frowning, eyebrow-raising, etc.). Hence, high DP on the location of channel Fz is most certainly due to noise, rather than related to the subject's MI. Furthermore, in case of strong muscular artifacts such as those produced by jaw movements, researchers often observe high separability in high-β and γ frequency bands (Fig. 3), usually evident in many neighbouring channels. This separability cannot be possibly attributed to MI, yet, an exclusively data-driven automatic feature selection method will be bound to select these features as optimal, thus resulting in a classifier model unable to detect actual MI patterns. FL-AFS is also designed to resolve cases of subjects with low SMR aptitude (see supplementary material and Fig. S2).

It must be highlighted that, as is common in FL system design, the final hyper-parameterization and architecture have been to some degree the result of "trial-and-error" tweaking, showcasing the flexibility of the FL framework to easily compromise knowledge by different experts, or update the model as novel expert knowledge is built up by experience. Specifically, our initial intuitions were revised through visual inspection of the effects that different choices have on the FL-AFS-modified DP maps on the data of [3] and [21]. The suitability of the final choices was confirmed, again by visual inspection of the modified DP maps, with a few sessions of the best- and worst-performing able-bodied and AT patients in our reported dataset, as a proofing mechanism before result extraction. No further fine-tuning and no formal, data-driven hyperparameter tuning took place.

For example, readers may note that two of the defined fuzzy sets for the FF output variable ("average", "good") are not used at all. We originally intended for each fuzzy rule to map antecedents to a consequent as:

- (all) three "good" inputs lead to "very good" FF
- (any) two "good" inputs lead to "good" FF
- (any) one "good" input leads to "average" FF
- (all) three "bad" inputs lead to "bad" FF

However, it was soon realized that this scheme would not lead to sufficient "fake" DP suppression. Conversely, requiring all three inputs to be "good" so as to infer "very good" FF, and mapping all other cases to "bad" FF delivered the desired outcome and was the solution we converged to (Table I). In our final architecture, the activation of the first rule is "driving" the final crisp output towards high FF, while the remaining 7 rules act as "penalizers" of different combinations of deficiencies (i.e., not-so-relevant location/band, low separability, etc.), dragging the crisp output towards lower fitness.

Similarly, the exact shape of the MBs of the inputs was determined by studying the effects of different choices on actual MI data [3], [21]. In certain cases, this led to final results that were counter-intuitive in the beginning, like the MB for the "good" fuzzy set for the Frequency Band input, where β features have been eventually assigned higher weight than μ ones. This has been found to be necessary mainly to serve our second goal (of meaningful/learnable selected features) for end-user participants, for whom β features are prevalent.

H. Classification

The classification of PSD feature vectors after selection is effectuated with a Linear Discriminant Analysis (LDA) model (main paper results). Automatic shrinkage with the OAS method [42] is applied to all class-wise covariance matrices in case they are not positive semi-definite. The LDA's common covariance matrix is estimated as the average of the class-wise covariance matrices. We extract and report two types of accuracy: per run or per session. In both cases, for closed-loop (online) runs/sessions the LDA classifier used to classify the data has been trained on the data of the previous run/session, respectively. Classification accuracy for the first (calibration) session is extracted through 5-fold cross-validation on the session's own data (without shuffling, and keeping PSD samples of the same trial in the same fold to avoid confounds due to training-testing data dependence), as no previous data exists in this case. Feature selection is naturally also done on the data of the previous run/session prior to training the LDA model. In the supplementary material, we show that the proposed feature selection method delivers improvements independent of the classifier subsequently used, by additionally employing Support vector machine (SVM) and Random Forest classifiers.

I. Evaluation

We evaluate the performance of the FL-AFS algorithm across three different aspects: SMR BCI performance, selected feature stability and avoidance of class-bias. In all cases, the respective evaluation metric is calculated for FL-AFS and compared to DD-AFS. We standardize the finally reported result as the difference $M_{FL-AFS} - M_{DD-AFS}$ for each metric M, so that positive values for accuracy and feature stability, and negative values for class bias indicate the superiority of the proposed FL-based method over the conventional data-driven method. Furthermore, we present this difference across

all combinations of a number of selected features $N_f = 2, 5, 10, 15, 20, 30, 50$ and noise level $p_{nl} \in [0 : 10 : 100]\%$. For statistical analysis, all sessions/runs executed by each subject are pulled together within each subject's dataset and averaged. Subsequently, averages and statistical testing are reported across subjects. Paired, two-sample t-tests ($\alpha = 0.05$) are performed for each evaluated measure with Bonferroni correction to account for multiple comparisons due to canvassing several values of selected feature set cardinality and noise level.

