Chapter 2

Preliminaries and Related Work: Literature Review

2.1. Introduction

This thesis is motivated by the diverse and rich literature works performed on the applications of different combinatorial techniques and machine learning methods in search space optimization. Overall, the procedure of recognizing research gaps in the field of robust BCI design and corresponding solutions is the fundamental cornerstone in modern optimization algorithms and serves as the basis for the successful implementation of machine learning models. Specifically, this thesis builds upon the prior research studies that seek the role of different information-theoretic paradigms and metaheuristics in the performance improvement of conventional BCI systems. We also draw strong conclusions from the baseline works in terms of their limitations and the scope of involvement of emerging concepts for dealing with the challenges in MI-specific EEG signals classification. In this chapter, we discuss the various concepts that are frequently employed in our experiments. A detailed description of all the concepts is given in the subsequent subsections.

2.2. Combinatorial Optimization for Search Space Reduction

Combinatorial optimization is a subfield of mathematical optimization that consists of finding an optimal object from a finite set of objects, where the set of feasible solutions is discrete or can be reduced to a discrete set. These procedures are effective when the solution search space is typically too large to search exhaustively using pure brute force. Compared to the conventional exact methods where the solution is always deterministic, combinatorial optimization seeks to improve an algorithm by using mathematical methods either to reduce the size of the set of possible solutions or to make the search itself faster [23]. Therefore, in principle, any sort of search algorithm or metaheuristic can be used to solve a given optimization problem that can recognize irrelevant solutions.

In optimization theory, combinatorial optimization algorithms are used to improve the global performance of the model by effectively correlating the candidate solutions with available instances. In this context, the following assumptions are made to determine optimal solutions for different issues discussed in Chapter 1:

I. The multichannel nature of EEG signals is mainly responsible for inducing an abundant amount of redundant and irrelevant information that increases the overall complexity and reduces the classification accuracy of BCI systems.

II. Maintaining a good trade-off between redundancy and relevancy is one of the major challenges to dealing with high-dimensional data. Therefore, a strong interactive mechanism is required to determine the proximity between priorly known candidate solutions and newly selected solutions.

III. Although the motor cortex is mainly responsible for stimulating MI signals other brain regions may also play a significant role in controlling muscular movement activities.

IV. Neural signals are always subject-specific, therefore, finding a global set of metrics to compare the results of different experiments is crucial. In this case, heterogeneous performance measures that are independent of each other may be used in performance comparison.

In search-space optimization, assumptions I and II deal with key problems: (1) Optimal channels subset selection, and (2) Significant feature selection. Mathematically, both are similar and therefore, their solutions are interchangeably used to solve both problems. Due to a large number of available optimization methods, it is difficult to determine the best method for a given optimal selection problem. A series of reviews that describe the various available optimization schemes in this context are available on the internet [24, 25]. In general, these approaches consist of four main stages: subset generation, subset evaluation, stopping criteria, and validation of results. In each iteration of the search process, a subset of the candidate feature set is generated from the original features, and its appropriateness is measured by an evaluation criterion. The subset generation process and its evaluation are repeated until a predetermined stop criterion is reached. At the end of this process, the best subset of the selected feature is validated on the test dataset. For better understanding, these methods can be categorized into the following three groups:

I. Filter methods
II. Wrapper methods
III. Hybrid methods

A detailed discussion of the aforementioned categories is given in subsequent subsections.

2.2.1. Filter-based Search Space Optimization Methods

The filter methods are popular optimization methods that utilize statistical data-dependent techniques to design optimal solutions. These methods are stable, classifier-independent, and relatively fast but avoid the relevance of dimensions while selecting channels [26]. In addition, these methods suffer from poor classification accuracy because they do not consider the role of the classifier in evaluating the quality of computed solutions. They utilize various information-theoretic concepts such as Mutual Information (MI), entropy, and Information Gain (IG). Also, they explore different distance and correlation measures to find the relationship between the dimensions. Recently, numerous filter methods have been proposed to obtain a good solution for both channel and feature selection problems. A generalized pseudocode of a filter-based feature selection algorithm is given in Algorithm 2.1. A sample dataset and filter-based feature selection Python code is given in Link 1 in the footer. In Table 2.1, we recapitulated the filter methods–--based on research studies published during the last 10 years.

