k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation

by

© Jacob White

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Engineering

Department of Engineering and Applied Science

Memorial University of Newfoundland

May 28, 2024

St. John's, Newfoundland and Labrador, Canada

Abstract

In passive brain computer interface (BCI) studies, a common approach is to collect data from mental states of interest during relatively long trials and divide these trials into shorter "epochs" to serve as individual samples in classification. While it is known that using k-fold cross-validation (CV) in this scenario can result in unreliable estimates of mental state separability (due to autocorrelation in the samples derived from the same trial), k-fold CV is still commonly used and reported in passive BCI studies. What is not known is the extent to which k-fold CV misrepresents true mental state separability. This makes it difficult to interpret the results of studies that use it. Furthermore, if the seriousness of the problem were clearly known, perhaps more researchers would be aware that they should avoid it. In this work, a novel experiment explored how the degree of correlation among samples within a class affects EEG-based mental state classification accuracy estimated by k-fold CV. Results were compared to a ground-truth (GT) accuracy and to "block-wise" CV, an alternative to k-fold which is purported to alleviate the autocorrelation issues. Factors such as the degree of true class separability and the feature set and classifier used were also explored. The results show that, under some conditions, k-fold CV inflated the GT classification accuracy by up to 25%. It is our recommendation that the number of samples derived from the same trial should be reduced whenever possible in single-subject analysis, and that both the k-fold and block-wise CV results are reported.

Acknowledgements

I would like to thank my supervisor, Dr. Sarah Power for her guidance and understanding throughout the difficult times working during the pandemic, and for her continued support afterwards. Without her contributions, this work would not have been completed.

I would also like to thank my fellow master's student, Hadi, for always being there to help during recording sessions, and to my friends for helping where they could.

Statement of Co-Authorship

Conceptualization, Jacob White (J.W.) and Sarah D. Power (S.D.P.); methodology, J.W. and S.D.P.; software, J.W.; formal analysis, J.W. and S.D.P.; investigation, J.W.; resources, S.D.P.; data curation, J.W.; writing—original draft preparation, J.W.; writing—review and editing, S.D.P.; visualization, J.W.; supervision, S.D.P.; project administration, S.D.P.; funding acquisition, S.D.P.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BCI brain computer interface. 1, 2, 6

CV cross-validation. 2, 19

DBS deep brain stimulation. 8

DE differential entropy. 15

ECG electrocardiography. 9

ECoG electrocorticography. 8, 9

EEG electroencephalogram. 1, 8

EMG electromyography. 9

EOG electrooculography. 9

fMRI functional magnetic resonance imaging. 8

ICA independant component analysis. 13

KNN k-nearest neighbour. 16, 18

LDA linear discriminant analysis. 16, 17

MEG magnetoencephalography. 9

mRMR minimal-redundancy-maximal-relevance. 16

NIRS near-infrared spectroscopy. 8

pBCI passive brain computer interface. 1, 6

RMS root mean square. 15

SNR signal-to-noise ratio. 12

SVM support vector machine. 16, 17

Chapter 1

Introduction

1.1 Problem Statement

A brain computer interface (BCI) is a system that translates information obtained via neurophysiological signals into commands for controlling an external device. Active BCI systems aim to provide individuals with severe motor disabilities a movement-free means of communication and environmental control by translating intentionally-modulated patterns of brain activity into control commands for external devices. Passive brain computer interface (pBCI) systems are potentially more broadly applicable; such systems are not for the intentional control of external devices, but rather they aim to enhance human-computer interactions by providing implicit information about the user's mental state (e.g., cognitive or emotional) [1]. An example would be a system that monitors a driver's level of fatigue/alertness and provides an alarm when a state of drowsiness is detected.

Due to the advantages it offers in terms of non-invasiveness, cost-effectiveness, portability, and high temporal resolution, the most common technique used in BCI research for obtaining the neurophysiological signals is electroencephalogram (EEG)

[2, 3]. How the recorded data is collected and used is largely based on the stage of development the system is in, and can be broadly grouped into two methodologies: online and offline systems. For a BCI to be functional and useable, online methods are used so that the system is capable of processing incoming information and evaluating the user's mental state in real-time. Therefore online implementations are the ultimate goal to be reached when developing BCIs, but to reach that point, the preliminary studies that develop and compare strategies make use of offline implementations. In offline implementations, a set of data is collected before any processing takes place.

Machine learning and statistical techniques are used to interpret the neurophysiological signals, and can be used to classify the mental states of interest. To evaluate the performance of classifiers trained offline, cross-validation (CV) techniques are used, the most common of which is randomized k-fold. In randomized k-fold CV, the data set is randomly partitioned into k subsets, and all but one of those subsets are used to train the classifier, while the remaining subset is used for testing. This process is repeated k times until all subsets have been used for testing the classifier exactly once, and the overall performance of the classifier (commonly the accuracy), is estimated as the mean performance of all k iterations of the CV process. This strategy, while very common, presents issues when used with temporally-structured data. As explained by Roberts [4], when data within a class is autocorrelated in time, it causes non-independence of residuals which violates core assumptions of many statistical methods. These violations can lead to overfitting and overly-optimistic results when verifying and comparing models.

Unfortunately, despite the potential issues, it is still very common to see k-fold CV used in pBCI studies with longer experimental trial durations that allow for within-class temporal correlations to be problematic. For example, in active BCI motor imagery studies trial durations are typically in the range of 5-10 seconds with only a single sample taken from each trial. Comparatively, in a pBCI study attempting to classify emotion, trial durations can be upwards of 60 seconds or more with multiple samples taken from each trial. Having multiple samples taken from a single trial leads to the issue that Roberts [4] discussed such that if samples within a class are temporally correlated, the classification results you get via k-fold CV might make it seem like the mental states are more separable than they truly are due to the influence of the correlated samples. Also discussed in [4] was an alternative to k-fold cross-validation, known as block-wise cross-validation that purports to be better as it alleviates temporal correlations. However, like k-fold CV, block-wise CV has not been empirically tested to determine how factors such as the strength of within-class correlation, class separability, and various classification algorithm specifics impact cross-validation biases. Understanding the potential magnitude of these effects, and under what conditions they exist is critical as offline studies serve to inform future and online studies. Therefore prior results must be able to be interpreted properly so that future studies are not based on unreliable results, and so that best practices are used going forward.

1.2 Research Objectives

The objective of this research is to investigate the reliability of EEG-based mental state classification results obtained using two common cross-validation techniques, namely k-fold and block-wise, when within-class temporal autocorrelation is present in the data. In this investigation, the effects of several different experimental and analytical factors will be considered, including:

1. The degree of within-class temporal correlation;

2. The true separability of the classes;
3. The classification algorithm used, and
4. The type of features used for classification.

1.3 Thesis Organization

The remainder of this thesis is organized into four chapters as described below.

Chapter 2 details the literature review for topics relevant to this thesis research. Passive brain-computer interfaces, signal acquisition and processing techniques, and classification and cross-validation methods are introduced and discussed. Relevant studies are also discussed to frame the purpose of the research objectives.

Chapter 3 contains the methods, results and discussion for the first study, which was an exploration of pre-existing data sets that guided the development of the novel study described in Chapter 4. The analytical methods used throughout both studies are introduced.

Chapter 4 contains the methods, results and discussion for the second study, which was based on an original experiment designed to meet the specific objectives of the work, as stated above.

Chapter 5 summarizes the primary conclusions of the study. It provides recommendations and guidelines for pBCI studies that follow the discussed data collection and analysis patterns. Study limitations and potential future work are also discussed.

Please note that parts of Chapters 1, 2, and 5, and the entirety of Chapters 3 and 4 of this thesis have been taken from a manuscript titled "k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation," by J. White and S.

D. Power that was published in the special issue “Advanced Machine Intelligence for Biomedical Signal Processing,” of the journal Sensors [5].

Chapter 2

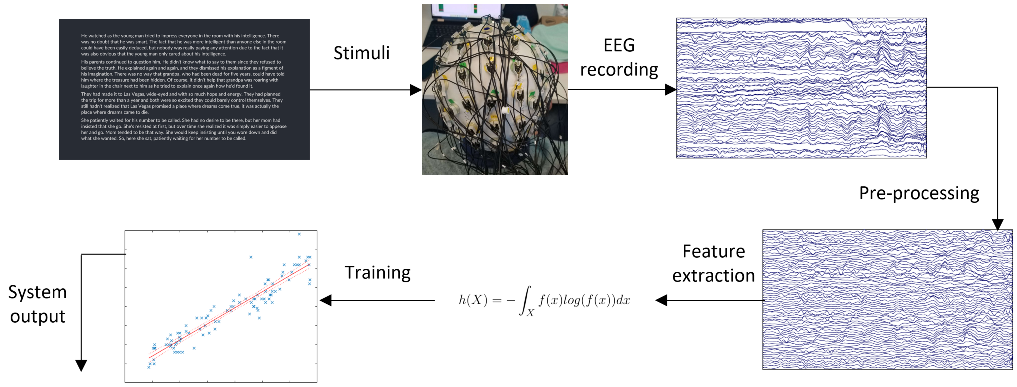
Literature Review

As noted at the end of Section 1.3, portions of this chapter and others were taken from J. White and S. D. Power, “k-fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation,” Sensors, vol. 23, no. 13, 2023, issn: 1424-8220. doi: 10.3390/s23136077, [5].

2.1 Brain Computer Interfaces (BCI)

A BCI is a system that translates information obtained via neurophysiological signals into commands for controlling an external device. Active BCIs aim to provide individuals with severe motor disabilities a movement-free means of communication and control, whereas passive BCI (pBCI) systems aim to enhance human-computer interactions by relaying implicit information about the user’s mental state (e.g., cognitive or emotional) to a computer [1, 6], and using this information to adapt the interaction to the user’s state in some useful way

Typically, a practical BCI system, whether active or passive, works as follows: 1) signals are acquired from the user via EEG, 2) the raw signals are pre-processed to remove unwanted artifacts, 3) features of the signals that are useful in discriminating the different mental states the BCI is meant to detect are calculated over many short segments/epochs of EEG to create samples, 4) this feature data is fed to a machine learning algorithm trained on previously collected neural data to classify/predict the current mental state, and 5) appropriate commands are sent to modify the connected external device as appropriate. This process is depicted in Figure 2.1.



He watched as the young man tried to impress everyone in the room with his intelligence. There was no doubt that he was smart. The fact that he was more intelligent than anyone else in the room could have been easily deduced, but nobody was really paying any attention due to the fact that it was also obvious that the young man only cared about his intelligence.

His parents continued to question him. He didn't know what to say to them since they refused to believe the truth. He explained again and again, and they dismissed his explanation as a figment of his imagination. There was no way that grandpa, who had been dead for five years, could have told him where the treasure had been hidden. Of course, it didn't help that grandpa was roaring with laughter in the chair next to him as he tried to explain once again how he'd found it.