The immediate effect on SMR BCI performance is assessed through classification accuracy, computed as detailed above. Accuracy reflects the percentage of correctly classified samples (of both classes) over the total number of samples in the run/session.

Feature stability in the context of SMR BCI systems refers to the consistency of the selected features across different runs or sessions. Stable features are crucial for the robustness and reliability of a SMR BCI. Feature stability is quantified thanks to the Jaccard similarity coefficient, a statistical measure assessing the similarity of two sets A, B as the size of their intersection over the size of their union: $J(A, B) = |A \cap B|/|A \cup B|$. Here, $J(S_i, S_{i+1})$ measures the similarity of the features selected in two consecutive runs/sessions, S_i and S_{i+1}, and effectively communicates the percentage of selected features that were common. Hence, a Jaccard Index value close to $J = 1.0$ indicates high stability, meaning most of the features are consistent across consecutive runs/sessions, while $J = 0.0$ signifies that the selected features tend to be entirely different every time a new feature selection round is performed.

We also report a measure of class-bias as the distribution of accuracy across the two MI classes. Our assumption is that, by promoting the selection of physiologically relevant features rather than only the most discriminant ones, the final feature set will be more balanced with respect to features that are responsive for both MI tasks. On the contrary, our observation is that, often, a subject's N most separable features will tend to exhibit ERD/ERS only, or at least primarily, for one of the two MI tasks in the taskset. This leads to "biased" control during 2-class closed-loop SMR BCI where a subject can deliver successfully one of the two mental commands, but struggles with the other one. The class balance metric offered here is derived by normalizing row-wise the standard confusion matrix of the classification outcome, so that the diagonal elements c_{11} and c_{22} of the normalized confusion matrix C express the class-wise accuracy rates. The final balance B is given by the absolute value of their difference: $B = |c_{11} - c_{22}|$.

III. Results

A. Classification Accuracy Results

The heatmaps in Fig. 4(a) and Fig. 5(a) show the run-wise and session-wise average accuracy difference between the novel FL-AFS and the conventional DD-AFS technique for able-bodied participants, AT, and rehabilitation patients, respectively. LDA classifiers and r^2 discriminancy for DD-AFS and the DP input of FL-AFS are used throughout in the main paper. In areas where the heatmap is green or, even more, yellow, there is a high, statistically significant (paired, two-sample t-test at the 95% confidence interval with Bonferroni correction) average accuracy difference in favour of the fuzzy method. On the contrary, blue encodes a slight, not statistically significant difference. For both session-wise and run-wise plots, the heatmaps show that FL-AFS delivers higher accuracy, especially for feature set cardinality greater than 5 features and noise levels within the interval 10%-50%. It is evident that the overall effect in accuracy changes with varying levels of noise and different numbers of features used; however, importantly, statistically significant and noteworthy gains are already observed for 0% noise (i.e., when the SMR data are not artificially contaminated with artifacts). The effect of our algorithm on accuracy can reach approximately 8% and 4% on average (run-wise and session-wise, respectively) for the most susceptible combinations of feature number and noise level.

Supplementary Fig. S4 and S5 provide the actual accuracy values per condition. The classification accuracy results remain virtually unaffected after replacing LDA with SVM or Random Forest classifiers (always with r^2, see the supplementary Fig. S6(a), S7(a), S8(a) and S9(a)), or after replacing r^2 by alternative FF metrics for DD-AFS and for the DP input of FL-AFS (Fisher score, Gradient Boosting and XGBoost are tested with LDA classifiers, please see the supplementary Fig. S10(a), S11(a), S12(a), S13(a), S14(a) and S15(a)).

Fig. 6 visualizes histograms of per-subject average (across each subject's runs) accuracy difference between FL-AFS and DD-AFS for 0%, 10%, 20%, 30%, and 40% noise level and 10 selected features. It can be seen that individual subjects may in fact benefit much more from the proposed algorithm than the average effects reported above, with certain subjects exhibiting impressive accuracy enhancement of up to 14%. More importantly, the few subjects that show no improvement won't suffer significant adverse effects (i.e., there are very few subjects where the application of FL-AFS leads to reduced accuracy, and the reduction is minimal). In-depth inspection reveals that the non-improving subjects are mostly the already "good" participants whose selected features were already strong and relevant; hence, the subset of users where our approach has no ground for improvements (ceiling effect).