2.2.2. Wrapper-based Search Space Optimization Methods

Wrapper-based techniques are another popular class of search-space optimization methods that rely on machine learning algorithms to find an optimal solution from a set of feasible solutions. These methods rely on the predictive ability of the applied classifier to estimate the quality of the selected solutions. These methods have better generalization ability and comprehensive search strategy than filter methods [27]. This approach provides better classification accuracy

---

[1] https://drive.google.com/drive/folders/1yDCoWxHbRx_8saQmVjeV-RD4HhyXliEE?usp=sharing

than the former class since it maximizes the interaction between a classifier and the data of selected channels, thereby making it computationally expensive, slow, and prone to overfitting. Wrapper methods use iterative search procedures that repeatedly compute responses to the model and then use the resulting model performance estimate to guide the selection of the next subset to evaluate. Compared to filter-based methods, wrapper methods intend to explore the dynamics of the search space so that they can evaluate the different aspects of the given problem instance. Therefore, these methods often rely on the application of different metaheuristic algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Firefly Algorithm (FA), and Harmony Search (HS) method in search space optimization related to channel and feature selection problems. A detailed study of the wrapper-based methods is summarized in Table 2.2.

2.2.3. Hybrid Search Space Optimization Methods

Hybrid techniques are the combination of the above two techniques and eliminate the pre-specification of a stopping criterion [27]. The hybrid techniques were developed to deal with large datasets. Algorithm 2.1 shows a typical hybrid technique algorithm that utilizes both an independent measure *M* and a mining algorithm *A* to evaluate the fitness of a subset. Also, these methods are efficient in computing dependencies between channels with lower computational costs than the wrapper methods since they don't compute optimal solutions iteratively. The algorithm starts its search from a given subset S_0 and tries to find the best subset in each round while also increasing the cardinality. The parameters/variables γ_{best} and θ_{best} correspond to cases with and without a classifier respectively and are calculated in each round. The quality of results from a mining algorithm offers a natural stopping criterion. However, it is complicated to develop hybrid optimization techniques because the amalgamation of incompatible components may reduce the global performance of the model.

Because of their effectiveness and advantages over filter and wrapper methods, numerous hybrid methods have been proposed to solve the search space optimization problem. A few prominent research techniques are listed in Table 2.3.

Algorithm 2.1.: **The Pseudocode of Generalized Filter-based Feature Selection Methods**

```
1. Input:
  - X: Feature matrix with n samples and m features
  - y: Target variable vector

2. Calculate the relevance score for each feature:
  - For each feature i in range(m):
    - Apply a relevance metric (e.g., correlation, mutual information) between X[:, i] and y
    - Assign the relevance score to the feature

3. Select the top-k features based on their relevance scores:
  - Sort the features based on their relevance scores in descending order
  - Select the top-k features with the highest scores

4. Output:
  - Selected features: The subset of features with the highest relevance scores
```