They had made it to Las Vegas, wide-eyed and with so much hope and energy. They had planned the trip for more than a year and both were so excited they could barely control themselves. They still hadn't realized that Las Vegas promised a place where dreams come true, it was actually the place where dreams came to die.

She patiently waited for his number to be called. She had no desire to be there, but her mom had insisted that she go. She's resisted at first, but over time she realized it was simply easier to appease her and go. Mom tended to be that way. She would keep insisting until you wore down and did what she wanted. So, here she sat, patiently waiting for her number to be called.

Figure 2.1: BCI system overview.

The description of a practical BCI given above describes real-time, or “online”, classification. In online classification, the training data has been previously collected and used to pre-train the classifier, and then the user’s mental state is predicted in real-time as they use the device. The performance of the classifier can be assessed based on the accuracy of those real-time predictions versus the user’s true state. However, when investigating BCI systems for new applications (e.g. when considering previously unexplored mental states, user populations, or environmental conditions) or when testing/comparing new classification algorithms even in established applications, it is often more useful or convenient to perform the classification “offline”. In offline analysis, a relatively large set of neural data is collected from a group of participants that includes the different mental states of interest and any relevant experimental conditions, and the data is then stored to be analyzed later; there is no “real-time” prediction of the user’s mental state. This allows the researcher to explore and compare different techniques for classifying the mental states of interest, or examine the effects of different experimental conditions, much more efficiently than would be possible with online analysis. Offline analysis is often used in proof-of-concept studies to determine the best approaches to use in subsequent online studies.

2.2 Data Acquisition

Several devices can be used to obtain the neurophysiological signals that enable BCIs to function, and can broadly be categorized into invasive techniques, and non-invasive techniques. An invasive technique traditionally is applied via a break in the skin, or through an opening in the body. For BCI applications an invasive technique “must cross the scalp-skull level” [7], and could involve placing electrodes directly on the brain’s surface or in the tissue to be able to collect data directly from the brain, for example, electrocorticography (ECoG) or deep brain stimulation (DBS) [7–11]. The opposite of these are non-invasive techniques, which, as the name suggests, do not enter the body. Non-invasive techniques are typically applied on the surface of the skin, and for BCIs choices include EEG and near-infrared spectroscopy (NIRS) [7, 12, 13] among others.

Another consideration about the recording technology is its spatial and temporal resolution. With non-invasive techniques, there is often a trade-off between the two; EEG will provide high temporal resolution, but poor spatial resolution, whereas NIRS provides high spatial resolution, but poor temporal resolution [7, 14]. Other non-invasive techniques like functional magnetic resonance imaging (fMRI) [15, 16] and magnetoencephalography (MEG) [17, 18] can provide both a high temporal and spatial resolution, however, the bulk and cost of these techniques prohibit them from being widely applicable [14]. Invasive techniques like ECoG can obtain a better signal quality [19], however, the inherent risks prohibit them from common usage on humans [7]. BCI systems can receive supplementary physiological signals from non-brain sources such as electromyography (EMG), electrocardiography (ECG), or electrooculography (EOG) to improve performance in some cases, these systems are known as hybrid BCIs [7, 20].

2.2.1 Electroencephalography (EEG)

The most popular recording technology for BCIs is EEG [3, 7, 12]. For most practical EEG applications, multi-channel systems are used to record signals over the surface of the entire head. Common amounts range from 16-256 channels, and an example can be seen in Figure 2.2 (a). These are placed in standardized locations [21, 22] (see Section 2.2.3), typically pre-placed via an EEG cap which allows for greater convenience and improved readings due to superior fit. Looking at the system from the perspective of a single electrode however, one can see that there are four main components: signals obtained from the electrode are sent to an EEG amplifier, where the signal is conditioned so that it can be digitally sampled via an analogue-to-digital (A/D) converter and then finally sent to a computer for storage and further processing [21].

2.2.2 Electrode Options

Three broad categories exist for EEG electrodes: passive-wet, active wet or dry electrodes. Passive electrodes, simply put, are just electrical contacts. In order to

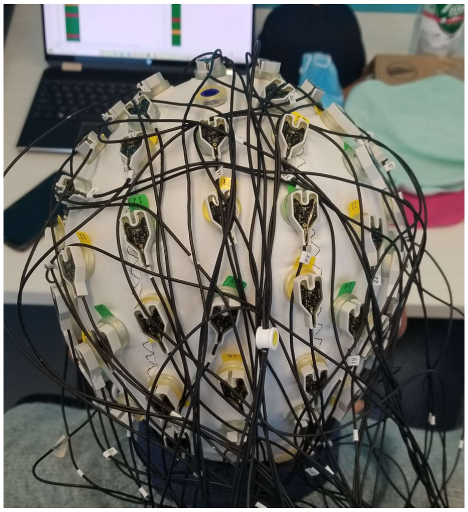


Figure 2.2: A 64-channel EEG-cap

obtain clear readings from passive electrodes a low electrode-tissue interface impedance is required, typically up to a value of $5k\Omega$ [21, 23]. This is obtained by applying an electrolytic gel/paste to the interfacing region to facilitate increased contact and conduction, hence, passive-wet.

Active-wet electrodes contain impedance conversion and active noise isolation circuitry. These additions allow active electrodes to measure signals with a low signal-to-noise ratio (SNR) while also having a much higher input interface impedance than passive electrodes, up to $50k\Omega$ [23, 24]. Like passive-wet electrodes, active-wet electrodes also require conductive gel. Active-dry electrodes, on the other hand, require no conductive medium to help facilitate connection but instead use force to push the electrode against the scalp. This system allows for faster setup times but introduces more noise into recordings compared to the wet electrodes [24, 25]. Active electrodes can also provide feedback to the user via integrated LEDs that convey the interface impedance - this allows for a faster setup as the user does not have to switch their focus from the electrodes to a computer screen constantly.

2.2.3 Standardized EEG Placement Systems

To ensure consistency and reproducibility, the International 10-20 system was developed to standardize electrode placement for EEG. The 10 and 20 refer to the distances placed between electrodes being 10% and 20% of the total distance across the skull [12, 22, 26, 27]. Extensions to the 10-20 system, the 10-10 and 10-5 systems, are used for higher resolution EEG systems that contain more electrodes [26, 27]. Alpha-numerical designations are given to each electrode, with letters representing specific coronal planes, whereas numbers designate sagittal planes [12, 22]. The 10-10 system is depicted in Figure 2.3.

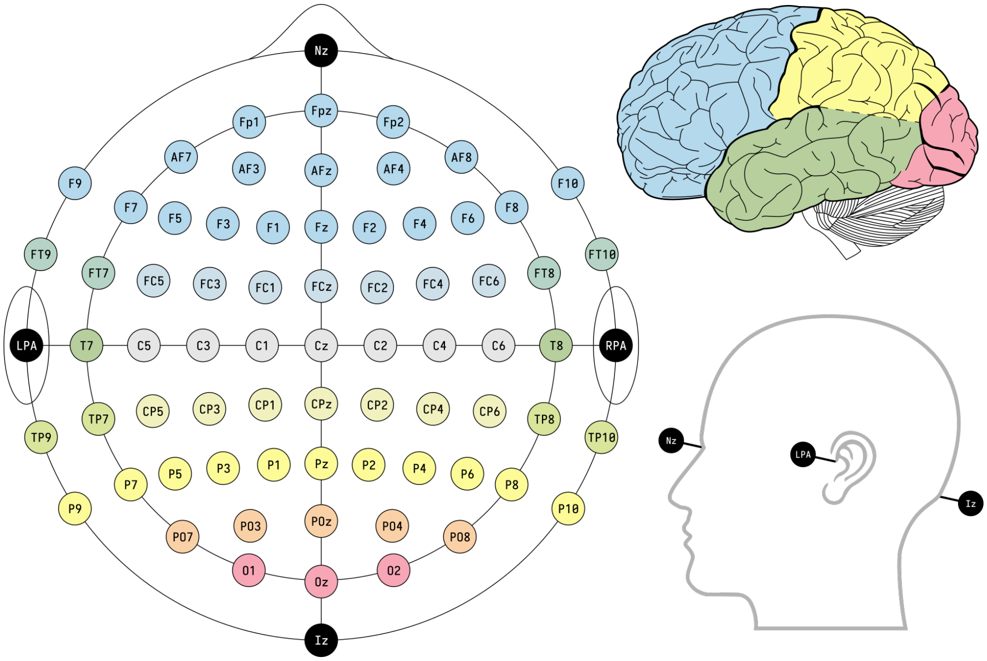


Figure 2.3: 10-10 electrode placement guide for higher-channel EEG systems.

2.3 Data Processing

2.3.1 Signal Pre-Processing

Due to EEG signals often being noisy and containing artifacts, the low signal-to-noise ratio (SNR) can lead to important information being obscured [28]. As such, rather than utilizing the raw EEG data, the recorded neurophysiological signals undergo a pre-processing procedure to prepare the data for further processing. Common sources of artifacts can broadly be grouped as physiological or non-physiological [28–30], some examples of which are explained below:

1. **Ocular Artifacts:** Eye movement and blinks can generate artifacts in EEG signals, particularly in the frontal electrode regions. They can introduce high-amplitude and high-frequency voltage changes that are often much larger than the EEG signal itself. Ocular artifacts can also be recorded via an electrooculogram which can be used to denoise the EEG signal [28, 31].

2. **Muscle Artifacts:** Muscle activity from talking, clenching, or swallowing among other possibilities, can contaminate EEG recordings [28, 29, 31]. Muscle artifacts can exist across a broad range of frequencies and are present across the entire scalp which can make them particularly difficult to eliminate unlike ocular artifacts [28, 30, 31].

3. **Power Line Interference:** Electrical interference from power lines, typically at 50 or 60 Hz, can contaminate EEG recordings [32, 33]. This interference manifests as rhythmic voltage fluctuations in the EEG signal.

4. **Electrode Impedance:** as the conductive gel dries after being applied, the

impedance between the scalp/electrode interface can change. High electrode impedance is known to increase noise and reduce signal quality [30, 34].

Several techniques are available to help negate these artifacts. For ocular and muscle artifacts, independant component analysis (ICA) is a ubiquitous technique used to identify sources of EEG activity that are localized to specific areas around the head [35], and is used to reject sources of eye and muscle activity [28–31, 36]. Power line interference can be removed simply through the usage of low-pass or notch filters, entirely removing the frequency that may cause issues, however, notch filters can cause ringing in the time-domain due to their sharp cut-off frequencies thus introducing an extra source of artifacts [32]. As such, alternatives to common filtering have been explored to remove power line interference, with extended-Infomax ICA [33] and spectrum interpolation [32] appearing as alternatives. As for conductive medium deterioration, while not an algorithmic solution, regular electrode impedance checks and proper electrode preparation techniques can help minimize impedance-related artifacts.

2.3.2 Feature Calculation

While raw EEG data can be directly used to classify mental states, generally the EEG data is subdivided into short time segments, less than 10 seconds, called "epochs." Summarizing and more informative values, known as features, are then calculated from the epochs to represent the EEG signals in a compact form. When features are grouped together for classification, the combined unit for a given epoch is known as an n-dimensional feature vector, where n is the total number of features.