As shown in Fig. 7, several subjects for whom the classification accuracy was below the 58% chance-level threshold (at the 95% confidence interval [3]) and who were unable to control the BCI, would have been enabled to raise their performance above this threshold with FL-AFS. Therefore, our approach also contributes to the alleviation of the serious issue of non-universal applicability to SMR BCI, where several subjects fail to control the BCI system after training [7]. Such improvements are more pronounced for 2-30 features used and 10-50% noise level. Fig. 8 represents the average improvements of the proposed FL-AFS over DD-AFS for different classifiers and different DP methods. For detailed improvements of FL-AFS see the supplementary Table S1 - S4. Additionally, the supplementary Fig. S16 and S17 illustrate the FL-AFS sensitivity to the noise level.

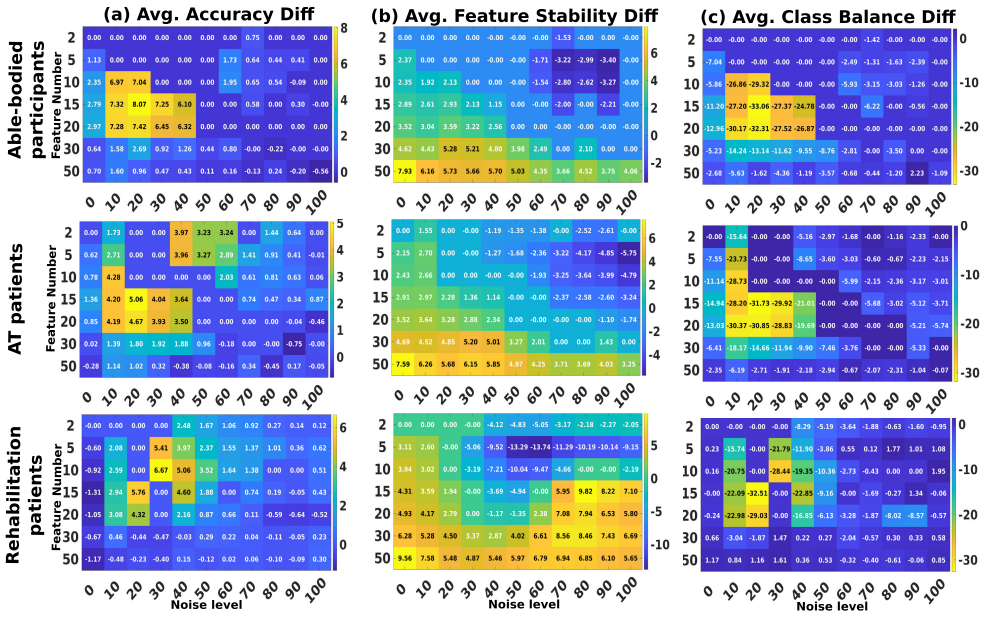

Fig. 4. Run-wise comparison of average (a) Accuracy, (b) Feature Stability and (c) Class Balance differences. x-axis represents the noise level and y-axis represents the feature number. Positive values (yellow/orange) indicate statistically significant superiority of FL-AFS over DD-AFS. Cells with 0 value indicate no statistical significance. All other cells indicate statistically significant differences at the 95% confidence interval after Bonferroni correction (paired, two-sample t-tests).