Table 2.1. A study on filter-based search-space optimization techniques

Methods	Dataset characteristics	Application	Comments
Mutual Information (MI), Minimum Redundancy and Maximum Relevance (MRMR), Laplacian Derivative, Spectral Features [28]	1. The EEG dataset was collected from 11 participants with 22 channels. 2. Three classes (movements of the legs, joints, and feelings) were used in the classification.	Channel Selection	The Minimum Redundancy and Maximum Relevance (MRMR) and the Mutual Information-based Channel Selection (MICS) methods efficiently computed the relevancy but were limited by excessive redundancy; thereby ignoring the relevance of channels individually.
Improved Sequential floating forward selection (I-SFFS) [29]	1. BCI Competition IV datasets 1 (59 channels), 7 participants, 3 classes (left hand, right hand, and foot.) 2. BCI Competition III datasets Iva (118 channels), 5 participants, 2 classes (right hand" and "foot")	Channel Selection	1. Based on the distribution of channels in the cerebral cortex, the adjacent channels would be treated as one measure for selection. 2. The proposed methodology was not superior in terms of classification accuracy compared to the original, SFFS method.
Pearson Correlation Coefficients, Regularized Common Spatial TPattern (RCSP) [30]	1. BCI Competition IV datasets 1 2. BCI Competition III datasets Iva 3. BCI Competition III (60 channels), 3 participants, 4 classes (left hand, right hand, tongue or foot)	Channel Selection	1. The key problem of this experiment is ignorance of the candidate channels. 2. In other words, it avoids the presence or effect of other significant channels outside of the two being explored in channel selection
Threshold-based Normalized Mutual Information (NMI) measure [31]	DEAP database, 32 channels, 32 participants, 2 classes (valence and arousal emotion)	Channel Selection	Although the selected channels significantly contributed to enhancing classification accuracy it ignored the relevance of the channels with respect to candidate solutions.
Non-negative Matrix Factorization (NMF), Mutual Information [32]	BrainNet BNT 36 (EMSA, Brazil) dataset, 19 channels, 2 classes (Rest and Pedal state)	Channel Selection	The performance of this procedure was limited because NMF is an algorithmically more complex method to implement, and its convergence can be slow

Granger Causality, common spatial pattern (CSP) and regularized CSP (RCSP) [33]	BCI2000 instrumentation system dataset, 64-channel EEG, 109 subjects, 2 classes (left and right hand MI)	Channel Selection	1. Granger causality lacked effective forecasting of interdependency between two or more channels. 2. This methodology does not provide any insight into the relationship between the variables hence it is not true causality, unlike 'Cause and Effect analysis.
Mutual Information-based Feature Selection (MIFS) [34]	Only statistical results are available	Feature Selection	1. Not suitable for high-dimensional datasets. 2. Mutual Information is sensitive to the scales of the variables. If the scales of different variables are not consistent, MI might not accurately capture their relationships.
Minimal-Redundancy Maximal-Relevance (MRMR) and MIFS [35]	6 gene expression data sets: NCI, Lymphoma, Lung, Child Leukemia, Leukemia, and Colon	Feature Selection	The MRMR and MIFS effectively compute the redundancy between the features but are limited by estimating too much redundancy
Feature Interaction Maximization (FIM)- Three-way interaction maximization approach [36]	Three multiclass UCI datasets: Gas sensor, Musk, Libra movement	Feature Selection	1. Superiority over Information Gain (IG), Minimal-Redundancy Maximal-Relevance (MRMR), and Information Gain-based Feature Selection (IGFS) methods. 2. Performance is limited by a lack of information about the interaction between the features and the classifier, and the selection of redundant and irrelevant features
Markov blanket approach [37]	Not known	Feature Selection	1. The proposed approach realized better classification performance than MRMR, conditional mutual information minimization, correlation-based feature selection, and variable neighborhood search method. 2. This approach lacks causality and, therefore, doesn't capture causal relationships. In addition, this approach is not suitable for noisy data.

Table 2.2. A study on wrapper search space optimization techniques

Methods	Dataset characteristics	Application	Comments
Genetic Algorithm (GA), Artificial Neural Network [38]	BCI competition III- ECoG (64 channels), 2 classes (tongue, finger movement)	Channel Selection	1. The higher-ranked (> Threshold score) channels were fed up into a multi-layer perceptron and categorized the neurons into four subspaces to generate if-else rules. 2. The proposed approach suffers from the "Averaging effect" because the GA depends on random number generators to select population, mutations, and crossover. 3. The experiment was validated on a small dataset. Therefore, the results may not be reliable.
Fuzzy Network, discrete wavelet transform (DWT), NSGA-II (Elitist Non-Dominated Sorting Genetic Algorithm), bi-objective function with two criteria (error rate and the number of channels) [39]	Private dataset, 14 channels, 21 participants, 5 classes of different Spanish words	Channel Selection	1. The method implemented a channel selection composed of two stages; the first one obtains a Pareto front and is approached as a multi-objective optimization problem dealing with the error rate and the number of channels; the second stage selects a single solution (channel combination) from the front, applying a fuzzy inference system (FIS). 2. The proposed methodology suffers from weak global optimization ability and poor convergence rate.
Genetic Algorithm (GA), Stockwell transform, and Bayesian linear discriminant analysis [40]	BCI Competition III dataset I, 64 ECoG Channels, 2 classes (left pinky finger, tongue)	Channel Selection	1. Performance on the EEG dataset is still unknown. 2. The proposed methodology suffers from weak global optimization ability and poor convergence rate.