Features can be calculated across the entire frequency range or can be subdivided further into frequency bands. Five commonly used frequency bands and some of their potential use cases are described below:

- **Delta** (δ)**, 1-4 Hz:** are associated with deep sleep, slow-wave sleep, quality of sleep [37, 38], and basic adult processes [39].
- **Theta** (θ)**, 4-8 Hz:** are often seen during light sleep, and are also associated with memory and the processing of new information [40, 41]. They have also been shown to be related to movement through one's environment [42]
- **Alpha** (α)**, 8-12 Hz:** are prominent when an individual is awake but relaxed with closed eyes [21]. Alpha activity is often used as an indicator of cognitive workload [43, 44], and is also related to long-term memory recollection [40, 41, 43].
- **Beta** (β)**, 12-30 Hz:** are associated with motor and cognitive processes [41], and are prominent in wakefulness [21]. They have also been noted when stopping memory recall [45].
- **Gamma** (γ)**, 30-50 Hz:** are the highest frequency categorization, and are involved in higher-level cognitive processes. Gamma activity is believed to represent the cognitive processing of sensory information [41, 46], and serves a function in working memory [41, 47].

Commonly used features include, but are not limited to, band power, time point, and connectivity features [48]. The features used in this thesis are discussed briefly in the following sections.

2.3.2.1 Band power

Band power features represent the average power of the EEG signal within a specified frequency range, such as the commonly used five frequency bands discussed above, and are widely used to determine when activity is occurring [48, 49]. They are used for a variety of classification scenarios, from motor imagery tasks [50], to mental workload tasks [51].

2.3.2.2 Differential Entropy

Classifying the emotion of the user is an area of interest in pBCI research, and to this end, differential entropy (DE) was found to be an effective feature [52, 53]. DE is defined as the logarithm of the energy spectrum of the signal as per Equation (2.1), for a time series X that obeys the Gaussian distribution $N(\mu, \sigma^2)$. It is suggested that DE should be calculated for each of the five frequency bands [52].

$$h(X) = \frac{1}{2}log(2\pi e\sigma^2) \tag{2.1}$$

2.3.2.3 Statistical Features

Statistical features capture various statistical properties of the EEG signals, such as mean, variance and standard deviation, root mean square (RMS), skewness, or kurtosis [49, 54]. These features provide information about signal amplitude distribution, symmetry, or shape characteristics. Statistical features are often used as basic descriptors in EEG classification tasks and can be combined with others.

2.3.3 Feature Selection

Once features have been calculated, the feature selection process typically follows in which only a subset of the total features calculated are ultimately used. Reducing the number of features used, also known as reducing the dimensionality of the data, helps prevent overfitting by reducing noise and irrelevant patterns that might bias the classifier [55]. This in turn can result in improved performance as high-dimensioned data can suffer from what is known as “the curse of dimensionality;” as the number of features increases, the amount of data required for the model to generalize also increases exponentially [50]. Reducing the dimension of the data has the additional benefit of being more computationally efficient, as larger dimension datasets require more resources and time to train.

The minimal-redundancy-maximal-relevance (mRMR) algorithm is a feature selection method that aims to identify the most informative and relevant subset of features from a larger set of available features [56]. MRMR is widely used in machine learning tasks and has been shown to be effective in pBCI classification [57] to improve classification accuracy and reduce the dimensionality of feature sets.

2.4 Classification

Classifiers serve as the means of turning recorded brain activity into actionable commands for BCIs. For EEG-based BCIs, common classifiers include linear classifiers, neural networks, and nearest-neighbour classifiers. Of these, linear classifiers are the most popular [48] and include options such as linear discriminant analysis (LDA), or support vector machine (SVM). As the name suggests, linear classifiers find a linear plane to divide the classes. Neural networks and nearest neighbour classifiers, such as k-nearest neighbour (KNN), allow BCIs to approximate non-linear functions to separate classes if needed. The following sections elaborate on how and under what conditions the mentioned classifiers perform best.

2.4.1 Linear Discriminant Analysis (LDA)

To be able to determine which class a previously unseen feature vector belongs to, LDA creates a decision boundary to separate the classes by using linear hyperplanes [50]. A hyperplane is a subspace that is one dimension less than the space it exists in, for example, in a 3-d space the hyperplane would be a 2-d plane. For binary classification, like in this work, newly presented feature vectors are classified by the side of the hyperplane they are on. For an example of this, see Figure 2.4. The LDA algorithm assumes a gaussian distribution for the data, and that all classes share the same covariance matrix [50, 58]. Despite EEG data being nonlinear [59], LDA often performs well and has low computational requirements, which is why it remains a popular choice for EEG-based BCI applications [48, 50, 60].

2.4.2 Support Vector Machine (SVM)

Similar to LDA, SVMs also use hyperplanes to identify classes. However, the hyperplane chosen is the one that maximizes the distance from the nearest training points of the classes, therefore creating an optimal decision boundary between the classes as seen in Figure 2.4. SVMs can also make use of kernel functions, that allow them to better handle non-linear data [58], as well as higher-dimensioned data [50, 58]. SVMs have been shown to be more effective for BCI systems than LDA [50], but also come with some additional disadvantages, such as being more difficult to find the optimal decision boundary when the dataset is noisy [58].

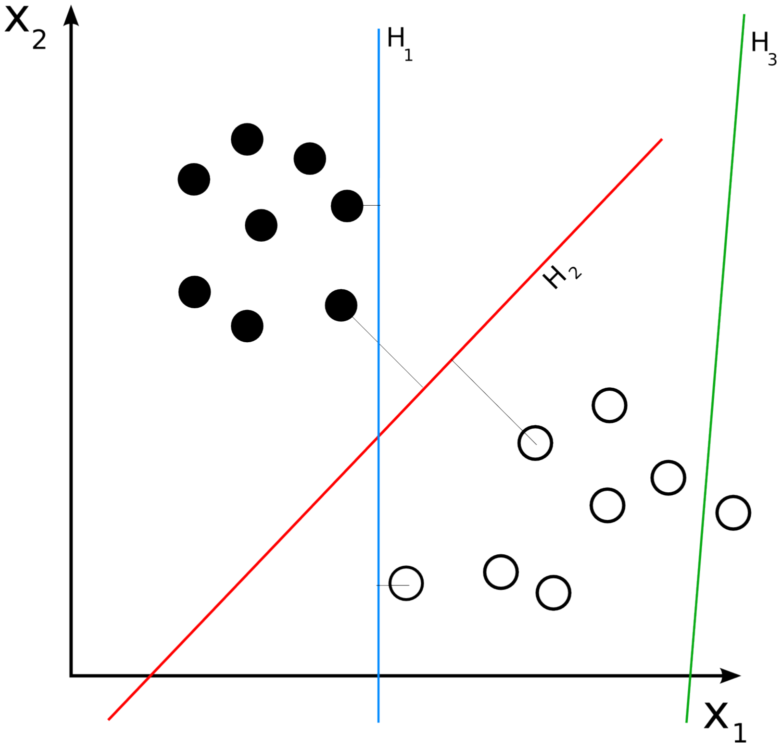


Figure 2.4: Linear hyperplanes that separate class 1 (solid dots) from class 2 (hollow dots). Boundary H_1 successfully divides the two classes and could be the decision boundary of LDA. Boundary H_2 divides the two classes optimally as it lies on the plane that maximally separates the classes, such a boundary could be found using an SVM classifier. Boundary H_3 fails to separate the classes.

2.4.3 K-Nearest Neighbour (KNN)

The KNNs algorithm at its core works by assuming similar things exist within close proximity to each other. KNN algorithms will classify unknown samples by finding the dominant class among its k nearest neighbours of the training data set [50], as seen in Figure 2.5 [61]. There are multiple ways to calculate what is considered “nearest,” but a common method is to calculate the euclidean, or straight line, distance between the point and each neighbour. Weighting neighbours that are closer to the sample in question can also be used to aid in determining which surrounding class is the most dominant. KNN can handle noisy data with the right selection of k [58], however, it is known that KNN classifiers are very sensitive to the curse of dimensionality [50], or in other words, the size of the feature vector compared to the training data set. For this reason, they are not very popular for BCI applications that use high-dimension feature vectors [50].

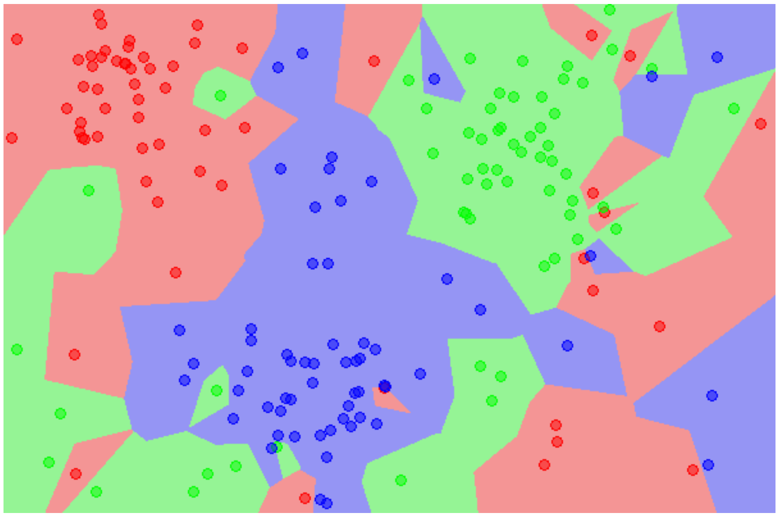


Figure 2.5: A depiction of the regions (colours) that would determine which class an unseen point belongs to based on the training data. [61]

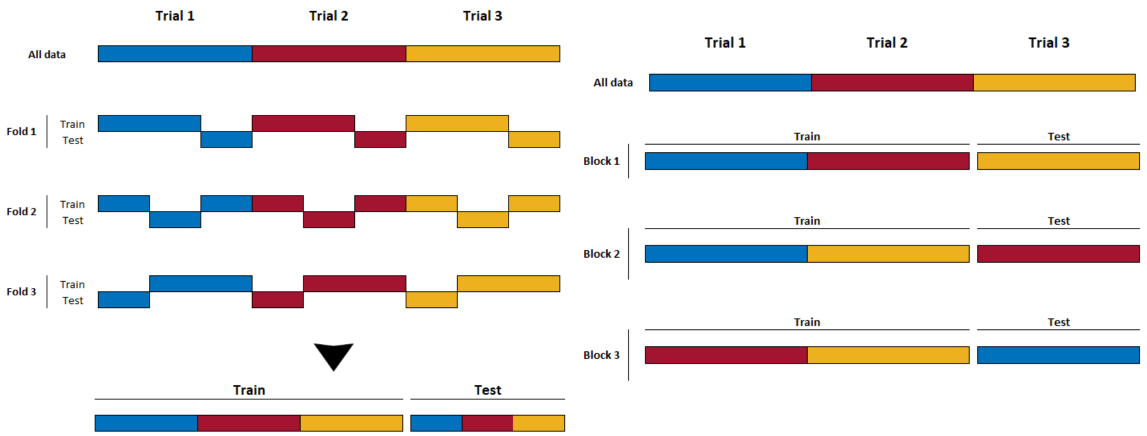
2.5 Cross Validation (CV)

One important consideration in offline studies is how the data is divided into training and test sets to estimate the classifier performance. This can be performed in a single train/test split of the samples or by cross-validation (CV). The most common CV technique is k-fold CV, in which the full dataset is randomly partitioned into k subsets of samples, and one of those subsets is retained for testing the classifier, while the remaining $k-1$ subsets comprise the training set. This process is repeated k times until all subsets (and thus, all individual samples) have been used for testing the classifier exactly once. The overall classifier performance is then estimated as the average of the resulting k classification accuracies from each step of the CV. Because there can be significant variation in the accuracy obtained with different train/test splits, this method yields a more generalizable estimate of classifier performance than taking just a single split.