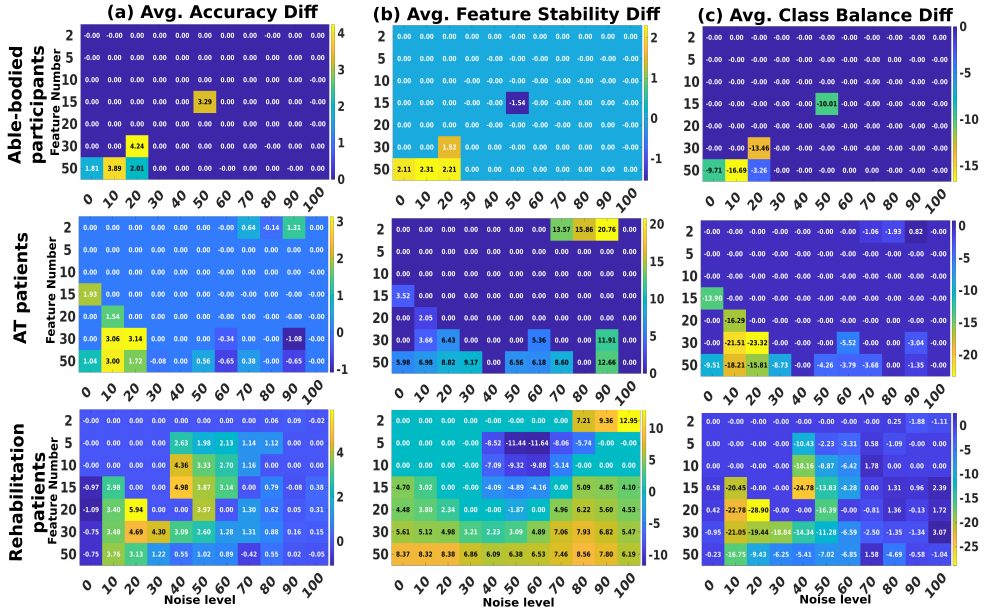

Fig. 5. Session-wise comparison of average (a) Accuracy, (b) Feature Stability and (c) Class Balance differences (see Fig. 4 for details).

B. Feature Stability Results

Fig. 4(b) and Fig. 5(b) demonstrate that the proposed FL-based feature selection method has a strong effect also concerning feature stability. Only statistically significant differences in Jaccard index values are visualised in the figure's heatmap, so that 0-valued cells indicate a small, insignificant difference between the two methods in terms of feature stability. It is noteworthy that statistically significant gains (2-5% in magnitude on average) are also noted for 0% noise level for cardinality above 20 features. Overall, the proposed method exhibits increased feature stability with a larger number of features, which is reasonable, since the modified FF in our approach mainly impacts the medium- and low-DP features (high-DP features will usually maintain high modified FF unless found in a completely irrelevant EEG channel and band for a given taskset). It is interesting to observe that run-wise stability gains are larger than session-wise gains; this could be anticipated, since non-stationarity effects, which may substantially alter the DP of features, are known to mainly occur in-between sessions rather than within a BCI session [21]. Similarly to accuracy, the feature stability improvement is independent of the classifier and DP metric used (supplementary Fig. S6- S9(b) and S10(b)- S15(b), respectively). In summary, FL-AFS helps preserve substantial overlap of features across runs and sessions, boosting stability

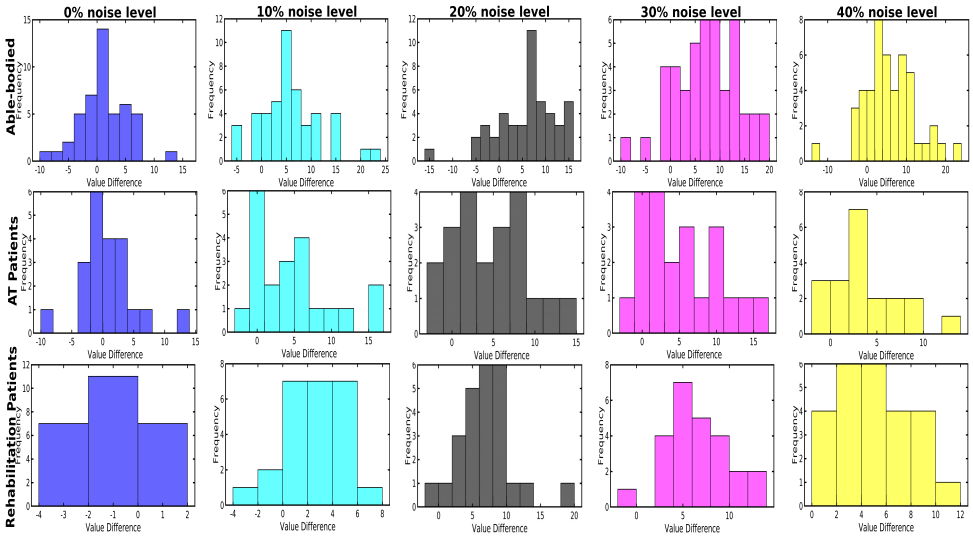

Fig. 6. Classification accuracy difference histograms between FL-AFS and DD-AFS for Able-bodied users (top) AT patients (middle) and Rehabilitation patients (bottom), for 10 selected features and 0% (blue), 10% (cyan), 20% (black), 30% (magenta), or 40% (yellow) noise level.

