

A combined EEG motor and speech imagery paradigm with automated successive halving for customizable command selection

Natasha Padfield^a, Tracey Camilleri^b, Simon Fabri^b, Marvin Bugeja^b and Kenneth Camilleri^{a,b}

^aCentre for Biomedical Cybernetics, University of Malta, Msida, Malta; ^bDepartment of Systems and Control Engineering, University of Malta, Msida, Malta

ABSTRACT

The classification performance of endogenous electroencephalogram (EEG) brain-computer interfaces (BCIs) can be improved by hybridizing the paradigm through the use of commands from multiple paradigms. Hybrid paradigms using motor imagery (MI) and speech imagery (SI) have shown promise, but there is a lack of research into: i) their effectiveness when compared to pure MI and SI for multiclass problems, and ii) automated command selection. This study investigates multiclass MI and SI hybrid paradigms and compares the results to those obtained using pure MI and SI. Performance was assessed using F1 score and accuracy. The performances of all possible hybrid paradigm designs were assessed. The analysis indicated that hybridization does not always guarantee improved performance when compared to the pure paradigms, and there is inter-subject variation in the best paradigm. This confirmed the need for automated subject-specific hybrid paradigm designs. An automated hybrid paradigm selection technique using successive halving (SH) for expedited computational times was developed and results were compared to those obtained using a standard grid search. The SH approach resulted in an improvement in F1 score of 21.09% and 36.86% compared to MI and SI and led to a reduction in computational times of 82.80% compared to grid search.

ARTICLE HISTORY

Received 7 February 2024
Accepted 8 July 2024

KEYWORDS

Electroencephalogram;
brain-computer interface;
motor imagery; speech
imagery; successive halving



1. Introduction


Wider adoption of electroencephalogram (EEG)-based brain-computer interfaces (BCIs) depends, in part, on the use of reliable and robust paradigms for issuing commands. There are two major kinds of EEG-based BCI paradigms: exogenous, which require the subject to interact with an external stimulus, for example by looking at a flickering light, and endogenous, in which the subject executes commands themselves, without any stimulus. The most widely adopted endogenous paradigm is motor imagery (MI) [1], which requires subjects to imagine movements of limbs to execute commands. Currently, exogenous paradigms provide better classification performance when compared to endogenous ones. However, exogenous paradigms can be impractical, tiring for the user, and possibly unintuitive [2,3]. Thus, there is extensive research into improving the classification performance when using endogenous paradigms [1,4–7].

Various approaches for improving the classification of endogenous commands have been proposed. Predominantly, the focus of previous research has been on exploring different signal processing methods,

features, and classifiers [1,8]. Studies have also investigated the use of selection methods to choose subsets of features or EEG channels which result in improved performance [9–16]. Other approaches involve implementing novel pre-processing algorithms to clean the data through the removal of artifacts [17] or to improve its decipherability through decomposition methods [18]. Post-processing of the classifier output to stabilize classification performance [19] has also been investigated. These approaches all use computational methods to improve EEG classification.

An alternative approach for improving the classification performance of endogenous systems is to design the paradigm such that the commands themselves are more likely to be strongly decipherable [20,21]. This can be achieved through hybridization of the paradigm, which involves including commands from two or more endogenous paradigms [1,4–7,20–22]. Previous studies [20,21] have indicated that hybrid paradigms can outperform pure paradigms (which consist of commands from just one paradigm, such as MI alone) [20,21]. Two studies [20,21] have found that combining

CONTACT Natasha Padfield  natasha.padfield@um.edu.mt  Centre for Biomedical Cybernetics, University of Malta, Room 211, Engineering Building, Msida MSD 2080, Malta

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/2326263X.2024.2379009>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

MI and speech imagery (SI, which involves imagining speaking words) led to improved performance in binary classification problems when compared to using standard left- and right-hand MI classes. Two overarching gaps were identified in the literature related to hybrid paradigms: i) there was a lack of research into the effectiveness of hybrid paradigms in improving classification performance, and ii) there was a lack of automated command selection within hybrid paradigm frameworks. These two gaps and associated literature will be discussed in turn.

The first gap identified was a lack of exhaustive analysis of the impact of hybridization on classification performance. Most studies do not compare the hybrid paradigms to a pure paradigm at all [1,4–7,22] whilst those that do have based their studies on binary classification problems alone [20,21], which would be limiting in a practical BCI where multiclass classification is generally necessary to facilitate a variety of commands. Although these initial studies have indicated that hybridization can lead to improved performance, there are still questions as to whether hybridization will always lead to improved performance, and if not, how to effectively select suitable commands.

The second gap identified was in the area of automated command selection for hybrid paradigms. Three main approaches for command selection within hybrid paradigms were identified. The most straightforward involved allowing participants to pick which commands they would like to use, but this does not mean that they will pick those which lead to the best performance [4]. The second approach evaluated the classification performance of each command combination to select the most appropriate [5] using a grid search method. Although this approach is rigorous, it can be computationally expensive. The third approach was the interactive method by Kuzovkin et al. [7]. The user was allowed to explore different kinds of commands, with the interface giving feedback to the user on their classification performance and prompting them to change the command that was detected most poorly. This approach requires extended periods of concentration and puts the onus on the user to identify a set of commands that work well together. If the subject experiments with various commands throughout the interactive session, it is possible that a particular subset of commands experimented with could potentially lead to improved performance compared to the final command set settled on through the interactive process due to the ‘greedy’ nature of the process.

Although automated hybrid paradigm selection has not been investigated in depth in the literature, selection algorithms have been applied to other problems in EEG

processing, most notably to select the best EEG channels or features to optimize classification performance [9–16]. The most popular selection algorithms which could also be applied to hybrid paradigm selection can be divided into three main categories: i) exhaustive searches like the grid search approach, which evaluates all possible options [12,16]; ii) greedy search methods such as recursive feature elimination and forward selection [13,15], and iii) metaheuristic and evolutionary methods [9,10,13,14,23], with popular metaheuristic and evolutionary algorithms used in the literature including genetic algorithms [9,11,13], swarm-based optimization [9,10,14,23] and differential evolution [9,24]. Aljalal et al. [13] found that metaheuristic approaches involving genetic algorithms performed better than greedy methods based on backward and forward feature selection. The main limitation of exhaustive searches is that they are computationally expensive. In a practical scenario, the hybrid paradigm selection process would contribute to the delay between recording the training data from the subject and the subject being able to use the BCI. Thus, the selection process must be computationally efficient such that this delay is brief. Furthermore, although greedy and metaheuristic algorithms try to locate the global optimal solution, they are not guaranteed to explore the entire solution space, meaning that there is a risk that the best-performing hybrid command combination may not even be considered by the algorithm. Finally, metaheuristic algorithms in particular have various parameters that need to be set and there is no guarantee that parameters set at the design stage would be universally appropriate for all potential users of a BCI. If to try and ensure better performance for individual users, these parameters are tuned for each new user, this tuning process will contribute to an increase in the delay experienced by the user, which is contrary to the aim of introducing automated selection.

Successive halving (SH) [25–27] is a state-of-the-art selection algorithm that has been used as a hyperparameter tuning approach and was designed to be more computationally effective than traditional grid search hyperparameter tuning. In SH, the number of candidates is halved in each successive iteration, and the number of resources used is doubled. SH traditionally continues iteratively until one candidate has been selected. Thus, SH expedites the selection process by exploiting the fact that the evaluation process gets slower as the volume of data used increases. Therefore, when performing the initial survey of the whole candidate space in the first iteration, the smallest amount of data is used. Then, as the algorithm narrows down the candidate space, the amount of data used is increased to ensure

that the best candidate is selected. In previous works, SH has been applied to hyperparameter tuning for deep learning algorithms [25–27] and a naïve Bayes [26] classifier for different machine learning problems such as wine recognition [26], cancer detection [26], image classification [25], and electrical interface signal processing [27]. To the authors' knowledge SH has not been applied to the problem of hybrid paradigm selection, although it is highly suitable because it presents solutions to the shortcomings of other selection algorithms in the EEG literature, namely: i) SH is designed to explore the entire solution space, ii) by nature, SH is not heuristic and is non-parametric [25–27], meaning that it can be applied to the data of any new subject without the need to worry about parameters being sub-optimal or about additional delays to tune parameters to the individual subject, and iii) SH has been designed to be more computationally efficient than a grid search [25].

The objective of this study, based on this literature review, is to improve the classification of multiclass endogenous EEG commands. This is achieved through: i) exhaustively evaluating the impact of a MI and SI hybrid paradigm on multiclass classification performance when compared to the pure paradigms, and ii) proposing a computationally efficient approach, based on SH, to select commands for each subject which lead to improved performance when compared to the pure paradigm.

The rest of the paper is structured as follows: Section 2 presents the Materials and Methods, Section 3 presents the Results, Section 4 presents the Discussion, and finally Section 5 presents the Conclusion.

2. Materials and methods

2.1. Data recording

Before data recording, ethical approval was sought, and project details were submitted to and acknowledged by

the Faculty Research Ethics Committee at the Faculty of Media and Knowledge Sciences of the University of Malta (application number MAKS-2022-00012). The principles of the Declaration of Helsinki were respected. Informed consent was obtained: subjects were given an information sheet and signed a consent form to participate. The consent form explained that their EEG data may be made publicly available. Five healthy subjects participated in this study, two males and three females, with an average age of 24.4 years. Data was recorded at a sampling frequency of 2.048 kHz using the BioSemi ActiveTwo EEG measurement system [28] with 32 active gel electrodes. The data has been made publicly available at: <https://doi.org/10.60809/drum.24465871.v1>

EEG data was recorded for four MI classes and four SI classes. For the former, left-hand, right-hand, tongue, and legs MI, which constitute the standard hands, legs, and tongue paradigm [29], was recorded. For the SI classes, EEG data for the words 'left', 'right', 'up', and 'down' was recorded. These words were chosen because they can easily be used for cursor control in a graphical user interface (GUI) or to control a robotic device.

A total of 40 trials were recorded for each of the four MI and four SI classes. Each trial had a duration of 6s and was structured as shown by the timing diagram in Figure 1. First, a fixation cross appears on-screen, indicating to the subject to remain relaxed but aware that the next trial is forthcoming. The cue then appears in the form of an arrow, with its direction being associated with a particular task. The subject starts executing the task as soon as they see the cue, and continues even when it has disappeared until the fixation cross appears again. The cues consist of a left-facing arrow (for left-hand MI or 'left' SI), a right-facing arrow (for right-hand MI or 'right' SI), an upward-facing arrow (for tongue MI or 'up' SI), and a downward-facing arrow (for legs MI or 'down' SI).