While k-fold cross-validation is a very commonly used technique for evaluating machine learning algorithms offline, it can present issues with some types of data when the samples within classes are collected in close proximity in time, without randomization with the other class(es). For time-series data, similar to EEG, the process of randomly dividing all samples into k partitions results in the training and test sets containing samples from the same class that are highly correlated due to their proximity in time. This violates the assumption of independence that is critical to the validity of k-fold cross-validation [4]. The result is that the classifier could pick up differences between the classes that are actually just related to this temporal correlation of some samples, rather than to any true class-related difference.

This is typically not an issue in active BCI research, where most often (1) the trials collected are relatively short (less than 10 s) and, thus, only one EEG sample/epoch is calculated from each trial, and (2) the trial order of the mental states of interest (usually different mental tasks, such as motor imagery of different body parts) is randomized. However, for passive BCI studies, where the mental state (e.g., mental workload, fatigue) data often must be collected over longer trials (e.g., 30 s up to several minutes), this is a common issue. In such studies, in order to produce a sufficient number of samples in a reasonable period of time, multiple EEG samples/epochs are typically calculated from a single trial. For example, for a trial 1 minute in duration that represents a single mental state (e.g., high mental workload), it would be common to extract 60 individual samples calculated over consecutive 1-second, non-overlapping epochs. However, due to their proximity in time, these samples will likely be more highly correlated with one another than they would with samples from other trials, regardless of their class membership. As such, in a k-fold cross-validation analysis, when some of these 60 samples end up in the training set while others end up in the test set, this could result in the classifier being tuned to pick up these time-related similarities among samples instead of (or in addition to) any actually related to workload level. The consequence would be that the true mental state separability could be significantly overestimated.

An alternative approach that mitigates this issue in experiments with this trial structure and associated autocorrelation of samples is to perform block-wise (or trial-wise) cross-validation. In each step of block-wise CV, the trials are first randomly divided into a number of subsets b. The samples derived from the trials in one subset are held back for testing, while the samples from the remaining trials are used to train the classifier. This is repeated b times until all trials have appeared in the test set exactly once. The overall classifier performance is estimated as the average of the b resulting accuracies from each step. This partitioning strategy ensures that all samples from a single trial always remain together in either the training or test set, and, thus, temporal correlations will not influence the results as described above for k-fold CV. k-fold and block-wise cross-validation, as performed on datasets where multiple samples are extracted from a single class's trials, are illustrated in Figures 2.6a and b, respectively.



(a) (b)

Figure 2.6: (**a**) An example of k-fold CV; all folds result in epochs from a single trial being mixed into both the training and testing sets, like the example combination at the bottom. (**b**) An example of block-wise CV; by not breaking up the trial structure, epochs from a given trial remain exclusively in either the training or testing set.

While this issue has been acknowledged in some BCI papers [62–64] and it has been suggested that block-wise cross-validation should be the preferred approach [62, 65], the effects of using k-fold CV in such scenarios has not been explicitly investigated, and it is not clear how significantly (and under what circumstances) it can overestimate true mental state separability. Furthermore, it is not clear if block-wise CV will actually accurately estimate class separability. Unfortunately, despite the potential issues, it is still very common to see k-fold CV used in pBCI studies with longer experimental trial structures that allow for block-level temporal correlations to be problematic, so it is important that the effects are understood in order to help inform the best methodology for researchers to use in future studies and to help with the interpretation of the results of past studies that have used this approach.

2.6 Relevant Studies in Literature

The goal of this work is to improve our understanding of how common CV techniques perform under a variety of non-ideal circumstances, so as to better inform researchers on how to interpret results that they may see or to guide them towards better study design. To the best of our knowledge, no other studies have attempted to quantify the effects of intra-class auto-correlation on results obtained via cross-validation, as this work does. However, one other study has investigated alternate CV methods for use with EEG-based mental state classification.

In their study, Kingphai and Moshfeghi [63] acknowledge that the time series nature of EEG data does not suit traditional randomized CV and may violate the assumption of independence noted from [4], and proposed two alternative techniques to handle the evolving nature of mental workload and fatigue. The alternatives proposed were an expanding and rolling window strategy that does not take data from the future into account so as to not bias the forecasting of future mental states. These techniques were shown to provide high accuracies (upwards of 85-95% for the given tasks). While promising, this study did not compare these alternatives to results from k-fold CV, thus leaving uncertainty about if these alternatives provide superior classification accuracy, or rather a more “true” accuracy that is lower but more emblematic of the realistic capabilities of the model.

Chapter 3

Study 1: Preliminary Analysis on Existing Databases

3.1 Materials and Methods

3.1.1 SEED and DEAP Datasets

The SEED dataset [52, 53] is a popular EEG dataset for emotion recognition, consisting of fifteen subjects, each participating in three separate sessions. During each session, EEG was recorded while participants observed fifteen movie clips, each chosen to elicit one of three emotional categories: positive, neutral, and negative (five clips per category). The duration of each clip ranged from 3 to 4 min. Data collection was performed via an ESI NeuroScan System using a 62-channel EEG cap with the international 10–20 placement system at a sampling rate of 1000 Hz [53].

For the present study, only the first session from each subject was analyzed, and all trials were truncated to the length of the shortest trial, which was 185 s. Only four of the five trials for each emotion were used to allow for balanced trial-level randomization as an even number of trials per class is required (for reasons described later in section 3.1.3). Differential entropy (DE) features were calculated, since they have been shown to be effective for classifying emotional states [52]. Samples for classification were obtained by extracting 185 1-second non-overlapping epochs from each trial, and, for each one, calculating DE in the delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–50 Hz) frequency bands, for each of the 62 electrodes. This resulted in a 310-dimensional feature vector per sample. The minimum-redundancy maximal-relevance (mRMR) algorithm [56] was used during classification to reduce the feature set dimensionality to 30. The features of the training and testing sets were then z-score normalized separately, for each fold.

The DEAP dataset [66] is another popular EEG dataset for emotion classification, containing EEG data from 32 participants as they watched 40 1-minute music video clips. Unlike the movie clip stimuli used in the SEED experiment, the music clips shown during the DEAP trials were not specifically selected to elicit a particular emotion; instead, each participant rated the level of arousal, valence, like/dislike, dominance, and familiarity for each trial. Data was collected using a 32-channel EEG system with the international 10–20 placement at a sampling rate of 512 Hz.

For the present study, to facilitate binary classification, the trials were sorted in order of ascending valence rating and quantized into three levels: the top 10 trials were labelled as positive valence (+), the bottom 10 trials were labelled as negative valence (−), and the middle ten trials (i.e., trials 15–25) were labelled as neutral valence (0). Samples were calculated in the same way that they were for the SEED data: 60 1-second non-overlapping epochs were extracted from each trial, and differential entropy was calculated for the 5 standard frequency bands for each electrode. With 32 electrodes, this resulted in a 160-dimension feature vector for each sample. For consistency with SEED, the mRMR algorithm [56] was again used within each classification fold to reduce the feature set dimensionality to 30. Similar to SEED, the training and testing features were then separately z-score normalized for each fold.

For both the SEED and DEAP datasets, in the analysis described below, all binary combinations of the three emotional states were considered (i.e., positive vs. negative, positive vs. neutral, and negative vs. neutral).

3.1.2 k-Fold and Block-Wise Cross-Validation with True Class Labels

The first step in this preliminary analysis of SEED and DEAP datasets was to compare the classification accuracies estimated via both k-fold and block-wise cross-validation. For both datasets and each of the binary classification problems, 5 runs of 4-fold cross-validation for SEED and 5 runs of 5-fold cross-validation for DEAP were performed on each individual subject's data. The average of the runs-by-folds accuracies was calculated as an estimate of classifier performance.

Next, for both datasets and each of the binary classification problems, block-wise cross-validation was performed as follows: 5 runs of 4-block cross-validation for SEED and 5 runs of 5-block cross-validation for DEAP. For SEED, this resulted in a block consisting of three complete trials per class used for training and one complete trial per class for testing, and for DEAP, a block containing eight complete trials per class for training and two complete trials per class for testing. Block creation continues until all trials have been used for testing once. As with k-fold cross-validation, the average of the runs-by-blocks accuracies was calculated as an estimate of classifier performance. Note that the number of training and test samples for each step was the same between the k-fold and block-wise CV approaches.

For all scenarios, three different classifiers were considered: linear discriminant analysis (LDA), linear support vector machines (SVM), and k-nearest neighbour (KNN).

3.1.3 k-Fold and Block-Wise Cross-Validation with Randomized Class Labels

While useful, the results of the analysis above only tell us whether the accuracies estimated by the two different cross-validation methods are different from one another, but not whether the k-fold cross-validation is actually over-estimating the accuracy due to the trial structure of the experiment. To further investigate the effect of the temporal correlation among samples extracted from the same trial on the estimation of classification accuracy, we repeated the k-fold and block-wise cross-validation, but this time, we randomly shuffled the class labels first. The class labels were shuffled in two ways:

- **Trial-level randomization.** In this case, half of the trials from Class 1 were randomly selected, and all samples extracted from those trials were re-labelled as Class 2, after which the same was carried out for Class 2. This completely masked any true differences between the classes, while leaving the temporal correlations related to the trial structure intact. As half of the trials are swapped, this necessitates that the number of total trials per class be even to maintain balanced randomization.

- **Sample-level randomization.** In this case, half of the samples from Class 1 were randomly selected and re-labelled as Class 2, and then the same was carried out for Class 2. This again completely masked any true differences between the classes, but it also eliminated the temporal correlations related to the trial structure.

By combining the two CV techniques and two label randomization techniques, three additional tests of interest were included in the analysis: k-fold CV with sample-level randomization, k-fold CV with trial-level randomization, and block-wise CV with trial-level randomization. Through investigation of these hybrid k-fold and block-wise CV techniques with class label randomization tests, the potential effect of the trial structure on the classifier performance can be seen in the absence of any true class differences.