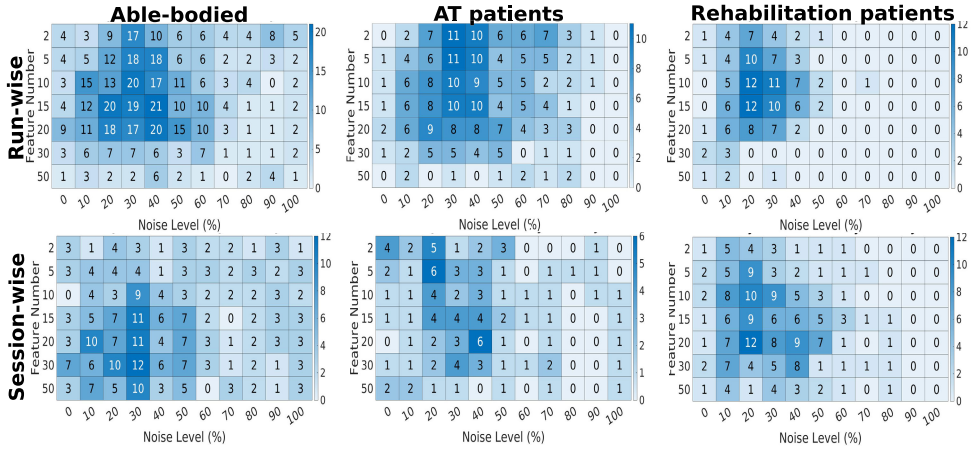

Fig. 7. Run-wise (top) and session-wise (bottom) count of participants exceeding chance classification threshold after FL-AFS.

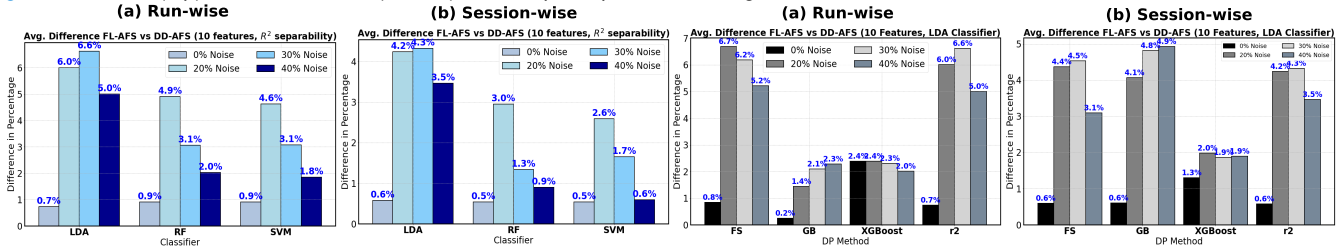

Fig. 8. Difference between FL-AFS and DD-AFS.

in the selected features by as high as 7% in the best-case scenario. Additionally, it must be underlined that feature stability improves already at low levels of noise, as well as in the absence of noise.

C. Class-Balance Results

Fig. 4(c) and Fig. 5(c) illustrate that the FL-AFS method outperforms DD-AFS also in maintaining class balance (note that, in this case, negative difference indicates superiority of FL-AFS). A pronounced and statistically significant (always with paired, two-sample t-tests, Bonferroni corrected) difference in favour of the fuzzy method is evident when the number of features ranges from 5 to 30, particularly at lower noise levels. Consistently with the other evaluation variables, class-balance is also invariant to the type of classifier and DP/FF metric used (see supplementary Fig. S6(c)- S15(c)). Notably, the proposed method manages to preserve class balance even under conditions with a high count of features and elevated noise levels. An impressive outcome is that class re-balancing in some scenarios reaches or exceeds 30%. Practically, this means that the fuzzy method is able to render a completely biased 2-class BCI into a very well-balanced system allowing the user to deliver both BCI commands with similar ease.