For data collection, four separate runs were carried out, where a 'run' refers to a series of 20 trials per class

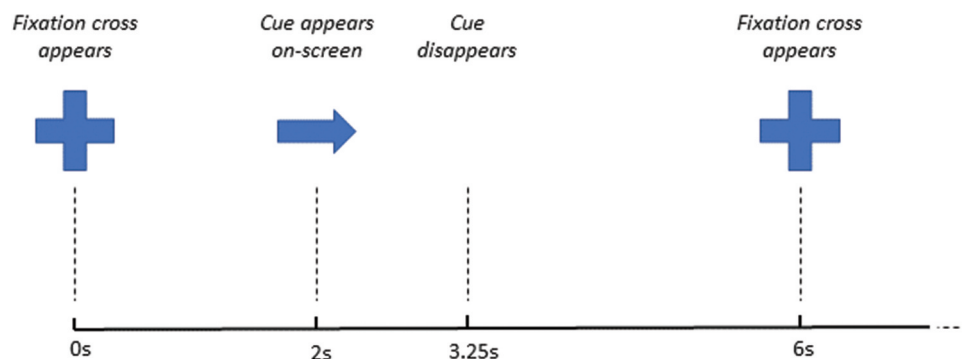


Figure 1. Timing diagram for each 6s-long trial.

being recorded. Between each run, subjects were given a short break of 3–5 minutes. In the first and third runs, MI data was recorded, and then in the second and fourth runs, SI was recorded. At the start of a run, subjects were given one minute to settle down before the trials began.

2.2. Data pre-processing

Channel Cz was the recorded data reference channel. The data was then re-referenced by calculating the mean value across the channels at each time point and subtracting the resulting vector from all channels (common average referencing). A bandpass filter with a passband between 1 Hz and 120 Hz was applied to the data, followed by a 50 Hz notch filter. The data was then down-sampled to 1 kHz. Each trial was then augmented by dividing it into 32 segments using overlapping windows of length 2s and spaced out by 0.063s. These parameters were chosen since they replicate the buffering used in online BCI systems in the literature [19]. For classification, 24 channels out of the 32 available in the BioSemi setup were used. Channel Cz was removed since it was the reference, the anterior-frontal and frontopolar channels (four in total) were removed due to their high correlation with artifacts [30], and the electrodes O1, O2, and O3 were not used in this analysis due to poor signal quality for one of the subjects. The 24 electrodes used were: F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, and Fz.

2.3. Feature extraction and classification

Power spectral density (PSD) features together with a support vector machine (SVM) classifier were used for classification. This processing pipeline was chosen due to the widespread established use of this feature-classifier pairing within the literature for MI and SI classification [1,31–33]. All signal processing was implemented in MATLAB.

PSD features were extracted from each 2s-long segment of data as follows. For each channel in the dataset, the Welch PSD was calculated using a 500 samples-long (i.e. 0.5s long) Hann window, with an overlap of 250 samples and a zero-padding vector of size 500. PSD values for frequencies from 1 Hz to 120 Hz, spaced by steps of 1 Hz, were extracted. The average powers within specific frequency bands were then calculated, thus obtaining a set of features from each channel. The features from all channels were concatenated to form the final feature vector.

Throughout this paper, separate results are presented using three different approaches for segmenting the frequency bands as described below. For each approach, baseline results using pure MI and SI paradigms were obtained, as well as results using the hybrid paradigm. The goal was to assess the versatility and generalizability of the hybrid paradigm using the three different feature types and identify which frequency feature works best with the hybrid paradigm.

The first approach used for frequency band segmentation involves the extraction of features from the standard, established EEG frequency bands, namely the delta, theta, alpha, beta, and gamma bands [34,35]. This standard method of frequency band stratification has been widely used in analysis throughout the EEG literature [8,34,35]. It is fairly common for the alpha, beta, and gamma frequencies to be further stratified into sub-bands [34,35]. Hence, the following divisions were used: delta (1–4) Hz, theta (4.5–7.5) Hz, low alpha (8–10) Hz, high alpha (10.25–13) Hz, low beta (13.25–18) Hz, medium beta (18.25–21) Hz, high beta (21.25–30) Hz, low gamma (30.5–45) Hz, medium gamma (45.5–58) Hz, high gamma (58.5–85) Hz, and ultra-high gamma (85.5–120) Hz. This first method is referred to as the ‘Standard Bands’ approach throughout this paper.

The second approach was used in the work of Koizumi et al. [31], who investigated an SI classification approach, in which PSD features were extracted from 10 Hz frequency bands, delineated as follows: $f < 10$ Hz, $(10 \leq f < 20)$ Hz, $(20 \leq f < 30)$ Hz, \dots , $(110 \leq f < 120)$ Hz, where f is the frequency. This method is referred to as the ‘10 Hz Bands’ approach.

The final feature extraction approach involved features from the alpha and beta bands only. Extensive previous research suggests that the alpha and beta bands are particularly important for MI, and have also been linked to SI [8]. Using features in these two bands only and comparing them to results from the Standard Bands and 10 Hz bands approaches will indicate whether they capture sufficient salient information for the hybrid classification system or whether bands outside of these also have substantial influence. Furthermore, since MI was the primary benchmarking paradigm in these experiments, and these two bands are particularly important for MI classification, generating baseline MI results for these bands was important to ensure that frequencies outside of these bands do not confound the benchmarking results. This method is referred to as the ‘Alpha-Beta Bands’ approach throughout the rest of the paper.

An SVM with a radial basis function kernel was used as the classifier. Multiclass SVM classification was carried out using a one-vs-one configuration

through the inbuilt *fitecoc* function [36]. This function reframes the multiclass classification problem into a collection of binary classification problems using an error-correcting output codes model [36]. The box constraint parameter and kernel scale parameters of the basis classifiers were tuned for values in the range N^k for N set to either 2 or 10, and k values from 2 to 6, increased in steps of 2. Tuning was carried out using a grid search on the training data, with the parameter set resulting in the greatest classification performance being applied to the test set. Four class classification was carried out in this analysis due to the four tasks the subjects carried out as described in Section 2.1.

2.4. Experimental methodology

2.4.1. Cross-validation

Subject-specific training was carried out, meaning that classifiers were trained using only the data from the individual subject. Ten-fold cross-validation was used for evaluating classification performance. The trials were divided into 10 groups in a stratified way, meaning that there were 4 trials of each class in each group. Trial segmentation was carried out such that the time chronology for each class was preserved. For each fold, one group was selected for testing, and the remaining 9 groups were aggregated into the training dataset and used for hyperparameter tuning of the SVM classifier. Once the hyperparameters were selected based on cross-validation in the training dataset, the final classifier was trained with the training data. Then, the performance with the test set was evaluated and stored, and the process was repeated such that each of the 10 groups was considered as the test set. The final test-set performance was obtained by averaging the results over the 10 folds. The approach used for tuning and evaluation is in keeping with established techniques used in the literature [37–39].

2.4.2. Performance measures

Macro classification performance measures were used. This means that for each class in the classifier, the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were extracted. Then, the performance measure was calculated individually for each class and averaged across the classes to obtain an overall classification measure for the classifier.

It should be noted that although the classifier is trained and tested on a balanced dataset, the macro performance measures approach obtains values in

a one-vs-rest fashion for each class before averaging across the four classes, meaning the performance calculation is carried out on imbalanced data (i.e. one class vs the three other classes grouped as one). Factoring this, the performance measures were chosen to accurately represent the performance.

The F1 score was the leading performance measure used in this study. It is deemed as a more informative statistic for BCI classification than the traditional classification accuracy, which can be skewed for unbalanced data [40]. F1 score has also been used as a performance measure in similar studies [7,41,42]. The equation for F1 score is as follows:

$$F_1 \text{ score} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

The sensitivity and positive predictive value (PPV) were also recorded, and calculated as follows [39]:

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$PPV = \frac{TP}{TP + FP} \quad (3)$$

The sensitivity indicates the rate at which commands were correctly identified, and the PPV is a complementary measure to test for oversensitivity [40]. Typically, there is a trade-off in the values of sensitivity and PPV [40]. The F1 score is a balanced average of these two values which can be expressed as follows:

$$F_1 \text{ score} = \frac{2 \times PPV \times \text{sensitivity}}{PPV + \text{sensitivity}} \quad (4)$$

An increase in the F1 score thus indicates an overall increase in performance, based on these two statistics. The sensitivity and PPV value are also reported to identify the underlying mechanisms through which the F1 score changes.

The balanced accuracy [43–46], which is an average of the sensitivity and specificity, captures the inherent trade-off between the two: ideally, both measures are high, leading to better overall performance, which is in turn captured by the balanced accuracy. The balanced accuracy is complementary to the F1 score since it takes into consideration the true negatives, which the F1 score does not. Moreover, since macro-averaging of performance measures inherently introduces a class imbalance in the calculations, the balanced accuracy is a more appropriate measure of performance than the standard

classification accuracy [43–46]. The equations for specificity and balanced accuracy are as follows:

$$\text{specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{balanced accuracy} = \frac{1}{2}(\text{sensitivity} + \text{specificity}) \quad (6)$$

The final performance measure used is computational execution time. A computational time for each subject was calculated by recording the execution time of each of the ten folds and then finding the average time. To obtain a global figure, the execution times obtained for each subject were then averaged. This means that computational times in this study were obtained based on 50 readings. Computational time readings were recorded on a desktop computer with an Intel®Core™ i7–10700 CPU running at 2.90 GHz, 36GB of RAM, and the Windows 10 Pro 64-bit operating system. Experiments were run using MATLAB 2023a. Other applications were closed, and non-essential background processes were suspended to ensure accurate readings.

Percentage changes in performance measures were calculated by taking the difference between the nominal value and the comparison value, and then dividing by the nominal value.