Binary Harmony Search (BHS) algorithm, Common Spatial Pattern (CSP) [41]	1. Private dataset	Channel Selection	1. In this approach, the new harmony was improvised by the existing Harmony Memory Consideration Rule (HMCR) and pitch adjustment operator. 2. The test results reveal that BHS achieved better classification accuracy in less computation time than the steady-state genetic algorithms. 3. Effective parameter tuning is a time and resource-consuming procedure that makes the proposed algorithm slow.
Three variants of multiobjective Genetic Algorithms [42]	1. BCI competition III - (64 channels), 2 participants, 36 classes. 2. BCI Competition III datasets Iva (118 channels), 5 participants, 2 classes (right hand" and "foot") 3. BCI Competition Data Set IVc-(118 channels), single participant, 2 classes (left hand, right foot)	Channel Selection	1. Conventional Genetic Algorithm (GA), steady-state GA, and NSGA-II are adopted for the experiment. 2. Proper parameter tuning and obtaining an effective convergence rate in all three variants were noticeable challenges in this experiment.
Improved Binary Gravitation Search Algorithm (IBGSA), (Time, Frequency, and Wavelet) feature domain [43]	BCI competition IV-2a-(22 channels), 9 participants, 4 classes (left or right hand, either feet or tongue movements)	Channel Selection	1. The major problem of this algorithm was its implementation on only a single dataset. 2. GSA is good at finding the global optimum but has the drawbacks of slow convergence speed and getting stuck in local minima in the last iterations.

Cosine similarity measure support vector machines (CSMSVM) [44]	Time series data collected from the rolling element bearing test	Feature Selection	1. Cosine-similarity distance of individual features from the SVM marginal hyperplane was calculated and then ranked according to their closeness level. 2. Selecting an effective kernel, and relevant similarity measures were major issues. 3. Implementation only on a single dataset is a major issue of the work.
Ant Colony Optimization (ACO) with differential evolution algorithm [45]	1. BCI dataset- 56 channels- 2 classes (right or left finger) 2. Myoelectric Signals (MES)- 8 channels- 7 classes (hand open, hand close, wrist flexion, wrist extension, supination, pronation, and rest)	Feature Selection	1. The performance of the proposed methodology depended on several hyperparameters. 2. Applying a global tuning procedure is complicated for adjusting all the hyperparameters is highly time-consuming and resource-intensive.
Jaya Optimization Algorithm [46]	1. 10 benchmark UCI datasets	Feature Selection	1. Performance was superior compared to the genetic algorithm, particle swarm optimization algorithm, and differential evolutionary. 2. The results were not compared with recently introduced algorithms