Again in all scenarios, three different classifiers were considered: linear discriminant analysis (LDA), linear support vector machines (SVM), and k-nearest neighbour (KNN).

3.2 Results

Table 3.1 shows all the classification accuracies (i.e., k-fold CV with true labels, block-wise CV with true labels, k-fold CV with trial-randomized labels, block-wise CV with trial-randomized labels, and block-wise CV with sample-randomized labels), averaged across 32 participants for DEAP and 15 for SEED, for each of the three classifiers (LDA, SVM, KNN) and binary classifications.

Figure 3.1 depicts a comparison of the classification accuracies for the k-fold and block-wise CVs. For both the SEED and DEAP datasets, the k-fold CV was found via paired *t*-tests to be significantly greater than the block-wise CV for all cases.

Figure 3.2 compares the random-labelled CV accuracies for the trial-randomized with k-fold, trial-randomized with block, and sample-randomized with k-fold CVs. Significance for these tests was determined via one-sample *t*-tests between the CV accuracies and chance (50%).

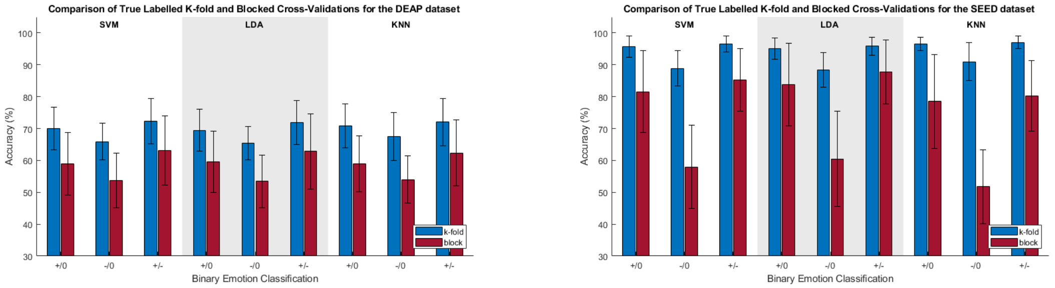
For both the DEAP and SEED datasets, trial-randomization with k-fold CV was

Table 3.1: Classification accuracies and statistical significance for the primary tests on DEAP and SEED. Columns denoted by ^{1} were compared together via a paired t-test, and columns denoted by ^{2} were compared to chance (%50) via one-sample t-tests. Bolded p-values indicate statistical significance. Under randomized labels, t.r. k-f, t.r. bl., and s.r. k-f, are short for trial-randomized k-fold, trial-randomized block, and sample-randomized k-fold CV, respectively.

			True Labels			Random Labels					
Dataset	Classifier	Classification	k-fold^1	block^1	^{1}	t.r. k-f^2		t.r. bl.^{2}		s.r. k-f^2	
DEAP	SVM	Pos-Neu	70.00	58.91	**p<.001**	63.28	**p<.001**	47.91	**p=.027**	50.05	p=.830
		Neg-Neu	65.90	53.75	**p<.001**	63.44	**p<.001**	51.43	p=.400	49.82	p=.533
		Pos-Neg	72.29	63.13	**p<.001**	62.70	**p<.001**	50.82	p=.512	49.84	p=.478
	LDA	Pos-Neu	69.44	59.63	**p<.001**	62.49	**p<.001**	50.16	p=.922	50.19	p=.487
		Neg-Neu	65.37	53.42	**p<.001**	64.57	**p<.001**	50.39	p=.793	50.14	p=.553
		Pos-Neg	71.83	62.81	**p<.001**	63.86	**p<.001**	50.06	p=.964	50.16	p=.378
	KNN	Pos-Neu	70.83	58.89	**p<.001**	65.07	**p<.001**	50.26	p=.823	50.17	p=.391
		Neg-Neu	67.45	54.01	**p<.001**	64.30	**p<.001**	50.03	p=.972	49.84	p=.371
		Pos-Neg	72.02	62.31	**p<.001**	63.80	**p<.001**	49.05	p=.370	50.21	p=.212
SEED	SVM	Pos-Neu	95.70	81.57	**p<.001**	83.86	**p<.001**	48.03	p=.484	49.63	p=.317
		Neg-Neu	88.89	57.95	**p<.001**	86.75	**p<.001**	42.88	**p=.028**	49.63	p=.480
		Pos-Neg	96.54	85.22	**p<.001**	83.83	**p<.001**	45.97	p=.143	49.71	p=.351
	LDA	Pos-Neu	95.10	83.80	**p<.001**	84.21	**p<.001**	41.35	**p=.018**	49.88	p=.699
		Neg-Neu	88.42	60.48	**p<.001**	84.79	**p<.001**	46.25	p=.278	50.06	p=.840
		Pos-Neg	95.85	87.86	**p=.001**	83.44	**p<.001**	45.44	p=.185	49.75	p=.288
	KNN	Pos-Neu	96.44	78.51	**p<.001**	91.52	**p<.001**	46.70	p=.560	50.21	p=.358
		Neg-Neu	90.94	51.73	**p<.001**	92.14	**p<.001**	48.17	p=.541	50.11	p=.723
		Pos-Neg	97.04	80.24	**p<.001**	91.53	**p<.001**	44.08	p=.217	49.83	p=.386

found to be significantly greater than chance for all scenarios. With any real class differences masked due to the class label randomization, it stands to reason that the classifier had been biased due to "high class-specific temporal correlations" among samples in the training and test sets resulting from the random partitioning of samples.

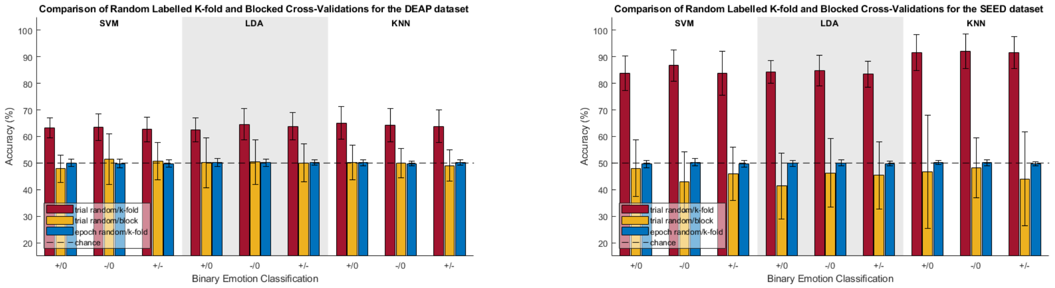
When using trial-randomization with blocked CV, most scenarios were not found to be significantly different from chance except for three outliers. The positive–neutral classification with the SVM classifier for the DEAP data, and the negative–neutral classification with the SVM classifier and positive–neutral classification with the LDA classifier for the SEED data were all found to be significantly less than chance.



(a) DEAP

(b) SEED

Figure 3.1: Comparison of true-labeled classification accuracies when using k-fold and blocked CV. +/0, −/0, and +/− are the positive/neutral, negative/neutral, and positive/negative classifications, respectively.



(a) DEAP

(b) SEED

Figure 3.2: Comparison of randomized-labeled classification accuracies of k-fold and blocked CV. +/0, −/0, and +/− are the positive/neutral, negative/neutral, and positive/negative classifications, respectively.

3.3 Discussion

This initial study provides the ground-work evidence of a latent problem within pBCI experiments that use longer duration trials in which many samples are extracted. As previously discussed, k-fold CV is a k-step process in which the samples are randomly divided into k partitions (maintaining class balance within each partition), and, at each step, one of the partitions is retained as the test set while the remaining $k-1$ partitions are used to train the classifier. The overall classifier performance is estimated as the average of the resulting k accuracies from each step. For both the SEED and DEAP data, relatively long trials of the different mental states of interest were collected, and we have extracted many different samples (60 for DEAP, 185 for SEED) from each trial. Thus, the random partitioning of the data in k-fold CV will result in the test set containing samples that are highly temporally correlated with samples in the training set. The concern is that this will bias the classifier and yield inflated classification accuracies that do not necessarily reflect the true separability of the mental states.

This concern was shown to be real, as can be seen when comparing the accuracies of the k-fold and blocked CV methods in Figure 3.1. Across all classification categories for the SVM classifier, with true class labels, the blocked CV accuracy was found to be significantly lower than its k-fold counterpart. Providing additional, and perhaps even more compelling, evidence of the temporal correlation effects were the randomized label tests, as seen in Figure 3.2. Whenever k-fold CV was used after trial-level class label randomization, results were found to be significantly greater than chance. With the SEED database, the k-fold CV results for the true and trial-randomized labels were all within approximately 5% of each other, and for the negative/neutral classification, the classification accuracy for the randomized labels was actually 1.2% higher than for the true labels (rand: 92.1%; true: 90.9%). With all real class differences masked due to class label randomization, these randomized test results indicate that k-fold CV results can be significantly inflated entirely due to the trial structure and associated underlying autocorrelation of the samples within trials. This casts doubt on the reliability of classification results obtained via k-fold CV under similar experimental conditions.

Chapter 4

Study 2: Original Experiment

4.1 Objectives

While Study 1 was able to show the presence of time-correlation effects on the cross-validation processes, the pre-existing datasets used were insufficient to be able to provide answers as to how much of an effect trial length, among other factors, has on these cross-validation techniques. As such, an original study was conducted to better answer these questions. By using multiple trial lengths within the same experiment, this study aimed to determine how trial length and class separability influence the outcomes of cross-validation techniques, how susceptible common classifiers and features sets are to time correlations in cross-validation, and whether current cross-validation techniques are over- or under-estimating a potential "ground-truth" accuracy.

4.2 Materials and Methods

4.2.1 Participants

In all, 12 healthy participants were recruited for this study (aged 24.75 ± 1.64 years; 11 male; 11 right-handed). Inclusion criteria required participants to be between 18 and 65 years old, have normal or corrected-to-normal vision and hearing, have no history of neurological disease, disorder, or injury, and have no cognitive impairment. Participants were asked not to exercise, smoke, or consume caffeine, alcohol, or other drugs within four hours of starting the experimental session. All participants provided written informed consent before participating. This study was approved by the Interdisciplinary Committee on Ethics in Human Research (ICEHR) at Memorial University of Newfoundland, NL, Canada.

4.2.2 Instrumentation

EEG signals were acquired via a 64-channel actiCHamp system (Brain Products GmbH, Gilching, Germany) at a sampling rate of 500 Hz. Electrodes were placed according to the international 10–10 system. The reference and ground electrodes were set as FCz and FPz, respectively. Electrode impedances were initially lowered to $\leq$ 10 kΩ and were then checked periodically during the session and reduced as necessary.