2.4.3. General evaluation of the hybrid paradigm

A general evaluation of the impact of the MI-SI paradigm hybridization on classification performance was first carried out. In this analysis, every possible hybrid pairing of two MI and two SI tasks was considered, amounting to 36 combinations. For each hybrid pairing, the SVM classifier was tuned, and the average F1 score and balanced accuracy were obtained using the 10-fold cross-validation approach previously described in Section 2.4.1.

Benchmarking results were obtained using: i) the pure MI paradigm which consists of the four MI classes, and ii) the pure SI paradigm, which consists of the four SI classes.

2.4.4. Automated hybrid paradigm selection

In a practical BCI, the best combination of four commands for the hybrid paradigm should be automatically selected from the subjects' training data. Two approaches were applied for this purpose: i) a grid search approach, and ii) a successive halving (SH) approach.

2.4.4.1. Grid search hybrid paradigm selection. The grid search method used the same 10-fold cross-validation tuning approach used to tune the classifier. Specifically, in each fold, the cross-validation algorithm looped through all the possible hybrid combinations of two MI and two SI commands, tuning an SVM classifier for each combination, and then using the training dataset, identified the best hybrid design based on classification performance. The term 'design' refers to the specific combination of two MI and two SI commands. The test set was then used to evaluate the performance of the selected hybrid pairing and SVM parameters. This mimics how tuning would be carried out in a practical application of the hybrid paradigm, where the best pairing would be selected using the training data.

2.4.4.2. Successive halving hybrid paradigm selection.

The SH-based method used in this study is an iterative process, which involves halving the number of candidate solutions (i.e. hybrid paradigm designs) considered in each iteration, whilst doubling the amount of data used for processing. The selection process is designed such that at the end of the iterations, two remaining candidate solutions are being considered, and 100% of the data is used to decide between them. Since the hybrid paradigm consists of two MI and two SI commands, and these commands are chosen from a pool of four MI commands and four SI commands, there are 36 candidate paradigms in total.

Figure 2 shows how SH was applied to the hybrid paradigm selection problem in this study. Since 36 can be successively halved five times before a value of two remains, five iterations are carried out. To calculate how much data was used in each iteration, 100% was successively halved five times, to give a starting data percentage of 6%. Thus, in the first iteration, 6% of the data is used to evaluate all 36 candidates and select the best half (i.e. 18), which are promoted for consideration in the second iteration. In the second iteration, the data is doubled, meaning 12% of the data is used to identify the best half of the candidates (i.e. 9), which are then promoted to the third iteration. This continues until the final iteration is reached, in which 100% of the data is used to evaluate the 2 remaining candidates. Note that since during four out of the five iterations, SH is using a subset of the training data, and is selecting candidates based on these randomly selected subsets, SH has the potential to select a different candidate than the grid search approach, which uses 100% of the data to evaluate all 36 candidates.



Figure 2. Successive halving applied to the hybrid paradigm selection problem. The green circles visualize the number of candidates being considered in each iteration, and the purple circles visualize the amount of data being used to evaluate the candidates. In each iteration, the best-performing half of the candidates are selected for promotion to the next iteration.

For each iteration, the data is randomly selected in a stratified way such that the data used to train is always balanced. Evaluation is based on the classification performance, with the candidates exhibiting the best performance being promoted.

2.4.5. Statistical analysis

Statistical tests throughout the results section consist of comparisons between two groups of data. Before carrying out these tests, both groups were checked for normality using an Anderson-Darling test. If both groups of data were normal, a t-test was used to compare the two groups. Otherwise, its non-parametric counterpart, the Wilcoxon rank-sum test was used. A 0.05 level of significance was adopted, meaning that p -values below 0.05 were considered to indicate a statistically significant result.

For the statistical analysis, the ten cross-validation results obtained for each subject were concatenated to make result vectors of size 50, which were then compared. Statistical analysis was carried out to compare:

- The classification performance of the selected hybrid paradigm to the pure MI.

- The computational times for hybrid paradigm selection using the grid search method and using the SH method.
- Separate comparisons for the F1 score, balanced accuracy, sensitivity, and PPV were carried out.

3. Results

3.1. Results for general evaluation of the hybrid paradigm

The results in this section were obtained using the methodology described in Section 2.4.3. The scatter plots in Figure 3 show the F1 score and balanced accuracy results, averaged across subjects, obtained for each of the 36 hybrid paradigm combinations with the three different feature extraction approaches: Alpha-Beta Bands, Standard Bands, and 10 Hz Bands. The x-axes denote the code number for the hybrid paradigm combination, with Table A1 in the Supplementary Material showing all the possible hybrid combinations and their associated identifying codes. In Figure 3, the average results obtained when using the pure MI paradigm are denoted by the black horizontal lines, and similar results for the pure SI paradigm are denoted by the red horizontal lines.

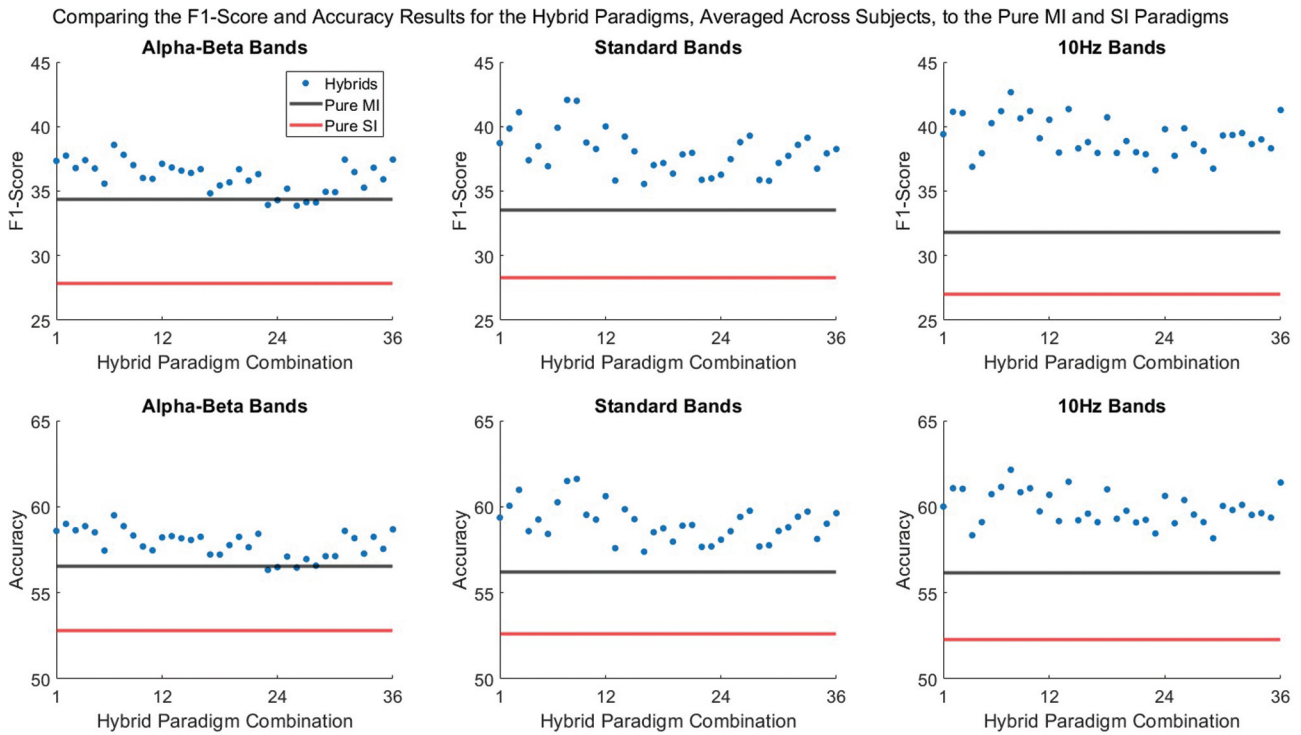


Figure 3. Comparing the F1 score and balanced accuracy results, averaged across subjects, for the hybrid paradigms (blue scatter points) to the performance obtained with the pure MI (black line) and SI (red line) paradigms. Results are shown for the three different feature extraction approaches. The x-axes denote the hybrid paradigm combination code as in supplementary table A1.

Table 1. The F1 score and balanced accuracy of the best-performing paradigm for each feature extraction approach. The numbers in brackets next to the hybrid paradigms denote the hybrid paradigm code number used along the x-axis of the plots in Figure 3. These codes are also used in Table A1.

Feature Extraction Method	Average F1 Score	Hybrid Paradigm
Alpha-Beta Bands	39.08%	Right-hand; left-hand; 'left'; 'down' (5)
Standard Bands	41.55%	Right-hand; tongue; 'right'; 'up' (8)
10 Hz Bands	41.14%	Tongue; legs; 'up'; 'down' (36)

The best performing hybrid paradigms, on average, for each feature extraction method are shown in Table 1, together with their associated F1 scores.

Additionally, the authors identified two intuitive command sets that emerge from the combinations under consideration, namely: i) left-hand MI, right-hand MI, SI 'up' and SI 'down' (called 'I1'), and ii) legs MI, tongue MI, 'left' and 'right' (called 'I2'). These command sets were chosen as the most intuitive because they have strong directional commands associated with the left and right directions, and left and right directional commands are important in many BCIs, from cursor control in a GUI to controlling the turning direction of a robot. Figure 4 compares the performance of these intuitive paradigm sets to the results for pure MI and pure SI.

Figure 5 then shows the F1 scores obtained for each of the 36 hybrid paradigms for Subjects 1 and 2 to illustrate inter-subject variability that can exist in the performance of different hybrid paradigms. Results for each of the three feature extraction approaches are shown, and the baseline results for the pure MI and pure SI paradigms are represented by the black and red horizontal lines, respectively.