Table 2.3. A study on hybrid search space optimization techniques

Methods	Dataset characteristics	Application	Comments
1. Minimum redundancy-maximum new classification information (MR-MNCI) method to filter the features, 2. Information Gain- binary butterfly optimization algorithm (IG-bBOA) to search for candidate feature subsets 3. Similarity-based ranking method to select the final feature subsets. [47]	6 UCI multiclass classification datasets	Feature Selection	1. The proposed three-stage hybrid feature selection method mitigated the issues of the conventional sigmoid-based binary butterfly optimization algorithm in terms of estimating the redundancy and relevancy of features. 2. The computational complexity of the proposed research work was very high
1. Genetic algorithm 2. Embedded regularization 3. hybrid $L_{1/2}$ + L_2 regularization approach [48]	5 Open source gene expression microarray datasets	Feature Selection	1. Novel chromosome representation (intron+exon) for global and local optimization procedures was proposed. 2. Validation is performed only on bioinformatics datasets so its generalization is questionable.
1. Mutual information maximization (MIM) 2. Adaptive genetic algorithm (AGA) [49]	6 Open-source UCI biological datasets	Feature Selection	1. Limited scope because the proposed methodology was validated only on biological datasets. 2. The performance of the approach was dependent on multiple hyperparameters making it computationally expensive to implement on real data.
1. binary flower pollination algorithm (FPA) 2. β-hill climbing (called FPAβ-hc) [50]	EEG dataset- (64 channels) 109 persons each with 14 different cognitive tasks	Channel Selection	1. Performance was superior to five state-of-the-art algorithms but hyperparameter tuning makes it slow and resource dependent.

1. Multi-view learning-based sparse optimization 2. CSP features with the $L_{2,1}$ -norm regularization [51]	1. BCI Competition III datasets Iva (118 channels), 5 participants, 2 classes (right hand and “foot). 2. BCI Competition IV Dataset I (59 channels), 7 participants, 2 classes (left hand, right hand)	Channel Selection	1. These results were found to be suitable for data sets with large small, midlevel, and large numbers of channels. 2. The procedure is quite time-consuming because of hyperparameter tuning and requires additional samples for validation.
1. Clustering-based multitask learning framework 2. Affinity propagation clustering scheme 3. Common Spatial Pattern (CSP) [52]	1. BCI Competition IV data set IIb (3 channels), 9 participants, 2 classes (right and left hand). 2. BCI Competition IV data set 1 (59 channels), 4 participants, 2 classes (left hand, right hand) 3. BCI Competition III data set IIIa (60 electrodes), 3 participants, 4 classes (left hand, right hand, foot, or tongue movement)	Channel Selection	1. Such a clustering procedure without feature selection may not provide the optimal solution since the true underlying subclasses present in the data may differ only with respect to a subset of the features. 2. Implementing the cross-validation requires additional samples for performance validation and is generally time-consuming, which limits the practicability of BCI systems, to some extent.

2.3. Dataset Details

This thesis explores two categories of datasets for the validation of the proposed search-space optimization techniques employed to solve the above-mentioned challenges. In the first category, three EEG datasets from publicly available BCI competitions are employed to validate the effectiveness of the proposed channel selection method. Dataset 1 (DS1) [53] consists of fewer electrodes (22 channels), Dataset 2 (DS2) [54] presents a moderate number of electrodes (59 channels), and Dataset 3 (DS3) [55] is densely packed with the highest number of electrodes (118 channels). To date, these datasets have been used in various BCI studies for results comparison. These results are standards and openly available to show the superiority of newly performed experiments.

In the second category, twenty benchmark datasets from the University of California Irvin (UCI) repository [56] for classification to validate the performance of the proposed feature selection method. The main motive behind selecting these datasets is that they are high-dimensional and encompass various research domains. The details of both categories of datasets are given below:

2.3.1. Datasets Used in Channel Selection Problem

2.3.1.1. Dataset 1 (BCI Competition IV- 2008 – II A)

This dataset (DS1) comprises EEG signals collected from 9 healthy participants. This spectrum consists of 22 EEG channels and 3 EOG channels with the left mastoid as the reference. It is a four-class MI task-based dataset where class 1 represents the left-hand movement, the right-hand gesture constitutes class 2, class 3 comprises the motion of both feet, and class 4 deals with the tongue activity. This dataset consists of individual training and validation EEG samples for all nine subjects to corroborate any classification scheme. Hence, there is no need to decompose given data samples into the training and validation sets using any cross-validation technique. This dataset can be directly downloaded from Link 1 given in the footer. The data were recorded from each participant in two different sessions having 6 runs per session (12 runs for the individual participant). Each run consists of 48 trials (12 for each class), which estimates 288 trials in a single session. The dataset was recorded with Ag/Agcl electrodes implanted with an inter-electrode distance of 3 centimeters. All the extracted signals

---

[1] https://drive.google.com/drive/folders/1UQnoq3SleL0_pQZ-HZWRQetywGwYVBLo?usp=sharing

were sampled with a 250Hz frequency using a bandpass filter to collect frequencies between 0.5 Hz to 100 Hz. Throughout the session, the amplifier frequency was fixed to 100 µV. Additionally, a notch filter was also applied to attenuate the band spectrum below 20Hz. All processed data is restored in General Data Format (.gdf extension) as one file per subject and session. The experimental setup of each trial is shown in Figure. 2.1. (A).