IV. Discussion

The major objective of this study has been to explore the feasibility and benefits of integrating expert knowledge with a data-driven approach via FL to enhance feature selection, thereby improving SMR BCI performance and user learning. We have assumed that the proposed system could yield improvements in terms of classification accuracy, the prototypical and still most widely used SMR BCI performance metric, feature stability and ability to "unbias" the BCI classifier. Our results have confirmed these assumptions and show that the magnitude of the effects across all these performance aspects could improve closed-loop SMR BCI control. We have found trends suggesting that the performance gains will tend to be greater as the noise in the data increases and when the number of selected features is large. Importantly, enhanced classification accuracy is shown to be adequate for allowing several users with initially random performance to get in control of a SMR BCI. Furthermore, we have exemplified how the proposed method's modus operandi is to eliminate noisy features that fully data-driven feature selection methods fail to identify, thanks to taking into account prior neuroscience findings embedded into the model in the form of intuitive fuzzy sets and rules. It has been shown that the proposed method outperforms solely data-driven selection irrespective of the classifier and the FF approach followed thanks to comparative analysis among various feature ranking (r^2, Fisher Score, Gradient Boosting, XGBoost) and classification methods (namely, LDA, SVM, Random Forest).

The run-wise average accuracy (using LDA and r^2) of the proposed feature selection algorithm when no noise is artificially added is (for 10 selected features) 70.74%, compared to 68.39% for the standard data-driven method. While the accuracy gap between the two methods may not seem substantial on average, it is shown that the proposed system exhibits, at the same time, more consistent and stable features than its data-driven counterpart. Importantly, this +2.3% average accuracy effect is augmented to reach +7% for the same number of features when there is artifact interference in 10-20% of trials, something that could be commonplace in real-world BCI [3], and may approach +15% for single subjects. Another crucial observation is that the distribution of accuracy difference is intensely skewed towards positive values; this implies that the application of FL-based feature selection as designed here is largely risk-free, since, even in case no accuracy improvement is observed, there will almost never be deterioration with respect to a fully data-driven approach.

All three performance facets checked here (classification accuracy, feature stability, class-balance) seem to be positively affected by our method as the noise level increases. However, it must also be highlighted that at very high noise levels (60% and above "noise level" as defined here through the % of artifact-contaminated trials), both methods collapse. This should be attributed to the difficulty of any method to select features on and classify extremely noisy data. We strongly believe that the range of 0-50% noise level where our method is impactful is within the spectrum of "natural" noise levels that could be observed in realistic BCI scenarios, also depending on the particularities of each application.

Concerning stability, we have underlined that gains are anticipated to be larger for session-wise selection, as in-between session (i.e., longer-term) non-stationarity is known to greatly affect EEG feature distributions, more so than short-term variability. It should also be noted that feature selection is anyway more likely to happen in-between sessions, as too frequent feature exchange could disrupt the user's learning [3], [7]. That being said, our method's greater feature stability probably opens the road for more frequent feature selection at the initial stages of learning, where intense user learning may lead to different features naturally (i.e., not as a result of noise interference) emerging as optimal. In other words, our method paves the way for effective online/adaptive feature selection algorithms. Of note, an adaptive version of this algorithm–if an unsupervised separability/data-driven metric can be defined–could be itself fully unsupervised, given that expert knowledge is embedded in the architecture of our model and does not need to be learned by data.

Very importantly, it must be stressed that the added value in various types of BCI performance indicators shown and discussed here is merely the one that has been established by our "offline", open-loop analysis. It is reasonable to hypothesize that, up to the extent that increased initial accuracy, feature stability and class-balance indeed yield improved subject learning through closed-loop feedback training, the actual effect of this algorithm when applied in a mutual learning [2], [3] and/or co-adaptive [7], [21], [22], [23] user training protocol could in fact be much stronger. Increased impact on closed-loop BCI performance can be anticipated, first, because we show that large numbers of subjects with random accuracy after BCI calibration with a purely data-driven method (preventing them from proceeding with feedback training) will be able, thanks to our algorithm, to achieve above-chance accuracy and thus receive meaningful feedback in closed-loop [21], potentially further improving their BCI control aptitude. Second, we assume that the neurophysiologically relevant features whose selection is fostered with our methodology may also be easier to learn to modulate with neurofeedback training [43]. Our future research will seek to shed light on the benefits that can be reaped by fusing background knowledge with ML in closed-loop.