3.2. Results for automated hybrid paradigm selection

Whereas in Section 3.1 the results were focused on presenting the impact of hybridization on average performance across the population for all possible hybrid paradigm command combinations, the results

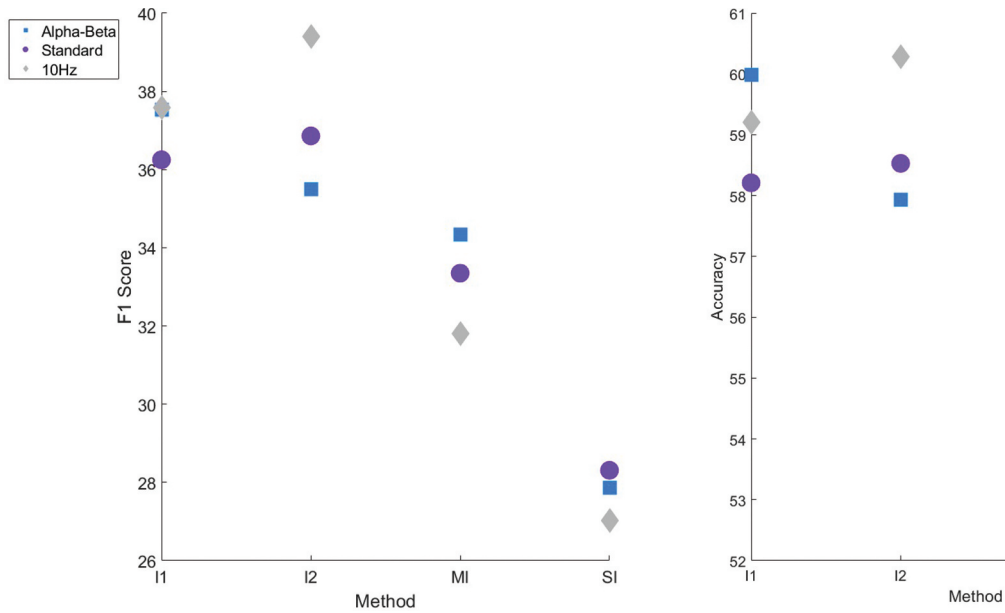


Figure 4. The classification performance of some intuitive command sets I1 and I2 compared to the pure MI and pure SI paradigms. Each data point represents the performance, averaged across all subjects.

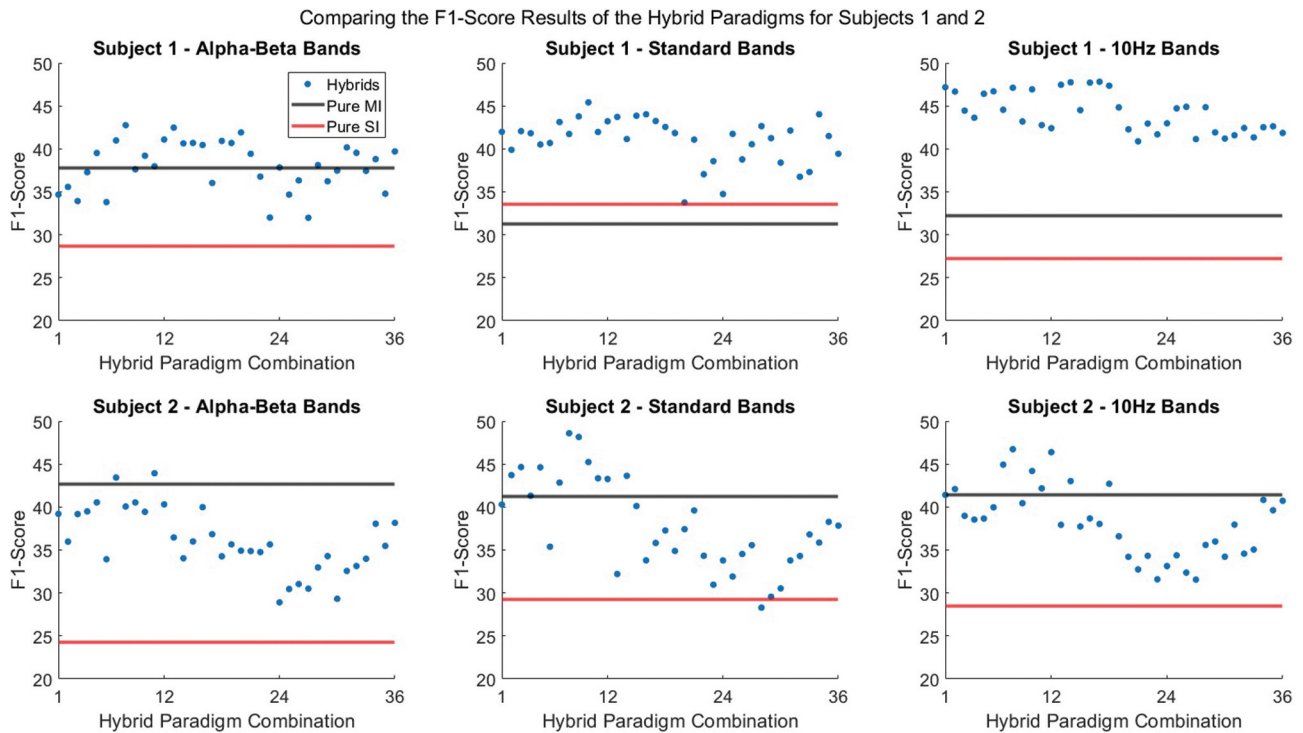


Figure 5. Comparing the performance of the hybrid paradigms (blue scatter points) for two subjects. The baseline results obtained using pure MI (black lines) and pure SI (red lines) for each subject are shown. The x-axes denote the hybrid paradigm combination code as in supplementary Table A1.

in this section capture the performance when the hybrid paradigm commands are customized for each subject through an automated selection process. The results in this section were obtained using the methodology described in Section 2.4.4.

Table 2 shows the F1 score results obtained using the hybrid paradigm designs selected through grid search and SH, and compares them to the pure MI and SI paradigms. Results for each subject and feature are shown. This table also contains p -values that were

Table 2. The F1 score results (%) obtained for individual subjects when using the pure MI paradigm, the pure SI paradigm, and the hybrid paradigms selected using grid search (GS) and successive halving (SH), for classification with the Alpha-Beta Bands, the Standard Bands, and the 10 Hz Bands features. In each row, the peak benchmarking result is highlighted in bold as well as the peak hybrid paradigm result. p -values were obtained through statistical comparisons between the results obtained using the automatically selected hybrid paradigm design and pure MI for classification. All results were generated using a Wilcoxon rank-sum test unless highlighted with an asterisk, in which case a t-test was used. p -values in bold denote statistical significance.

F1 Scores													
Subject	Alpha – Beta Bands				Standard Bands				10 Hz Bands				
	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)	
1	28.64	37.08	34.80	35.61	33.59	31.23	42.75	39.65	27.25	32.21	39.93	44.92	
2	24.22	42.64	40.24	42.76	29.27	41.26	46.67	47.90	28.52	41.41	40.52	45.43	
3	29.88	32.41	38.75	38.39	31.34	35.19	39.80	34.69	28.45	29.04	35.86	41.15	
4	23.37	29.82	37.47	34.00	21.81	31.45	42.36	46.55	23.39	28.63	46.47	48.39	
5	33.18	29.68	37.01	35.46	25.55	28.51	30.70	32.13	27.55	27.69	36.31	34.67	
Average	27.86	34.33	37.65	37.25	28.31	33.53	40.45	40.18	27.03	31.80	39.82	42.91	
(p -value)			(5.6×10^{-2})	$(5.4 \times 10^{-2}^*)$			(5.1×10^{-3})	$(1.4 \times 10^{-4}^*)$			(6.8×10^{-6})	(2.8×10^{-7})	

obtained by comparing the results for the selected hybrid paradigm designs to those obtained using pure MI classification. Pure MI was used since it was the best-performing benchmark approach. The methodology of this statistical analysis is explained in Section 2.4.5. This analysis was carried out for the results obtained using Alpha-Beta Band features, Standard Band features, and 10 Hz Band features. p -values that indicate statistical significance are in bold. Table 3 shows results for the additional performance statistics, namely balanced accuracy, sensitivity, and PPV averaged across subjects. The table also contains p -values comparing the results obtained using the selected hybrid paradigm designs to the results obtained with pure MI for all three of the feature types.

The best performance in terms of F1 score was obtained for MI classification when using the Alpha-Beta bands (34.33%), whereas, for the hybrid paradigm approach, the best performance was obtained when using the design selected through SH for 10 Hz Bands

features (42.91%). The results for these two scenarios were compared using statistical tests, with the p -values obtained being: 1.5×10^{-5} (for F1 score), 6.38×10^{-6} (for balanced accuracy), 9.1×10^{-7} (for sensitivity), and 1.0×10^{-3} (for PPV).

Figure 6 shows the distributions of the F1 score results for each approach, namely MI, SI, the hybrid paradigm with grid search, and the hybrid paradigm with SH. The box plots were created using all of the individual fold-wise results for all of the subjects, meaning that each box represents the distribution of 50 results.

Figure 7 compares the average computational times, in seconds, for the traditional grid search approach and the grid search with SH. A t-test was used to compare the computational times for each feature, when using the grid search method and when using SH. The p -value obtained in all three instances was 7.56×10^{-10} , which indicated that SH had a significant impact on computational time.

Table 3. Comparing various performance measures (%) for the hybrid paradigms selected using grid search (GS) and successive halving (SH), and the baseline MI and SI approaches. In each row, the peak benchmarking result is highlighted in bold as well as the peak hybrid paradigm result. p -values are shown in brackets and were obtained through statistical comparisons between the results obtained using the automatically selected hybrid paradigm design and pure MI for classification. All results were generated using a Wilcoxon rank-sum test unless highlighted with an asterisk, in which case a t-test was used. p -values in bold denote statistical significance.