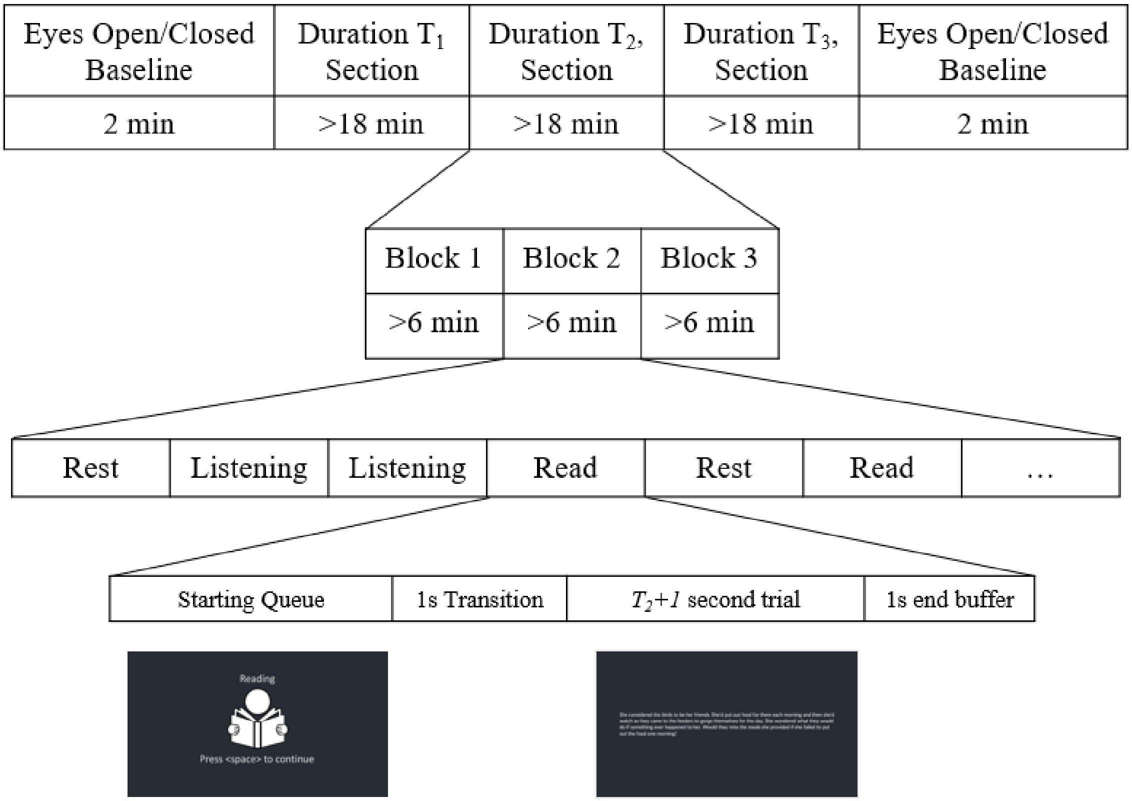
4.2.3 Experimental Protocol

Following set-up of the EEG system, participants were asked to complete two one-minute baseline trials—one with eyes open and one with eyes closed. During the rest of the session, participants completed three different types of trials: reading, listening, and rest. In order to investigate the effect of increasing the number of samples extracted from the same trial on the performance of the different cross-validation methods, 3 different trial lengths T_n (for $n = 5$, 15, and 60 s) were considered. The session was divided into three main sections, one for each of the trial lengths. There are six possible permutations of the three lengths, so each participant was pre-assigned a unique order in which they would complete the three sections (each permutation was completed by two participants).

For each section (i.e., trial length), a total of six minutes of data were recorded for each of the three tasks. These six minutes were divided into trials of the appropriate length for that section (i.e., for the T_5, T_{15}, and T_{60} sections, there were 72, 24, and 6 trials per task, respectively). Each section was divided into three blocks (two minutes of recording per task, per block). Participants were allowed to rest as needed between blocks. The order of presentation of the task trials was random in each section.

Trials began with an image and text cue, indicating the upcoming trial type and prompting the participant to press space when ready to start. Upon pressing space, a one-second blank transition screen appeared, followed by the appropriate task stimulus (described below) for $T_n + 1$ s. A one-second buffer (blank screen) was presented at the end, after which the next trial's prompt appeared. This protocol is summarized in Figure 4.1.

For the reading tasks, a passage (in English) was presented on the screen during the stimulus period. Participants were instructed to read at a comfortable pace in their head (not out loud) until the trial ended. The text passages, which were gathered from online free-to-use paragraph generators, were carefully selected to take slightly longer than the average adult reading speed for the given trial length; however, if a participant did finish before the end of the trial, they were asked to start reading again from the beginning of the passage. Auditory stimuli for the listening trials were generated via a paid online text-to-speech bot (from the same text passages used for the reading trials) so that all trials would have a consistent pacing, volume, and pronunciation. During these trials, the passages were played through computer speakers placed on the desk in front of the participant. For rest trials, a cross was presented on a blank screen for the participants to focus on (this cross was presented for the listening trials as well).



She considered the birds to be her friends. She'd put out food for them each morning and then she'd watch as they came to the feeders to gorge themselves for the day. She wondered what they would do if something ever happened to her. Would they miss the meals she provided if she failed to put out the food one morning?

Figure 4.1: Overview of the experimental protocol for recording multiple trial lengths.

4.2.4 Data Analysis

4.2.4.1 Pre-Processing

All pre-processing was performed on the aggregate of each participant's data. Pre-processing was completed in Matlab using the package EEGLab [36]. The following steps were used:

- Downsampling from 500 Hz to 250 Hz;
- Bandpass filtering from 0.5 Hz to 55 Hz;
- Artifact Subspace Reconstruction (ASR), using default settings, to remove channels that were poorly correlated with adjacent channels and to remove non-stationary high-amplitude bursts;
- Interpolating channels removed from ASR;
- Re-referencing data to the common average.

4.2.4.2 Feature Calculation

After pre-processing, the EEG data from each trial (of all trial lengths) was then segmented into five-second non-overlapping epochs. Next, the following features were calculated over each epoch: band power in the delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma bands (30–50 Hz), spectral entropy in the delta band, and root-mean-square (RMS) and variance across the entire spectrum (0.5–55 Hz), as they have been shown previously to be effective for reading tasks [51, 67]. All electrodes (excluding FPz as ground) were used for feature calculation. All features were z-score normalized based on the training set during each fold/block of cross-validation.

Note that samples were extracted from 5-second non-overlapping epochs because 5 seconds was the shortest trial length recorded. The T_5 trials were used to represent the "ground truth" class separability because, for these trials, only 1 sample was extracted per trial and because the task order was random; this means that, for the T_5 trials, the samples used for classification were completely randomized between classes across time (and thus contain no class-specific temporal correlations). The T_{15}- (with 3 samples extracted per trial) and T_{60} (with 12 samples extracted per trial)-length trials are meant to represent datasets with relatively low and relatively high (respectively) temporal correlations among samples within a class.

4.2.4.3 Classification: k-Fold and Block-Wise Cross-Validation

For each classification scenario considered, the following analyses were performed:

k-fold cross-validation: Samples were randomly divided into six (i.e., k = 6) equal subsets (balanced between classes). One of the subsets was retained for testing, while the remaining five subsets were used to train the classifier. Classification accuracy was then calculated. This was repeated six times, until all subsets (and thus, all individual samples) were used for testing the classifier exactly once. This whole process, starting with the random division of the samples into six subsets, was repeated five times. The overall estimate of classification accuracy was calculated as the average of the 30 resulting classification accuracies.

This six-fold CV process was performed with (1) true class labels, (2) trial-randomized class labels (see Section 3.1.3)), and (3) sample-randomized class labels (see Section 3.1.3);

Block-wise cross-validation: Trials were randomly divided into six equal subsets (balanced between classes). All samples from the trials in one of the subsets were retained for testing, while the samples from the trials in the remaining five subsets were used for training the classifier. Classification accuracy was then calculated. This was repeated until all subsets (and thus, all individual samples) were used for testing the classifier exactly once. This whole process, starting with the random division of the trials into six subsets, was repeated five times. The overall estimate of classification accuracy was calculated as the average of the 30 resulting classification accuracies. Unlike in the six-fold CV described above, in this approach, all samples extracted from a given trial always remain together in either the training or the test set.

This process was performed with (1) true class labels and (2) trial-randomized class labels (see Section 3.1.3).

4.2.4.4 Classification: Factors Considered

To move towards the objectives of Study 2, we investigated how the following factors influence the effect that temporal correlations have on the classification results from the k-fold and block-wise CV approaches:

- **Class separability:** Two different task pairs were considered: (1) read vs. rest, which represented the high-separability case, and (2) listen vs. rest, which represented the low separability case. These task pairs were chosen based on the results of pilot data; in preliminary analysis, the read vs. rest and listen vs. rest pairs were found to be separable, with approximately 96% and 68% accuracy, respectively, as determined using bandpower features and an SVM classifier. Note that read vs. listen was initially considered as well, but, because it was determined to have approximately the same separability as read vs. rest and, thus, provided no additional insights related to our research questions, it is not reported here;

- **Amount of class-specific temporal correlation:** Classification of the data from each trial length was considered separately, so that the results from each could then be compared. Because six minutes of data were collected (per task) for each trial length, once divided into five-second epochs, the number of samples used for classification was identical for the three different trial lengths. The only difference was the amount of temporal correlation that existed within the data for the 3 conditions; the 15 s length (with 3 samples per trial) represented relatively low temporal correlation, while the 60 s length (with 12 samples per trial) represented relatively high temporal correlation;

- **Feature types:** All calculated feature categories (i.e., band power, spectral entropy, RMS, variance) were treated separately and were not combined for classification. For all classification problems considered, feature set dimensionality was reduced to 30 via the mRMR algorithm [56] (this was performed using just the training set at the appropriate time within the cross-validation analyses);

- **Classifier:** The same as in the first study, three different classifiers were investigated: LDA, SVM, and KNN. Hyperparameters were optimized for each classification using Matlab's default automatic hyperparameter optimizer.

To summarize, we investigated the combination of 2 binary classification problems (the low and high separability cases), 4 feature sets (bandpower, spectral entropy, RMS, variance), and 3 classifiers (SVM, LDA, KNN), for a total of 24 classification scenarios. When considering the individual cross-validation tests, 5 were performed for the 15 and 60 s trial lengths (2 for true labels, 3 for randomized labels), and 2 were performed for the 5-second trial length (one for each the true and randomized labels, because k-fold and block-wise CV are identical in this case, as are sample-wise and trial-wise label randomization). This resulted in 288 individual cross-validation tests per participant.

4.2.5 Statistical Analysis

To better understand if and how factors such as amount of temporal correlation and true class separability influence the effect of the temporal correlation on the k-fold and block-wise cross-validation approaches, the following statistical tests were performed for every combination of class separability (low and high), amount of class-specific temporal correlation (3 samples per trial and 12 samples per trial), classifier, and feature set:

Classification with true labels: Accuracies for the k-fold and block-wise CV approaches were compared to the ground truth accuracy (as determined by the 5-second trials) via individual paired *t*-tests;

Randomized labels: Accuracies for the k-fold CV with trial-randomization and block-wise CV with trial-randomization were each compared to chance (50% for binary classification) via one-sample *t*-tests. Note that for the randomized label scenarios, the "ground truth" class separability is chance.