2.3.1.2. Dataset 2 (BCI Competition IV)

This dataset (DS2) was recorded from 7 healthy participants through 59 EEG channels. During the entire session, MI tasks were performed without any feedback. During the experiment, two participants performed left-hand (L) and foot (F) MI tasks while the rest accomplished the same motion with the right hand (R) and the left hand. In the first two runs, the visual cues are shown as left, right, or down keys. The cues were shown on the screen for 4 seconds, during which the subjects were asked to execute respective MI tasks. These periods are shown in the center of the computer screen interleaved with 2 seconds of a blank screen and 2 seconds with a fixed cross. All recordings were collected by using Ag/AgCl electrode cap. The calibration data with a sampling rate of 100 Hz consists of two runs, each of which contains 100 single observations. In our study, the first run of the calibration data was divided into training and testing samples of size 80 and 20 trials, respectively. The experimental setup of each trial is shown in Figure 2.1. (B). This dataset can be directly downloaded from Link 1 given in the footer.

2.3.1.3. Dataset 3 (BCI Competition III - Dataset IVa)

In this dataset (DS3), EEG signals are recorded from 5 participants (“aa”, “al”, “av”, “aw”, “ay”) through 118 channels. Each participant executes three MI tasks (left-hand, right-hand, right-foot), but data of only two (left-hand, right-hand) is used in the classification. The visual cues are shown for 3.5 seconds. For all participants, a total of 140 trials were collected and sampled with 100 Hz frequency. Initial 110 trials were used for training, and the rest were utilized in the evaluation session. The experimental setup of each trial is shown in Figure 2.1. (C). This dataset can be directly downloaded from Link 2 given in the footer.

---

[1] https://drive.google.com/drive/folders/1RMyav_rRlKqRoXxeXyt0zGVv_fbiM-hW?usp=sharing

[2] https://drive.google.com/drive/folders/17NlPLdkEY785Am-uifeW64Eixafw4A6v?usp=sharing