Measure	Alpha – Beta Bands				Standard Bands				10 Hz Bands			
	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)	SI	MI	Hybrid Paradigm (GS)	Hybrid Paradigm (SH)
Balanced Accuracy	52.80	56.56	59.49	59.26	52.60	56.19	61.20	60.86	52.28	56.16	60.47	62.93
Sensitivity	28.00	36.20	41.38	40.07	29.18	34.90	44.86	43.72	26.65	31.96	45.20	46.75
PPV	29.77	33.98	38.14	37.28	28.62	33.69	38.90	40.44	29.85	33.61	37.63	42.92
			$(1.9 \times 10^{-2}^*)$	(1.4×10^{-2})			(9.5×10^{-4})	$(1.1 \times 10^{-4}^*)$			$(4.9 \times 10^{-4}^*)$	$(6.5 \times 10^{-7}^*)$
			(5.2×10^{-3})	(2.3×10^{-2})			(8.7×10^{-6})	(5.9×10^{-5})			(1.8×10^{-9})	(6.0×10^{-9})
			$(8.8 \times 10^{-2}^*)$	(5.1×10^{-2})			$(2.8 \times 10^{-2}^*)$	$(2.8 \times 10^{-3}^*)$			(3.5×10^{-2})	(5.0×10^{-4})

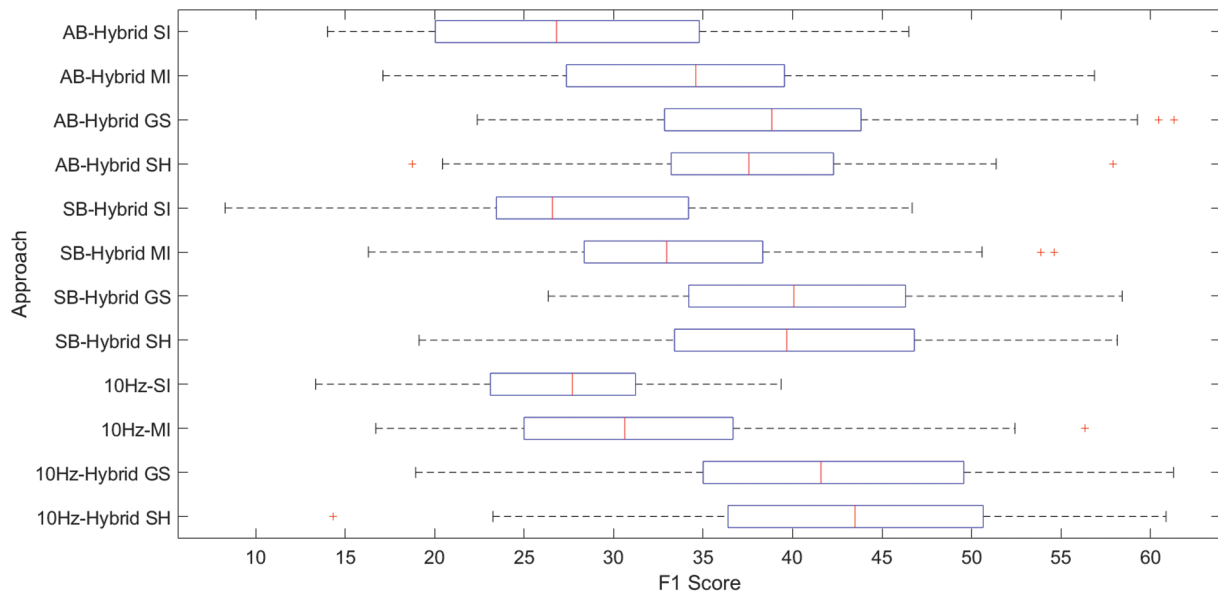


Figure 6. Box plots showing the distributions of the fold-wise F1 scores for the MI, SI, hybrid with grid search (GS), and hybrid with successive halving (SH) approaches. Results for all three types of features, namely the Alpha-Beta (AB), Standard Bands (SB), and 10 Hz Bands (10 Hz) are shown. The horizontal red lines denote the median of the distribution.

4. Discussion

4.1. General evaluation of the hybrid paradigm

This section discusses the results obtained during the general evaluation analysis of the MI-SI hybrid paradigm. Thus, it contains a broad analysis of the hybrid paradigm and illustrates the necessity for automated,

subject-specific command selection, which is then discussed in Section 4.2.

Considering the classification results for the pure paradigms only, in Figure 3, the MI paradigm outperformed the SI paradigm for all three feature extraction approaches. It is also evident that the pure MI paradigm performed best when only the alpha and beta frequency

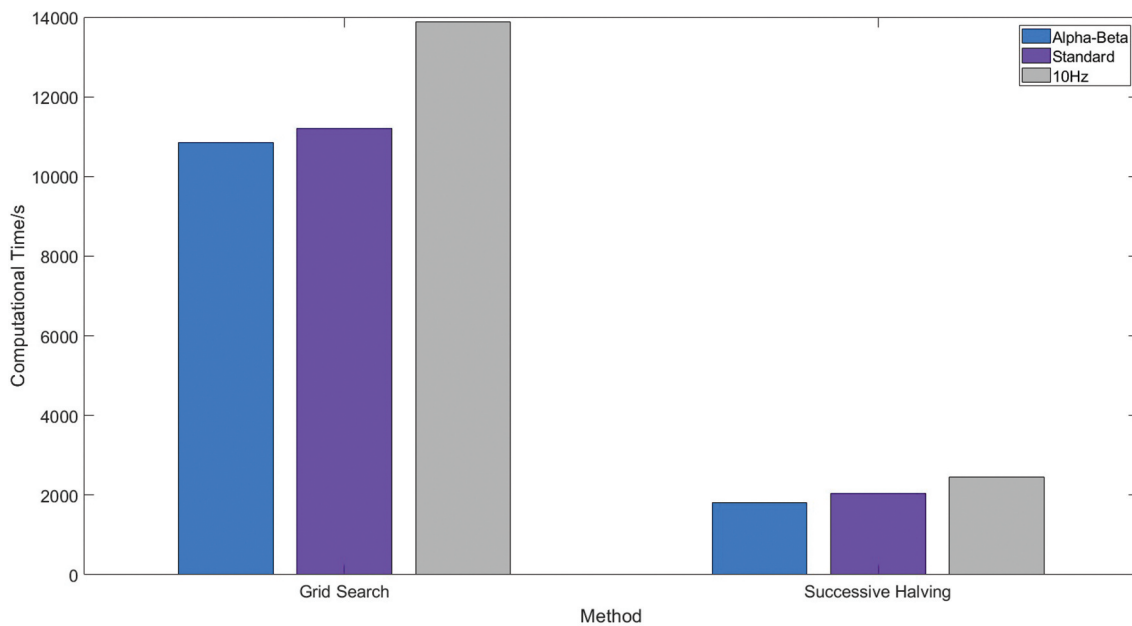


Figure 7. Comparing the average computational time in seconds for the hybrid paradigm design using the traditional grid search approach and successive halving.

bands were used. As discussed in Section 2.3 this was expected due to the relationship between MI and these particular bands.

It is evident from the plots in Figure 3 that the hybrid paradigm has a strong potential for outperforming the pure paradigms. When using the Standard Bands and 10 Hz Bands feature extraction approaches, all the hybrid paradigm combinations outperformed pure MI both in terms of F1 score and balanced accuracy.

Considering the results obtained when using the alpha and beta bands, the hybrid paradigms generally led to improved performance, however not as consistently. Furthermore, the hybrid results obtained using Alpha-Beta Band features tend to be clustered at lower values than those obtained using Standard Bands or 10 Hz Bands, with none of the results for Standard Bands or 10 Hz Bands falling below the baseline MI results. These results therefore indicate that frequencies outside the alpha and beta bands appear to be important for discrimination in the hybrid paradigms, and using wideband features has the potential to lead to notably improved performance.

Considering the results in Table 1, features extracted using Standard Bands gave the best F1 score, although the difference between the Standard Bands approach and the 10 Hz Bands approach was less than 1%. As expected from the plots in Figure 3, the Alpha-Beta Band features gave rise to the poorest-performing hybrid paradigm system. Based on this analysis, a hybrid paradigm classification system based on Standard Bands and the hybrid commands right-hand, tongue, 'right' and 'up', or 10 Hz Bands and the 'tongue', 'legs', 'up' and 'down' hybrid paradigm would be recommended. These resulted in improvements of 21.03% and 19.84%, respectively, in F1 score when compared to the best performing MI baseline F1 score of 34.33%, which was obtained using Alpha-Beta Band features.

It should be noted that some of the hybrid paradigm command sets are more intuitive than others. For example, in Table 1 the best performing hybrid paradigm consisted of the right-hand, tongue, 'right', and 'up' commands with Standard Bands features. This paradigm contains two commands conceptually associated with the term 'right'. Consider applying these commands to a practical BCI for cursor control which requires four commands: move left, move right, move up, and move down – whilst one of the two commands conceptually associated with the right can be associated with the 'move right' command, the other command will have to be associated with one of the other commands, which may be counter-intuitive. Of course, the intuition or lack thereof of these commands will depend on the application of the BCI. Furthermore, it should be

noted that in previous BCI work in the domain of paradigm selection, and also in the wider field of endogenous BCIs, the intuition of the paradigm is secondary to obtaining the optimal levels of control in the BCI through discriminative commands [1,5,6,47].

For this reason, Figure 4 shows the results obtained for two hybrid paradigm designs that were deemed to be more intuitive, and compares them to the MI and SI baseline performance. Considering the F1 scores, both of these designs outperformed the best-performing pure paradigm result, which was obtained for MI with the Alpha-Beta Band features. For both of the intuitive designs, the 10 Hz Bands features gave rise to peak performance. The F1 score for the right-hand, left-hand, 'up' and 'down' design (I1) outperformed the baseline MI approach by 9.45%, while the legs, tongue, 'left' and 'right' design outperformed the baseline MI approach by 14.80%. Improved performance was also observed in terms of balanced accuracy. Thus, although these more intuitive paradigms did not exhibit peak performance, they still outperformed the best baseline result.

Discussions up until this point have been focused on the average performance across subjects. Although these results give a generalized overview of the performance of the hybrid paradigm, they can obscure important inter-subject differences. Figure 5 illustrates these differences by comparing the classification performance for two particular subjects, specifically Subject 1 and Subject 2. Various hybrid paradigm designs would be suitable for Subject 1. In fact, in the case of the 10 Hz Bands approaches, all the hybrid paradigm designs led to improved performance when compared to the baseline MI and SI approaches. For the Standard Bands approach the vast majority of hybrid paradigms led to improved performance. As expected from the observations made in Figure 3, fewer hybrid paradigm designs led to improved performance when using Alpha-Beta Band features. Starkly different trends emerge for Subject 2. In this case, the majority of hybrid paradigm designs led to decreased performance when using features extracted using Standard, 10 Hz, and the Alpha-Beta Band features.

The paradigm format that results in peak performance also varies between subjects. Considering the results for 10 Hz Bands feature extraction, Subject 1 obtained a peak F1 score of 47.11% with the paradigm format {right-hand; legs; 'right'; 'up'}, while Subject 2 obtained a peak performance of 45.90% with the format {right-hand; tongue; 'left'; 'up'}.