4.3 Results

Table 4.1 shows the most pertinent classification results (i.e., ground truth, k-fold CV with true labels, block-wise CV with true labels, k-fold CV with trial-randomized labels, and block-wise CV with trial-randomized labels) for all combinations of class separability, trial length, classifier, and feature set. For the "true-label" scenarios, over-estimations and under-estimations from the k-fold and block-wise CVs compared to the ground-truth are represented in blue and red, respectively. For the trial-randomized k-fold CV and trial-randomized block-wise CV tests, blue and red are used to indicate over-estimation and under-estimation compared to chance (i.e., 50%). Statistically significant results, as determined by the statistical tests described in Section 4.2.5, are denoted by bold font. Note that, for the k-fold CV with sample-randomized labels analysis, as expected, all scenarios were not significantly different from chance (average accuracies ranged from 48.4–52.0%, so they are not shown in the table).

To aid in visualization of the trends in the different cross-validation analysis for the factors of class separability and amount of temporal correlation, Figure 4.2 represents all classification results for the SVM classifier and bandpower feature set (general trends across these factors were similar for all classifier/feature set combinations).

Table 4.1: Classification accuracies for all classifiers and feature sets. Blue/red intensities denote the overestimation/underestimation of the CV techniques compared to the ground truth (G.T.) for k-fold and block columns, and compared to chance (50%) for trial-random k-fold (t.r. k-f) and trial-random block (t.r. bl.) columns. Heat map intensities are calculated separately for the ranges in [1], [2], and [3]. Bolded values were found to be significantly different as per Section 4.2.5.

			High Seperability					Low Seperability				
Classifier	Trial Duration	Feature	G.T.	k-fold[1]	block[1]	t.r. k-f[2]	t.r. bl.[3]	G.T.	k-fold[1]	block[1]	t.r. k-f[2]	t.r. bl.[3]
SVM	15s	Band power	94.58	93.43	92.41	**55.67**	49.73	67.41	67.34	**62.07**	**55.06**	49.32
		Spectral Ent.	95.02	93.68	**92.29**	**57.14**	49.21	65.71	67.36	64.62	**55.08**	51.72
		RMS	91.53	90.09	**89.12**	**55.38**	49.54	62.14	66.64	63.11	53.41	46.96
		Variance	91.83	90.29	**89.11**	52.92	50.07	62.18	65.59	63.84	52.85	48.84
		15s avg	93.24	91.87	90.73	55.28	49.64	64.36	66.73	63.41	54.10	49.21
	60s	Band power	94.58	94.58	**91.57**	**71.82**	50.61	67.41	**77.37**	**56.30**	**73.33**	46.22
		Spectral Ent.	95.02	94.43	**91.17**	**70.84**	45.53	65.71	**77.80**	**55.16**	**71.84**	58.40
		RMS	91.53	90.47	**86.53**	**62.52**	**42.67**	62.14	**72.05**	59.51	**64.47**	48.56
		Variance	91.83	90.58	**86.62**	**62.40**	**41.74**	62.18	**71.16**	58.50	**60.37**	53.85
		60s avg	93.24	92.52	88.97	66.90	45.14	64.36	74.59	57.37	67.50	51.76
LDA	15s	Band power	95.05	93.23	93.14	**55.96**	51.78	66.72	67.55	64.54	**55.87**	47.82
		Spectral Ent.	95.24	93.32	**92.97**	**54.98**	50.34	66.18	68.26	65.52	54.46	52.62
		RMS	92.55	**90.31**	**89.55**	53.54	47.60	63.18	66.46	63.76	**53.15**	**46.82**
		Variance	91.67	**89.44**	**88.32**	**53.44**	48.37	64.10	67.15	64.36	53.41	51.68
		15s avg	93.63	91.58	91.00	54.48	49.52	65.05	67.36	64.55	54.22	49.73
	60s	Band power	95.05	94.05	**91.00**	**73.58**	50.39	66.72	**78.08**	**56.97**	**73.45**	48.43
		Spectral Ent.	95.24	93.92	**90.60**	**72.50**	**40.96**	66.18	**77.89**	**57.78**	**72.25**	55.96
		RMS	92.55	91.60	**87.63**	**63.61**	46.83	63.18	**72.63**	61.47	**64.99**	48.62
		Variance	91.67	90.59	85.74	**64.26**	**43.47**	64.10	**71.96**	**57.91**	**65.07**	51.24
		60s avg	93.63	92.54	88.74	68.49	45.41	65.05	75.14	58.53	68.94	51.06
KNN	15s	Band power	92.85	92.23	91.76	**54.06**	50.39	62.78	65.06	61.50	**54.86**	50.00
		Spectral Ent.	93.09	92.75	91.93	**57.57**	**46.70**	62.67	**66.57**	62.67	**56.08**	49.58
		RMS	88.51	86.22	84.86	**56.04**	49.71	59.19	**63.00**	59.86	**51.81**	47.99
		Variance	87.77	85.82	84.72	**53.74**	49.19	60.87	62.55	59.84	**53.47**	51.94
		15s avg	90.55	89.26	88.32	55.35	49.00	61.38	64.29	60.97	54.05	49.88
	60s	Band power	92.85	93.65	88.53	**78.67**	**41.70**	62.78	**78.00**	**51.50**	**76.66**	51.27
		Spectral Ent.	93.09	93.84	**88.32**	**75.10**	48.26	62.67	**74.71**	**53.83**	**73.56**	46.25
		RMS	88.51	87.29	**81.34**	**64.70**	45.24	59.19	**67.55**	55.71	**65.25**	48.82
		Variance	87.77	87.57	**80.72**	**64.19**	47.28	60.87	67.08	**54.81**	**62.94**	49.13
		60s avg	90.55	90.59	84.73	70.67	45.62	61.38	71.83	53.96	69.60	48.87

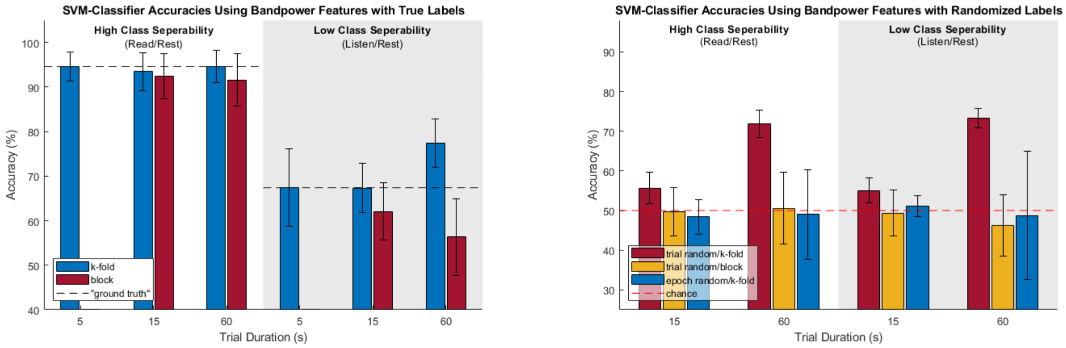


Figure 4.2: SVM-band power accuracies for true and random labelled cross validations.

4.4 Discussion

This study explored the limitations of conventional cross-validation techniques on single-subject EEG signal classification in scenarios where there was time-related correlation among samples within a class. Such scenarios are very common in offline passive BCI studies where, typically, relatively long trials of different mental states (e.g., low and high workload) are recorded, and multiple short epochs are extracted from these trials and serve as samples for classification. In this scenario, the use of k-fold cross-validation is problematic because the randomized division of samples into the k subsets means that, in any given "fold", some samples from a single trial end up in the training set while others end up in the test set. Additionally, because samples from a single trial will be highly correlated due to their proximity in time, this can potentially influence the classification accuracy and cause over-estimation of the true class separability. An alternative to k-fold CV is "block-wise" or "trial-wise" cross-validation, where trials are randomly divided into subsets, and short epochs are extracted from these trials to serve as samples for classification. In this case, all samples from a single trial always stay together in either the training or test set, and so the problem inherent in k-fold cross-validation should be eliminated. To the best of our knowledge, however, the actual effects of using k-fold CV in such scenarios had not been investigated, and it was not clear how significantly (and under what circumstances) it can overestimate true mental state separability. Furthermore, it was not clear if block-wise CV actually accurately estimates class separability. The motivation behind this work was to investigate the extent to which standard cross-validation techniques may misrepresent “true” class separability of EEG data, whether that may be over- or under- estimation, in scenarios where time-related correlations exist among samples within a class. The tests used in this study could be broadly applied to any EEG dataset using single-subject classification and a similar trial structure, to gauge the potential impact on the results.

For the case of high true class separability, k-fold CV produced results very close to the “ground truth” accuracies (as determined by the 5-second trials) which were between approximately 88% and 95%, depending on the specific classifier and feature set used. This was true for both the “low class-specific temporal correlations” and “high class-specific temporal correlations” conditions. There were only a few scenarios where the k-fold CV accuracy was significantly different from the ground-truth according to a paired *t*-test (and, in these cases, the k-fold CV result did not overestimate, but rather was actually *less* than the ground truth by about 2%). In the high true class separability case, block-wise CV did not perform quite as well as k-fold CV; while many results were fairly close, there were more scenarios where the results were significantly less than the “ground truth” accuracies. This was more pronounced for the case of “high class-specific temporal correlations,” where the blockwise CV result was, on average, 5% less than the ground-truth.

For the case of low true class separability, the amount of class-specific temporal correlation strongly affected the outcomes of the two CV methods. When there are low amounts of class-specific temporal correlation, both k-fold and block-wise CV provided results similar to the “ground-truth.” There were only three scenarios which were found to be significantly different from the ground-truth according to paired t-tests, two of which were over-estimations by k-fold CV (by about 4%) and one of which was under-estimation by block-wise CV (by about 5%). On average, k-fold and block-wise CV over- and under- estimated the ground-truth by 2.5% and 0.6%, respectively.

Lastly, for the case of low true class separability and high amounts of class-specific temporal correlation, the majority of both k-fold and block-wise CV methods were significantly different from the ground-truth. For k-fold CV, 11 of 12 scenarios were found to significantly overestimate the ground-truth and by a substantially greater degree (significant overestimations ranged from 7.86% to 12.09%) than in the previous case of low amounts of class-specific temporal correlation. For block-wise CV, 8 of 12 scenarios were found to significantly underestimate the ground truth, also by a substantially greater degree than the previous case (significant underestimations ranged from -6.05% to -11.11%).