Explainability of applied ML is emerging as a major prerequisite in neural and biomedical engineering [9]. Our method is shown to exploit the inherent transparency and intuitiveness of FL to easily incorporate any form of expert knowledge in the BCI design through simple definitions of membership functions and "human-readable" fuzzy rules. Increased model explainability makes the extension of this approach to additional MI tasksets, other BCI paradigms and objectives (e.g., explicit encoding of stability) straightforward for all prospective BCI designers and facilitates algorithmic hyperparameterization by maintaining a physical meaning for all the variables and parameters involved in the model (unlike, for example, the weights of an artificial neural network that are hard to interpret).

Concluding, this research takes a step towards bridging the machine learning versus "human factors" gap in BCI by

exemplifying how data-driven methods can be blended with neurophysiological insights and expert knowledge to improve the interaction.

REFERENCES

[1] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nat. Rev. Neurol.*, vol. 12, no. 9, p. 513, 2016.

[2] R. Leeb et al., "Transferring brain–computer interfaces beyond the laboratory: Successful application control for motor-disabled users," *Artif. Intell. Med.*, vol. 59, no. 2, pp. 121–132, Oct. 2013.

[3] S. Perdikis, L. Tonin, S. Saeedi, C. Schneider, and J. D. R. Millán, "The cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users," *PLOS Biol.*, vol. 16, no. 5, May 2018, Art. no. e2003787.

[4] A. Biasiucci et al., "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke," *Nature Commun.*, vol. 9, no. 1, Jun. 2018, Art. no. 2421.

[5] N. Birbaumer et al., "A spelling device for the paralysed," *Nature*, vol. 398, no. 6725, pp. 297–298, Mar. 1999.

[6] I. Iturrate, R. Chavarriaga, and J. del R. Millán, "General principles of machine learning for brain–computer interfacing," in *Handbook of Clinical Neurology*, vol. 168, N. F. Ramsey and J. del R. Millán, Eds., Amsterdam, The Netherlands: Elsevier, 2020, ch. 23, pp. 311–328.

[7] S. Perdikis and J. del R. Millán, "Brain–machine interfaces: A tale of two learners," *IEEE Syst. Man, Cybern. Mag.*, vol. 6, no. 3, pp. 12–19, Jul. 2020.

[8] J.-S. Bang, M.-H. Lee, S. Fazli, C. Guan, and S.-W. Lee, "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3038–3049, Jul. 2022.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[10] T. J. Ross, *Fuzzy Pattern Recognition*. Hoboken, NJ, USA: Wiley, 2010, ch. 11, pp. 369–407.

[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[12] F. Lotte et al., "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Apr. 2018, Art. no. 031005.

[13] S. Bhattacharyya, A. Sengupta, T. Chakraborti, A. Konar, and D. N. Tibarewala, "Automatic feature selection of motor imagery EEG signals using differential evolution and learning automata," *Med. Biol. Eng. Comput.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[14] M. A. M. Joadder, J. J. Myszewski, M. H. Rahman, and I. Wang, "A performance based feature selection technique for subject independent MI based BCI," *Health Inf. Sci. Syst.*, vol. 7, no. 1, p. 15, Aug. 2019.

[15] M. K. I. Molla, A. A. Shiam, M. R. Islam, and T. Tanaka, "Discriminative feature selection-based motor imagery classification using EEG signal," *IEEE Access*, vol. 8, pp. 98255–98265, 2020.

[16] J. Fruitet, D. J. McFarland, and J. R. Wolpaw, "A comparison of regression techniques for a two-dimensional sensorimotor rhythm-based brain–computer interface," *J. Neural Eng.*, vol. 7, no. 1, Feb. 2010, Art. no. 016003.

[17] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial–temporal patterns for single-trial EEG-based BCIs," *Pattern Recognit.*, vol. 45, no. 6, pp. 2137–2144, Jun. 2012.

[18] M. Bevilacqua, S. Perdikis, and J. d. R. Millán, "On error-related potentials during sensorimotor-based brain–computer interface: Explorations with a pseudo-online brain-controlled speller," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 17–22, 2020.

[19] M. Radman, A. Chaibakhsh, N. Nariman-zadeh, and H. He, "Feature fusion for improving performance of motor imagery brain–computer interface system," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102763.

[20] O. Z. Ying, S. A. B. Awang, and V. A. Vijean, "Optimization of irrelevant features for brain–computer interface (BCI) system," *J. Phys., Conf. Ser.*, vol. 1372, no. 1, Nov. 2019, Art. no. 012047.