Overall, these results illustrate that inter-subject differences in EEG data can lead to stark differences in performance for a hybrid paradigm. They also indicate that

while a hybrid paradigm can improve performance relative to the pure MI and SI benchmarks, this is not the case for all subjects. Furthermore, the paradigm formats that result in peak performance can vary between subjects. Thus, an automated subject-specific hybrid paradigm design selection approach would be beneficial. The next section discusses a generalizable hybrid paradigm design selection approach that can be used to select the best paradigm design for a particular subject.

4.2. Automated hybrid paradigm selection

This section discusses the results obtained using automated SH hybrid paradigm selection, in which subject-specific hybrid paradigm commands were selected for each subject. First, the classification performance obtained using automated hybrid paradigm selection is discussed, and then the computational efficiency of the automated selection approach is analyzed.

Considering the average results in the bottom row of [Table 2](#), for each feature type, the hybrid paradigm results led to an improved F1 score when compared to pure MI and pure SI. Comparing average performance across the features, a peak F1 score of 42.91% was obtained with the hybrid paradigm selected through SH for 10 Hz Bands features. This peak value was a 22.90% improvement on the best baseline result of 34.33%, which was obtained with pure MI and Alpha-Beta Band features. A peak performance of 28.31% was obtained for the SI paradigm with the Standard Bands features, meaning that the peak hybrid paradigm result was an improvement of 51.57% on this. Consider also the results for individual subjects. Using a hybrid paradigm led to improved peak performance when compared to the baseline approaches across subjects and features.

[Tables 3](#) shows the results for balanced accuracy, sensitivity, and PPV, averaged across subjects. Considering the average results for each feature, using a hybrid paradigm led to increased performance in all three statistics. As was the case for the F1 score results, peak performance was obtained for the hybrid paradigm selected through SH for 10 Hz Bands features. These peak values represented an improvement of 11.26% in balanced accuracy, 29.14% in sensitivity, and 26.31% in PPV when compared to the peak results obtained with MI (those for the Alpha-Beta Bands features). p -values recorded in [Section 3.2](#) also confirmed that these improvements were all statistically significant. The improvement in both sensitivity and PPV is particularly important: there is a risk that increased sensitivity in a classifier can lead to oversensitivity, which would

result in a corresponding drop in PPV. However, this is not the case in these results; in fact, the PPV increased, indicating that not only was the hybrid classifier more sensitive to command classes when compared to the pure MI paradigm, but it also had improved precision.

Comparing the average F1 score obtained by SH in [Table 2](#), and the average balanced accuracy results in to the corresponding baseline MI and SI results, the improvements in average F1 score and balanced accuracy respectively can be summarized as:

- 8.51% and 4.77% compared to MI, and 33.70% and 12.23% compared to SI, for Alpha-Beta Band features.
- 19.83% and 8.31% compared to MI, and 41.93% and 15.70% compared to SI, for Standard Band features.
- 34.94% and 12.05% compared to MI, and 34.94% and 12.05% compared to SI, for 10 Hz Bands features.

When averaging these observations across the different features, the average improvements in F1 score and accuracy were 21.09% and 8.38%, respectively, when compared to pure MI, and 36.86% and 13.33% when compared to pure SI.

The p -values in [Tables 2 and 3](#) confirm that, for the Standard Bands and 10 Hz Bands approaches, hybridization of the paradigm leads to significantly improved performance across all measures. In the case of the Alpha-Beta Bands approach, statistically significant changes in performance were observed in balanced accuracy and sensitivity.

The classification results discussed so far in this section are in agreement with the observations made in the general analysis in [Section 4.1](#), which indicated that at a population level, the Standard Bands and 10 Hz Bands had a strong performance with the MI-SI hybrid paradigm. However, the results discussed in this section go beyond those presented in the general evaluation by illustrating that the subject-specific commands selected through the automated SH approach lead to improved classification performance. Recall that the general evaluation highlighted the risk that for some subjects, the majority of hybrid paradigm command combinations can lead to diminished performance. Despite this risk, the SH approach selects commands that lead to improved performance for all subjects when using Standard Bands and 10 Hz Bands features, as shown in [Table 2](#).

The results obtained using the grid search and SH approaches can also be directly compared. Consider the

average F1 score results obtained using grid search and SH in Table 2. For each of the three feature extraction results (Alpha-Beta, Standard, and 10 Hz Bands), the results obtained using the traditional grid search were 37.65%, 40.45%, and 39.82%, which are similar to those obtained using SH: 37.25%, 40.18%, and 42.91%. The box plots in Figure 6 further confirm this similarity in results by illustrating how the distribution of results varies for both methods. Comparing the results for the hybrid paradigm with grid search and the hybrid paradigm with SH for each feature individually, there is a notable overlap in the distributions for each case. Considering the balanced accuracy results in Table 3, the results for the traditional grid search method were 59.49%, 61.20%, and 60.47%, which were similar to those obtained with SH: 59.26%, 60.86%, and 62.93%. From these results, it is notable that SH can sometimes lead to a notable increase in performance when 10 Hz features are used. This could be because SH may produce more generalizable results since it uses a subset of the training data for all of the iterations except the final one, meaning that it would be less likely to overfit the training data. Successive halving is considered to be an early-stopping method since it disqualifies the majority of the candidates after evaluating them on only a subset of the data [25,26]. Early-stopping methods by nature can prevent overfitting [26].

Considering the average computational time results in Figure 7, there are substantial differences between the traditional grid search approach and the SH approach. The figure presents the times in seconds because it is the International System of Units time measurement, however for better conceptualization of the latency the discussion will also mention hours and minutes. The grid search method always took longer than three hours to select the hybrid paradigm design regardless of the feature type used, whereas the SH approach always took under 45 minutes. Using the successive halving approach led to a reduction in average computational times of 83.34%, 81.75%, and 82.32% for the Alpha-Beta, Standard, and 10 Hz Bands features, respectively.

Computational times are important because, in a practical setup, the hybrid paradigm selection process introduces latency. In a practical system, the prospective user of the BCI would need to wait for the system to complete the selection process before they can use the BCI. Thus, shorter latencies use fewer computer resources and result in lower delays for the end-user. SH resulted in a statistically significant improvement in computation times.

4.2.1. Comments on performance variation with PSD feature type

This study focused on the use of PSD-based features and investigated the hybrid MI and SI paradigm in the context of features from three different PSD bands. Based on the classification performance results, the use of 10 Hz Bands in conjunction with a hybrid paradigm design approach based on SH would be recommended. This investigation into PSD-based features adds a supplementary element of novelty to this research since previous studies into hybrid MI and SI paradigms have not used PSD features, having favored other common features, namely: common spatial patterns, cross-correlation function, and phase locking value [20–22].

Comparing computational results between the different kinds of features, the Alpha-Beta Band features resulted in the lowest computational times, whereas the 10 Hz Bands approach had the highest computational times. This is likely because the feature vector used with the 10 Hz Bands was the largest, and that for the Alpha-Beta Bands approach was the smallest. Larger feature vectors contain more data for the classifier to process during training.

4.3. Limitations of this study and future work

The results indicate that a hybrid MI-SI paradigm can lead to significant improvement when compared to using the pure MI and SI paradigms, and should be considered for practical BCI implementations. However, the analysis has a core limitation: the relatively small sample size of five subjects. Although BCI-related studies with populations of this size are present in the recent literature [7,48–50], indicating that these results are still impactful and highlight a clear avenue for potential future studies, the statistical limitations of small sample sizes should be considered when interpreting the results. Future research that directly builds on this work should involve a larger number of participants to further validate the results presented in this study.

Another limitation is the fact that the experiments were conducted using just one kind of signal processing pipeline, consisting of power spectral density features and an SVM classifier, which have both been widely used for EEG classification [1,8,51]. This particular kind of classification pipeline was chosen due to its established use in the literature for the classification of both MI [52,53] and SI [8,31] data. Moreover, straightforward classification pipelines that do not involve any channel or feature selection have also been adopted in online BCI

implementations [54–58], possibly because these approaches are established and can be rapidly employed. Thus, the research presented in this study would be most relevant to these implementations. Notwithstanding this, feature and channel selection [9–14], as well as deep learning classifiers [1,6], are growing in importance in the literature, and it would be important for future work to evaluate the impact of SH command selection in pipelines involving these methods.

In this study, we constrained the hybrid paradigm to consist of two MI commands and two SI commands. The SH approach, however, is not restricted to this hybrid paradigm setup. It can be easily deployed in the same way shown in this paper to select the optimal commands for other hybrid paradigm command setups, such as one using three MI commands and one SI command, or three SI commands and one MI command. If one wishes to find the optimal hybrid paradigm design when considering all possible scenarios for the four commands (i.e. all possible combinations of two MI/two SI commands, three MI/one SI command, and three SI/one MI command), this would constitute a larger search which would take a longer time, but could lead to an improved result, and could be investigated as part of future work.

The impact of training the BCI user was also not covered in this study, which used data that was recorded from the subjects during a single recording session. It is a well-established phenomenon that users of both MI and SI-based BCIs can exhibit improved performance when they are trained over multiple sessions in using a BCI [59]. This training typically involves the user executing mental commands and being given feedback on how well the commands were classified [19,59]. In the future, it would be of value to study how the performance of users of a hybrid MI-SI BCI varies with user training and compare this to results obtained after training with the pure MI and SI BCIs. Training the user over multiple sessions is, however, time and resource-intensive. The results in this paper indicate that naïve (untrained) users perform better with the hybrid paradigm when compared to the pure paradigms, and this is a notable finding for studies that cannot afford extended training sessions, or aim to design plug-and-play BCIs.

Regarding future work, applying the findings of this study to an online BCI would be a direct next step. The commands would be selected offline by applying the proposed SH command selection method to the training data of the participant. Then, the selected commands could be used to control an online BCI to operate an

external device, such as a GUI. The performance of the online system with the selected commands could be compared to the performance with pure SI and MI paradigms.

Future work could also increase the variety of commands considered by the SH selection algorithm. In this paper, the search was limited to MI and SI commands, however, commands from other paradigms such as spelling and visual imagery, could be included in the search for the optimal command set. This future work could investigate whether introducing a wider variety of commands to the search leads to improved performance. Alternatively, the SH command selection algorithm could be applied to select the best commands within a particular pure paradigm. For example, it would be of interest to use SH to select the best commands in an SI system from pairs of synonyms. For instance, in an SI-based BCI for robot control, SH could be used to select the best commands from the synonym command pairs (up/increase), (down/decrease) and (stop/brake) to control the speed.