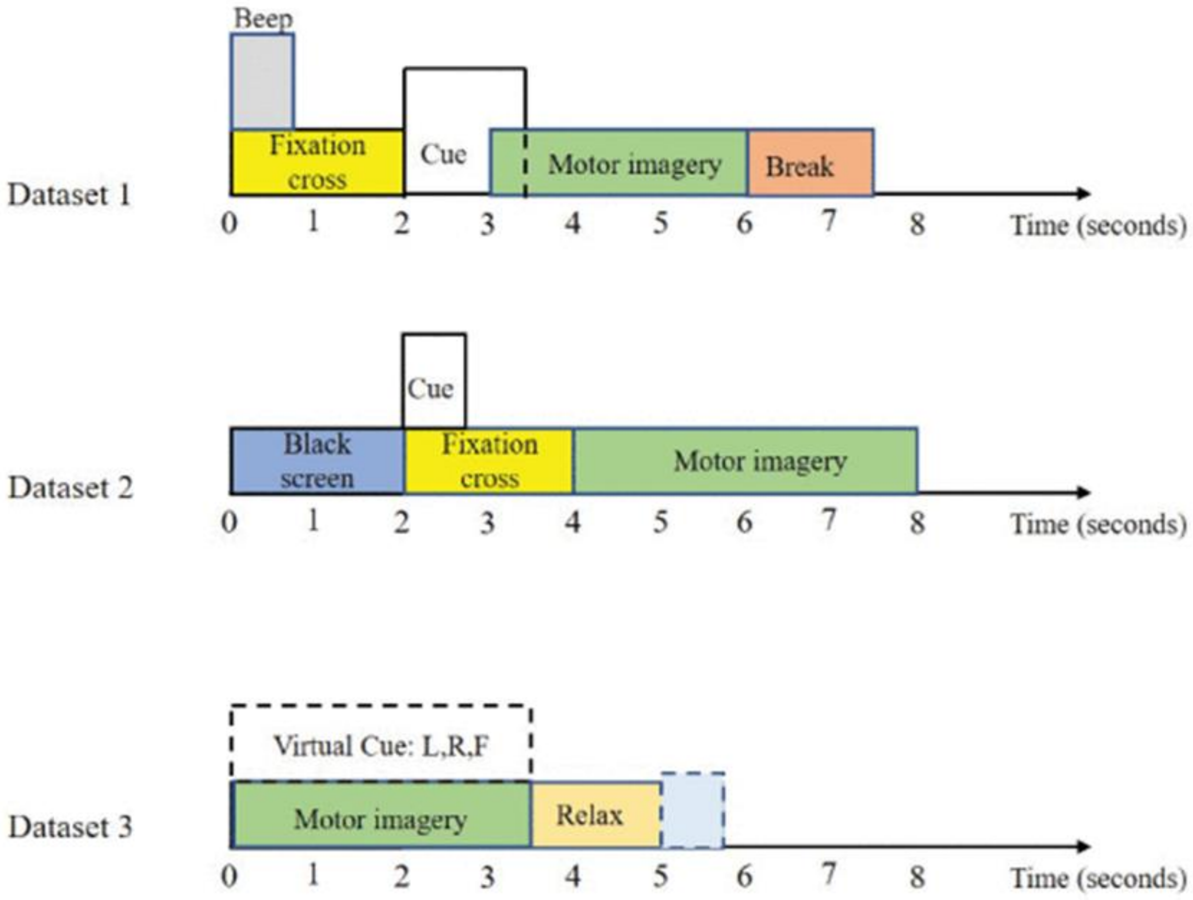

Figure 2.1. The timing sequence of the BCI experiment when only the motor imagery section executes from each dataset

2.3.2. Datasets Used in Feature Selection Problem

The details of the twenty datasets used to validate the performance of the proposed feature selection method are given in Table 2.4. Each dataset contains various characteristics in the context of attributes, sample size, and the number of classes. For example, Penglungew, TOX-171, and Yale are high-dimensional (>300) datasets but with fewer samples. Therefore, a proper cross-validation scheme is used to avoid the overfitting issue caused by the three datasets mentioned above. Similarly, CTG, Libras, OBS-Network, TOX-171, Vowel, Waveform, and Yale are multiclass (>2) datasets. All datasets are normalized before applying the proposed feature selection method. This dataset can be directly downloaded from Link 1 given in the footer.

[1] https://drive.google.com/drive/folders/1AZexl5PAfp7CJg_YdxcohFtWVzdtJJnj?usp=sharing

Table 2.4 Categorical distribution of UCI datasets used in this study

Index	Dataset	Number of features (D)	Number of samples (S)	No. of classes
1	Australian	14	690	2
2	Credit	20	1000	2
3	CTG	22	2126	3
4	Exactly	13	1000	2
5	Diabetic	20	1151	2
6	Hill Vally	100	606	2
7	Ionosphere	34	351	2
8	Libras	90	360	15
9	M-of-N	13	1000	2
10	OBS-Network	21	1075	4
11	Penglungew	325	73	2
12	QSAR	41	1055	2
13	Sonar	60	208	2
14	Spambase	57	4601	2
15	Spect	22	267	2
16	TOX-171	5748	171	4
17	Vote	16	300	2
18	Vowel	13	990	10
19	Waveform	21	5000	3
20	Yale	1024	165	15
Mean	-	383.70	-	-

[1] https://drive.google.com/drive/folders/1AZexl5PAfp7CJg_YdxcohFtWVzdtJJnj?usp=sharing