[21] S. Perdikis, R. Leeb, and J. D. R. Millán, "Context-aware adaptive spelling in motor imagery BCI," *J. Neural Eng.*, vol. 13, no. 3, Jun. 2016, Art. no. 036018.

[22] J. D. Cunha, S. Perdikis, S. Halder, and R. Scherer, "Post-adaptation effects in a motor imagery brain–computer interface online coadaptive paradigm," *IEEE Access*, vol. 9, pp. 41688–41703, 2021.

[23] J. Faller, R. Scherer, U. Costa, E. Opisso, J. Medina, and G. R. Müller-Putz, "A co-adaptive brain–computer interface for end users with severe motor impairment," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e101168.

[24] S. Perdikis, R. Leeb, R. Chavarriaga, and J. d. R. Millán, "Context-aware learning for generative models," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3471–3483, Aug. 2021.

[25] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ. Press, 1976.

[26] J.-A. Martinez-Leon, J.-M. Cano-Izquierdo, and J. Ibarrola, "Feature selection applying statistical and neurofuzzy methods to EEG-based BCI," *Comput. Intell. Neurosci.*, vol. 2015, Apr. 2015, Art. no. 781207.

[27] A. G. Trofimov, S. L. Shishkin, B. L. Kozyrskiy, and B. M. Velichkovsky, "A greedy feature selection algorithm for brain–computer interface classification committees," *Proc. Comput. Sci.*, vol. 123, pp. 488–493, Jan. 2018.

[28] Z. Tao, "Classification-oriented fuzzy-rough feature selection for the EEG-based brain–computer interfaces," in *Proc. 13th Int. Conf. Manage. Sci. Eng. Manage.*, J. Xu et al., Eds., Cham, Switzerland: Springer, 2020, pp. 295–307.

[29] R. Chatterjee, D. K. Sanyal, and A. Chatterjee, "Fuzzy-discernibility matrix-based efficient feature selection techniques for improved motor-imagery EEG signal classification," *bioRXix*, 2021, doi: 10.1101/2021.03.24.436722.

[30] P. A. Herman, G. Prasad, and T. M. McGinnity, "Designing an interval type-2 fuzzy logic system for handling uncertainty effects in brain–computer interface classification of motor imagery induced EEG patterns," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 1, pp. 29–42, Feb. 2017.

[31] A. Gupta and D. Kumar, "Fuzzy clustering-based feature extraction method for mental task classification," *Brain Informat.*, vol. 4, no. 2, pp. 135–145, Jun. 2017.

[32] N. Saga, A. Doi, T. Oda, and S. N. Kudoh, "Elucidation of EEG characteristics of fuzzy reasoning-based heuristic BCI and its application to patient with brain infarction," *Frontiers Neurorobot.*, vol. 14, p. 123, Jan. 2021.

[33] M. Sultana, C. Reichert, C. Sweeney-Reed, and S. Perdikis, "Towards calibration-less BCI-based rehabilitation," in *Proc. IEEE Int. Conf. Metrol. eXtended Reality, Artif. Intell. Neural Eng. (MetroXRAINE)*, Oct. 2023, pp. 11–16.

[34] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoustics*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.

[35] Z. Jin, F. Bourban, R. Leeb, and S. Perdikis, "Quantifying the impact and profiling functional EEG artifacts," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2022, pp. 629–634.

[36] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain–computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, Jul. 2001.

[37] W. Yi, S. Qiu, H. Qi, L. Zhang, B. Wan, and D. Ming, "EEG feature comparison and classification of simple and compound limb motor imagery," *J. NeuroEngineering Rehabil.*, vol. 10, no. 1, pp. 1–12, Dec. 2013.

[38] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 51, pp. 17849–17854, Dec. 2004.

[39] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, "Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks," *Neuroimage*, vol. 31, no. 1, pp. 153–159, May 2006.

[40] T. V. Avdeenko and E. S. Makarova, "Integration of case-based and rule-based reasoning through fuzzy inference in decision support systems," *Proc. Comput. Sci.*, vol. 103, pp. 447–453, Jan. 2017.

[41] L. A. Zadeh and R. A. Aliev, *Fuzzy Logic Theory and Applications: Part I and Part II*. Singapore: World Scientific, 2018.

[42] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, Oct. 2010.

[43] P. T. Sadtler et al., "Neural constraints on learning," *Nature*, vol. 512, no. 7515, pp. 423–426, Aug. 2014.