Finally, future work could build on this study by exploring alternative early-stopping methods that could be applied to the command selection problem. These could be compared to SH, both in terms of computational times to converge to the final command set, and in terms of the classification performance obtained with the selected commands. Algorithms suitable for such a study include Bayesian optimization with early stopping [60] and greedy early stopping [61].

5. Conclusion

The goal of this study was to improve the classification performance of endogenous BCI commands. This was achieved by first exhaustively investigating the impact of a hybrid MI-SI paradigm on multiclass classification performance when compared to the pure MI and SI paradigms, and then presenting a computationally effective SH-based command selection approach that produced command sets that were able to outperform the pure paradigms. This analysis was based on classification results obtained with a pipeline that used PSD features and an SVM classifier.

A general evaluation of the performance of the hybrid MI-SI command paradigm was carried out. In this analysis, the classification results for the pure four-class MI and SI command sets were used for benchmarking. The classification results for every possible combination (design) of two MI and two SI commands were compared to these benchmarks. It was evident that hybridization could lead to a notable improvement in performance. However,

there were also substantial inter-subject variations in paradigm design performance, as well as the risk that hybridization could lead to decreased performance for certain subjects, something previous literature had not indicated. This highlighted the necessity of an automated approach for selecting the best commands for individual subjects.

An SH-based approach for automated command selection was then presented. Using hybrid paradigm designs selected using this approach led to an average improvement of 21.09% and 36.86% in F1 score when compared to baseline MI and SI results, respectively. The SH approach was also compared to the traditional grid search approach. In terms of the performance of the selected designs, the results were similar. However, SH was found to be highly effective in accelerating the hybrid paradigm selection process, resulting in an 82.80% improvement in computational time.

The two main limitations of this study were the relatively small sample size used and the fact that experiments were only carried out on a single, conventional classification pipeline. Additional limitations include the fact that the impact of user training on classification performance in the proposed system was not assessed and that the system was limited to the selection of two MI and two SI commands.

Future work should focus on testing with a larger sample size, and with a more diverse range of classification pipelines. More in-depth investigations into the characteristic differences between MI and SI signals could also be carried out to improve classification with this hybrid paradigm. Moreover, testing with an online system would further validate the results. The SH command selection approach could also be applied to different hybrid or pure paradigms. Furthermore, other early-stopping techniques from the literature could be applied to the command selection problem.

Acknowledgements

This work was supported by the European Regional Development Fund 2014–2020 under Grant ERDF.01.124. The authors would like to acknowledge the project: “Setting up of transdisciplinary research and knowledge exchange (TRAKE) complex at the University of Malta (ERDF.01.124)”, which is being co-financed through the European Union through the European Regional Development Fund 2014–2020.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the European Regional Development Fund [ERDF.01.124].

ORCID

Natasha Padfield  <http://orcid.org/0000-0001-5533-4807>

Data availability statement

The EEG dataset used in this study is open-access, available at: <https://doi.org/10.60809/drum.24465871.v1> [62].

References

- [1] Padfield N, Camilleri K, Camilleri T, et al. A comprehensive review of endogenous EEG-Based BCIs for dynamic device control. *Sensors (Basel)*. [2022 Aug 1];22(15) NLM (Medline):5802. doi: 10.3390/s22155802
- [2] Baek HJ, Chang MH, Heo J, et al. Enhancing the usability of brain-computer interface systems. *Comput Intell Neurosci*. 2019;2019:1–12. doi: 10.1155/2019/5427154
- [3] Azadi Moghadam M, Maleki A. Fatigue factors and fatigue indices in SSVEP-based brain-computer interfaces: a systematic review and meta-analysis. *Front Hum Neurosci*. 2023 Nov;17:1248474. doi: 10.3389/fnhum.2023.1248474
- [4] Del Millán JR, Mouriño J. Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project. *IEEE Trans Neural Syst Rehabil Eng*. 2003 Jun;11(2):159–161. doi: 10.1109/TNSRE.2003.814435
- [5] Friedrich EVC, Neuper C, Scherer R. Whatever works: a systematic user-centered training protocol to optimize brain-computer interfacing individually. *PLoS One*. 2013;8(9):19–22. doi: 10.1371/journal.pone.0076214
- [6] Tang X, Li W, Li X, et al. Motor imagery EEG recognition based on conditional optimization empirical mode decomposition and multi-scale convolutional neural network. *Expert Syst Appl*. 2020 Jul;149:113285. doi: 10.1016/j.eswa.2020.113285
- [7] Kuzovkin I, Tretyakov K, Uusberg A, et al. Mental state space visualization for interactive modeling of personalized BCI control strategies. *J Neural Eng*. 2020;17(1):016059. doi: 10.1088/1741-2552/ab6d0b
- [8] Panachakel JT, Ramakrishnan AG. Decoding covert speech from EEG-A comprehensive review. *Front Neurosci*. [2021 Apr];15. doi: 10.3389/fnins.2021.642251
- [9] Martínez-Cagigal V, Santamaría-Vázquez E, Hornero R. Brain-computer interface channel selection optimization using meta-heuristics and evolutionary algorithms. *Appl Soft Comput*. 2022 Jan;115:108176. doi: 10.1016/j.asoc.2021.108176
- [10] Idowu OP, Adelopo O, Ilesanmi AE, et al. Neuro-evolutionary approach for optimal selection of EEG channels in motor imagery based BCI application.

- Biomed Signal Process Control. 2021;68 (January):102621. doi: [10.1016/j.bspc.2021.102621](https://doi.org/10.1016/j.bspc.2021.102621)
- [11] Padfield N, Ren J, Murray P, et al. Sparse learning of band power features with genetic channel selection for effective classification of EEG signals. *Neurocomputing*. 2021;463:566–579. doi: [10.1016/j.neucom.2021.08.067](https://doi.org/10.1016/j.neucom.2021.08.067)
- [12] Aljalal M, Molinas M, Aldosari SA, et al. Mild cognitive impairment detection with optimally selected EEG channels based on variational mode decomposition and supervised machine learning. *Biomed Signal Process Control*. 2024 Jan;87:105462. doi: [10.1016/j.bspc.2023.105462](https://doi.org/10.1016/j.bspc.2023.105462)
- [13] Maniruzzaman M, Hasan MAM, Asai N, et al. Optimal channels and features selection based ADHD detection from EEG signal using statistical and machine learning techniques. *IEEE Access*. 2023 Apr;11:33570–33583. doi: [10.1109/ACCESS.2023.3264266](https://doi.org/10.1109/ACCESS.2023.3264266)
- [14] Kouka N, Fourati R, Fdhila R, et al. EEG channel selection-based binary particle swarm optimization with recurrent convolutional autoencoder for emotion recognition. *Biomed Signal Process Control*. 2023 Jul;84:104783. doi: [10.1016/j.bspc.2023.104783](https://doi.org/10.1016/j.bspc.2023.104783)
- [15] Taheri Gorji H, Wilson N, VanBree J, et al. Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Sci Rep*. 2023 Feb;13(1):2507. doi: [10.1038/s41598-023-29647-0](https://doi.org/10.1038/s41598-023-29647-0)
- [16] Farokhah L, Sarno R, Fatichah C. Simplified 2D CNN architecture with channel selection for emotion recognition using EEG spectrogram. In: *IEEE Access*; 2023 May. doi: [10.1109/ACCESS.2023.3275565](https://doi.org/10.1109/ACCESS.2023.3275565)
- [17] Abdi-Sargezeh B, Foodeh R, Shalchyan V, et al. EEG artifact rejection by extracting spatial and spatio-spectral common components. *J Neurosci Methods*. 2021 Jul;358:109182. doi: [10.1016/j.jneumeth.2021.109182](https://doi.org/10.1016/j.jneumeth.2021.109182)
- [18] Fu Y, Li Z, Gong A, et al. Identification of visual imagery by electroencephalography based on empirical mode decomposition and an autoregressive model. *Comput Intell Neurosci*. 2022 Jan;2022:1–10. doi: [10.1155/2022/1038901](https://doi.org/10.1155/2022/1038901)
- [19] Tonin L, Bauer FC, Del Millán JR. The role of the control framework for continuous teleoperation of a brain-machine interface-driven mobile robot. *IEEE Trans Robot*. 2019;36(1):78–91. doi: [10.1109/TRO.2019.2943072](https://doi.org/10.1109/TRO.2019.2943072)
- [20] Wang L, Zhang X, Zhang Y. Extending motor imagery by speech imagery for brain-computer interface. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; 2013 Jul. p. 7056–7059. doi: [10.1109/EMBC.2013.6611183](https://doi.org/10.1109/EMBC.2013.6611183)
- [21] Tong J, Xing Z, Wei X, et al. Towards improving motor imagery brain-computer interface using multimodal speech imagery. *J Med Biol Eng*. 2023 Jun;43 (3):216–226. doi: [10.1007/s40846-023-00798-9](https://doi.org/10.1007/s40846-023-00798-9)
- [22] Wang L, Liu X, Liang Z, et al. Analysis and classification of hybrid BCI based on motor imagery and speech imagery. *Measurement*. 2019 Dec;147:106842. doi: [10.1016/j.measurement.2019.07.070](https://doi.org/10.1016/j.measurement.2019.07.070)
- [23] Yildirim E, Kaya Y, Kilic F. A channel selection method for emotion recognition from EEG based on swarm-intelligence algorithms. *IEEE Access*. 2021;9:109889–109902. doi: [10.1109/ACCESS.2021.3100638](https://doi.org/10.1109/ACCESS.2021.3100638)
- [24] Baig MZ, Aslam N, Shum HPH, et al. Differential evolution algorithm as a tool for optimal feature subset selection in motor imagery EEG. *Expert Syst Appl*. 2017;90:184–195. doi: [10.1016/j.eswa.2017.07.033](https://doi.org/10.1016/j.eswa.2017.07.033)
- [25] Li L, Jamieson K, Rostamizadeh A, et al. Massively parallel hyperparameter tuning. In: *NeurIPS Workshop on Machine Learning Systems*; Montreal. 2018.
- [26] Soper DS. Hyperparameter optimization using successive halving with greedy cross validation. *Algorithms*. 2022 Dec;16(1):17. doi: [10.3390/a16010017](https://doi.org/10.3390/a16010017)
- [27] Goay CH, Ahmad NS, Goh P. Transient simulations of high-speed channels using CNN-LSTM with an adaptive successive halving algorithm for automated hyperparameter optimizations. *IEEE Access*. 2021;9:127644–127663. doi: [10.1109/ACCESS.2021.3112134](https://doi.org/10.1109/ACCESS.2021.3112134)
- [28] BioSemi. ActiveTwo. BioSemi Website; [cited 2024 Feb 2]. Available from: <https://www.biosemi.com/products.htm>
- [29] Kaya M, Binli MK, Ozbay E, et al. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Sci Data*. 2018 Oct;5(1):180211. doi: [10.1038/sdata.2018.211](https://doi.org/10.1038/sdata.2018.211)
- [30] Romero S, Mañanas MA, Barbanoj MJ. Ocular reduction in EEG signals based on adaptive filtering, regression and blind source separation. *Ann Biomed Eng*. 2009 Jan;37(1):176–191. doi: [10.1007/s10439-008-9589-6](https://doi.org/10.1007/s10439-008-9589-6)
- [31] Koizumi K, Ueda K, Nakao M. Development of a cognitive brain-machine interface based on a visual imagery method. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2018 Jul. p. 1062–1065. doi: [10.1109/EMBC.2018.8512520](https://doi.org/10.1109/EMBC.2018.8512520)
- [32] Oikonomou VP, Georgiadis K, Liaros G, et al. A comparison study on EEG signal processing techniques using motor imagery EEG data. *Proc IEEE Symp Comput Based Med Syst*. 2017;2017-June(1):781–786. doi: [10.1109/CBMS.2017.113](https://doi.org/10.1109/CBMS.2017.113)
- [33] Kim C, Sun J, Liu D, et al. An effective feature extraction method by power spectral density of EEG signal for 2-class motor imagery-based BCI. *Med Biol Eng Comput*. 2018 Sep;56(9):1645–1658. doi: [10.1007/s11517-017-1761-4](https://doi.org/10.1007/s11517-017-1761-4)
- [34] Fleck JI, Green D, Stevenson J, et al. The transliminal brain at rest: baseline EEG, unusual experiences, and access to unconscious mental activity. *Cortex*. 2008;44 (10):1353–1363. doi: [10.1016/j.cortex.2007.08.024](https://doi.org/10.1016/j.cortex.2007.08.024)
- [35] Stinson B, Arthur D. A novel EEG for alpha brain state training, neurobiofeedback and behavior change. *Complement Ther Clin Pract*. 2013 Aug;19 (3):114–118. doi: [10.1016/j.ctcp.2013.03.003](https://doi.org/10.1016/j.ctcp.2013.03.003)
- [36] MathWorks. fitcecoc. MATLAB documentation. [cited 2023 Jul 12]. Available from: <https://www.mathworks.com/help/stats/fitcecoc.html#bufm0tv>
- [37] Cooney C, Korik A, Folli R, et al. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. *Sensors*. 2020 Aug;20(16):4629. doi: [10.3390/s20164629](https://doi.org/10.3390/s20164629)

- [38] Autthasan P, Du X, Arnin J, et al. A single-channel consumer-grade EEG device for brain-computer interface: enhancing detection of SSVEP and its amplitude modulation. *IEEE Sensors J.* 2020 Mar;20(6):3366–3378. doi: [10.1109/JSEN.2019.2958210](https://doi.org/10.1109/JSEN.2019.2958210)
- [39] Liu X, Lv L, Shen Y, et al. Multiscale space-time-frequency feature-guided multitask learning CNN for motor imagery EEG classification. *J Neural Eng.* 2021 Apr;18(2):026003. doi: [10.1088/1741-2552/abd82b](https://doi.org/10.1088/1741-2552/abd82b)
- [40] O'Reilly C, Nielsen T. Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools. *Front Hum Neurosci.* 2015;9(JUNE):1–19. doi: [10.3389/fnhum.2015.00353](https://doi.org/10.3389/fnhum.2015.00353)
- [41] Momenzadeh A. EEG-based emotion recognition utilizing wavelet coefficients. *Multimed Tools Appl.* 2018 Oct;77(20):27089–27106. doi: [10.1007/s11042-018-5906-8](https://doi.org/10.1007/s11042-018-5906-8)
- [42] Parekh V, Subramanian R, Roy D, et al. An EEG-Based image annotation system. 2018. p. 303–313. doi: [10.1007/978-981-13-0020-2_27](https://doi.org/10.1007/978-981-13-0020-2_27)
- [43] Peh WY, Yao Y, Dauwels J. Transformer convolutional neural networks for automated artifact detection in scalp EEG. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE; 2022 Jul. p. 3599–3602. doi: [10.1109/EMBC48229.2022.9871916](https://doi.org/10.1109/EMBC48229.2022.9871916)
- [44] Kumaravel V, Paissan F, Farella E. Towards a domain-specific neural network approach for EEG bad channel detection. In: *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE; 2021 Dec. p. 1–4. doi: [10.1109/SPMB52430.2021.9672305](https://doi.org/10.1109/SPMB52430.2021.9672305)
- [45] Baumgartl H, Bayerlein S, Buettner R. Measuring extraversion using EEG data. 2020. p. 259–265. doi: [10.1007/978-3-030-60073-0_30](https://doi.org/10.1007/978-3-030-60073-0_30)
- [46] Rahman Z, Ami AM. A transfer learning based approach for skin lesion classification from imbalanced data. In: *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE; 2020 Dec. p. 65–68. doi: [10.1109/ICECE51571.2020.9393155](https://doi.org/10.1109/ICECE51571.2020.9393155)
- [47] Kucukyildiz G, Ocaik H, Karakaya S, et al. Design and implementation of a multi sensor based brain computer interface for a robotic wheelchair. *J Intell Robot Syst.* 2017;87(2):247–263. doi: [10.1007/s10846-017-0477-x](https://doi.org/10.1007/s10846-017-0477-x)
- [48] Pawar D, Dhage S. Multiclass covert speech classification using extreme learning machine. *Biomed Eng Lett.* 2020 Mar;10(2):217–226. doi: [10.1007/s13534-020-00152-x](https://doi.org/10.1007/s13534-020-00152-x)
- [49] Liu Z, Wang L, Xu S, et al. A multiwavelet-based sparse time-varying autoregressive modeling for motor imagery EEG classification. *Comput Biol Med.* 2023 Mar;155:106196. doi: [10.1016/j.compbiomed.2022.106196](https://doi.org/10.1016/j.compbiomed.2022.106196)
- [50] Guo Y, Zhang Y, Chen Z, et al. EEG classification by filter band component regularized common spatial pattern for motor imagery. *Biomed Signal Process Control.* 2020 May;59:101917. doi: [10.1016/j.bspc.2020.101917](https://doi.org/10.1016/j.bspc.2020.101917)
- [51] Padfield N, Zabalza J, Zhao H, et al. EEG-based brain-computer interfaces using motor-imagery: techniques and challenges. *Sensors (Switz).* 2019;19(6):1–34. doi: [10.3390/s19061423](https://doi.org/10.3390/s19061423)
- [52] Oikonomou VP, Nikolopoulos S, Kompatsiaris I. Motor imagery classification via clustered-group sparse representation. In: *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*; 2019. p. 321–325. doi: [10.1109/BIBE.2019.00064](https://doi.org/10.1109/BIBE.2019.00064)
- [53] Pal M, Bandyopadhyay S. Many-objective feature selection for motor imagery EEG signals using differential evolution and support vector machine. In: *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*. IEEE; 2016 Jan. p. 1–6. doi: [10.1109/MicroCom.2016.7522574](https://doi.org/10.1109/MicroCom.2016.7522574)
- [54] Chen C, Zhou P, Belkacem AN, et al. Quadcopter robot control based on hybrid brain-computer interface system. *Sensors Mater.* 2020;32(3):991–1004. doi: [10.18494/SAM.2020.2517](https://doi.org/10.18494/SAM.2020.2517)
- [55] Wang C, Wu X, Wang Z, et al. Implementation of a brain-computer interface on a lower-limb exoskeleton. *IEEE Access.* 2018 Jul;6:38524–38534. doi: [10.1109/ACCESS.2018.2853628](https://doi.org/10.1109/ACCESS.2018.2853628)
- [56] Ehrlich SK, Cheng G. Human-agent co-adaptation using error-related potentials. *J Neural Eng.* 2018 Sep;15(6):066014. doi: [10.1088/1741-2552/aae069](https://doi.org/10.1088/1741-2552/aae069)
- [57] Gordleeva SY, Lobov SA, Grigorev NA, et al. Real-time EEG-EMG human-machine interface-based control system for a lower-limb exoskeleton. *IEEE Access.* 2020;8:84070–84081. doi: [10.1109/ACCESS.2020.2991812](https://doi.org/10.1109/ACCESS.2020.2991812)
- [58] Xu Y, Ding C, Shu X, et al. Shared control of a robotic arm using non-invasive brain-computer interface and computer vision guidance. *Rob Auton Syst.* 2019 May;115:121–129. doi: [10.1016/j.robot.2019.02.014](https://doi.org/10.1016/j.robot.2019.02.014)
- [59] Wang L, Huang W, Yang Z, et al. A method from offline analysis to online training for the brain-computer interface based on motor imagery and speech imagery. *Biomed Signal Process Control.* 2020 Sept;62:102100. doi: [10.1016/j.bspc.2020.102100](https://doi.org/10.1016/j.bspc.2020.102100)
- [60] Cho H, Kim Y, Lee E, et al. Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks. *IEEE Access.* 2020 Mar;8:52588–52608. doi: [10.1109/ACCESS.2020.2981072](https://doi.org/10.1109/ACCESS.2020.2981072)
- [61] Soper DS. Greed is good: rapid hyperparameter optimization and model selection using greedy k-fold cross validation. *Electronics.* 2021 Aug;10(16):1973. doi: [10.3390/electronics10161973](https://doi.org/10.3390/electronics10161973)
- [62] Padfield N, Camilleri K, Camilleri T, et al. “Motor and speech imagery EEG dataset,” drUM. [cited 2024 Jan 12]. Available from: https://drum.um.edu.mt/articles/dataset/Motor_and_Speech_Imagery_EEG_Dataset/24465871