Broadly speaking, the case of low class separability with low amounts of class specific temporal correlation (where we only used three samples per trial) performed adequately in comparison to the ground-truth; however, when high amounts of class-specific temporal correlation are present (for this study, 12 samples per trial), one can begin to see how k-fold CV becomes unreliable as the temporal correlation begins inflating the accuracies. To further highlight how the amount of class-specific temporal correlation can inflate accuracies, the randomized label tests from Section 3.1.3 were also used—of specific note are the trial-randomized k-fold (t.r. k-f) tests. This combination effectively negates any true class differences while retaining the amount of class-specific temporal correlation, and, in Table 4.1, all 24 scenarios of t.r. k-f were found to be significantly greater than chance. For the case of low true class separability, the t.r. k-f CV accuracies were often within a few percentage points of the true labeled k-fold CV accuracies, thus casting significant doubt on the veracity of those accuracy figures.

As mentioned in Section 4.3, epoch-level label randomization with k-fold CV tests were also performed to verify that, when no class differences and no time correlation were present, the outcomes of classifications were consistently chance. Trial-level label randomization with block-wise CV tests were performed to verify if block-wise CV was able to negate the time correlations present when using k-fold CV. This did generally result in accuracies much closer to chance, but the accuracies did not always average to exactly chance, rather showing some over- and under-shooting. It could be that block-wise CV introduces additional shortcomings into the cross-validation process that are not fully understood.

For the true labelled tests, the magnitude of over- and underestimation varied by feature. The bandpower and spectral entropy features consistently overestimated (with k-fold CV) and underestimated (with block-wise CV) the ground-truth accuracy by the greatest margins. On the other hand, RMS features consistently underestimated the ground-truth by a smaller margin when using block-wise CV, as compared to the other features, such that it was never found to be significantly different from the ground-truth in any of low class separability scenarios, regardless of the amount of class-specific temporal correlation. The choice of classifier also appeared to have a slight impact on the extent to which k-fold CV overestimated and block-wise CV underestimated the ground-truth accuracy, as can be seen in the feature average rows of Table 4.1. In the low class separability/long trial length case, where the differences from ground-truth were the greatest across the board, the overestimatation via k-fold CV and the underestimatation via block-wise CV were both larger on average for the KNN classifier than for either SVM or LDA. Given our previous explanation of the issue arising due to time-related correlations among samples from a single trial, it is not surprising that a classifier that predicts the class of a test sample based on the class membership of the nearest training samples would not be reliable.

With the evidence collected throughout this original study, it is the recommendation of this paper that, when reporting the results of analysis based on single-subject classification for pBCI applications that use a longer trial duration, authors should provide both the k-fold and block-wise cross-validation accuracies, as well as the accuracies for trial-level label randomization with k-fold cross-validation. Together, these metrics provide stronger evidence regarding the efficacy of new neural indicators presented to the field, as they indicate the influence of the possible compounding factor of time correlation between samples due to trial duration and experimental protocol. Nevertheless, it is worth nothing that, in passive BCIs, the block-wise CV scenario more accurately reflects the training/testing conditions for an online BCI, so even if it does underestimate the "true" class separability, it might represent what is practically achievable. Authors should also carefully consider trial duration during study design and attempt to reduce the duration of individual trials as much as possible, as this has been seen to have a positive effect on reducing time correlation effects.

Chapter 5

Conclusions

5.1 Primary Conclusions

We have evaluated two popular pBCI datasets (SEED and DEAP) using multiple cross-validation approaches and tests, and have also conducted an original study in which EEG signals were recorded while participants read and listened to selected text and audio prompts for three different trial durations (5 s, 15 s, and 60 s). After training three common classifiers (SVM, LDA, KNN) on four proposed features (bandpower, spectral entropy, RMS, variance), for each trial duration, the various cross-validation procedures were evaluated. This work evaluated how the degree of within-class temporal correlation and true class separability affect CV. In cases of high class separability, k-fold cross-validation produced a reasonably close estimate of the proposed ground-truth, while block-wise cross-validation produced a significant underestimation of the ground-truth. In cases of low class separability and longer trial durations, it was evident that the time correlations can begin to overtake the inherent class separability and cause an overestimation in classification accuracy when using standard k-fold cross-validation. This could lead to faulty conclusions about class separability to be drawn from the data. The results also showed that block-wise cross-validation may not be a perfect k-fold substitute, as, when used in the previous conditions, block-wise CV was found to significantly underestimate classification accuracies. This work also evaluated how cross-validation outcomes when using various classifiers and feature sets, were influenced when changing the amount of within-class temporal correlation and found that classifiers and features are not equally affected. This work also evaluated how several classifiers and feature sets were affected by the degree of within-class temporal correlation present, and found that classifiers and features are not equally affected. For instance, the KNN classifier produced more significantly divergent results than the LDA or SVM classifiers. By providing quantitative evidence, this work will make researchers more aware of the potential magnitude of the problem of using k-fold CV on data with temporal correlations, and encourage them to either avoid the problem all together through more appropriate experimental design, or if the problem is not feasible to avoid, report both the k-fold and block-wise results to allow readers to better interpret the results.

5.2 Study Limitations and Future Work

Though the experiment was very carefully designed and the analytical methods carefully chosen, there are some limitations of the study that are worth noting. Due to practical limitations on the duration of the experimental session, we were only able to use three different trial durations, representing three different "amounts" of temporal correlation, and also could only investigate three different levels of class separability (e.g., not separable/random, low separability, high separability). It would have been ideal to use additional trial durations and additional degrees of true separability to give a more precise picture of the conditions under which k-fold cross-validation results in a statistically significant divergence from the ground truth accuracy. It may be worth further investigating these factors in the future.

While we have focused on the common passive BCI experimental paradigm where multiple EEG samples from a given mental state are derived from a single long trial, this issue would also exist for paradigms where a single sample is taken per trial, but where the trials are not completely randomized across states/classes. Additionally, the results of this study are likely relevant to other commonly used physiological time-series signals, such as ECG, EMG, and fMRI, to name a few examples. Further investigation with these signal types may help to broaden the understanding of this issue.

Bibliography

[1] T. O. Zander and C. Kothe, “Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems in general,” *Journal of Neural Engineering*, vol. 8, no. 2, p. 025 005, 2011. DOI: 10.1088/1741-2560/8/2/025005.

[2] H. Berger, “Ueber das elektrenkephalogramm des menschen.,” *Journal für Psychologie und Neurologie*, 1930.

[3] C. Guger, B. Z. Allison, and N. Mrachacz-Kersting, “Recent Advances in Brain-Computer Interface Research—A Summary of the 2017 BCI Award and BCI Research Trends,” in *Brain-Computer Interface Research: A State-of-the-Art Summary 7*, C. Guger, N. Mrachacz-Kersting, and B. Z. Allison, Eds., Cham: Springer International Publishing, 2019, pp. 115–127.

[4] D. R. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, ISSN: 1600-0587. DOI: 10.1111/ecog.02881.

[5] J. White and S. D. Power, “k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation,” *Sensors*, vol. 23, no. 13, 2023, ISSN: 1424-8220. DOI: 10.3390/s23136077.

[6] P. Aricò, G. Borghini, G. D. Flumeri, N. Sciaraffa, and F. Babiloni, "Passive bci beyond the lab: Current trends and future directions," *Physiological Measurement*, vol. 39, no. 8, 08TR02, Aug. 2018. DOI: 10.1088/1361-6579/aad57e.

[7] A. Ortiz-Rosario and H. Adeli, "Brain-computer interface technologies: From signal to action," *Reviews in the neurosciences*, vol. 24, no. 5, pp. 537–552, 2013, ISSN: 0334-1763.

[8] S. Cousins, N. S. Blencowe, and J. M. Blazeby, "What is an invasive procedure? A definition to inform study design, evidence synthesis and research tracking," *BMJ Open*, vol. 9, no. 7, e028576, 2019, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2018-028576.

[9] A. M. Lozano *et al.*, "Deep brain stimulation: Current challenges and future directions," *Nature Reviews Neurology*, vol. 15, no. 3, pp. 148–160, 2019, ISSN: 1759-4758, 1759-4766. DOI: 10.1038/s41582-018-0128-2.

[10] A. L. Benabid *et al.*, "Deep brain stimulation: BCI at large, where are we going to?" *Progress in Brain Research*, vol. 194, pp. 71–82, 2011, ISSN: 1875-7855. DOI: 10.1016/B978-0-444-53815-4.00016-9.

[11] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140–154, 2011. DOI: 10.1109/RBME.2011.2172408.

[12] M. Soufineyestani, D. Dowling, and A. Khan, "Electroencephalography (eeg) technology applications and available devices," *Applied Sciences*, vol. 10, no. 21, 2020, ISSN: 2076-3417. DOI: 10.3390/app10217453.

[13] K. B. Beć, J. Grabska, and C. W. Huck, "Near-Infrared Spectroscopy in Bio-Applications," *Molecules*, vol. 25, no. 12, p. 2948, 2020, ISSN: 1420-3049. DOI: 10.3390/molecules25122948.

[14] M. R. Lakshmi, T. Prasad, and D. V. C. Prakash, "Survey on eeg signal processing methods," *International journal of advanced research in computer science and software engineering*, vol. 4, no. 1, 2014.

[15] R. Sitaram *et al.*, "Fmri brain-computer interface: A tool for neuroscientific research and treatment," *Computational intelligence and neuroscience*, vol. 2007, pp. 1–10, 2007, ISSN: 16875265. DOI: 10.1155/2007/25487.

[16] N. Weiskopf *et al.*, "Principles of a brain-computer interface (bci) based on real-time functional magnetic resonance imaging (fmri)," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 966–970, 2004. DOI: 10.1109/TBME.2004.827063.

[17] J. Mellinger *et al.*, "An meg-based brain-computer interface (bci)," *NeuroImage*, vol. 36, no. 3, pp. 581–593, 2007, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2007.03.019.

[18] L. Kauhanen *et al.*, "Eeg and meg brain-computer interface for tetraplegic patients," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 190–193, 2006. DOI: 10.1109/TNSRE.2006.875546.

[19] J. Wilson, E. Felton, P. Garell, G. Schalk, and J. Williams, "Ecog factors underlying multimodal control of a brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 246–250, 2006. DOI: 10.1109/TNSRE.2006.875570.

[20] G. Mueller-Putz *et al.*, “Tools for brain-computer interaction: A general concept for a hybrid bci,” *Frontiers in Neuroinformatics*, vol. 5, 2011, ISSN: 1662-5196. DOI: 10.3389/fninf.2011.00030.

[21] M. Teplan *et al.*, “Fundamentals of eeg measurement,” *Measurement science review*, vol. 2, no. 2, pp. 1–11, 2002.

[22] “Guideline 5: Guidelines for Standard Electrode Position Nomenclature. [Miscellaneous Article],” *Journal of Clinical Neurophysiology*, vol. 23, no. 2, pp. 107–110, 2006.

[23] K. E. Mathewson, T. J. L. Harrison, and S. A. D. Kizuk, “High and dry? Comparing active dry EEG electrodes to active and passive wet electrodes,” vol. 54, no. 1, pp. 74–82, 2017, ISSN: 00485772. DOI: 10.1111/psyp.12536.

[24] C. Fonseca *et al.*, “A Novel Dry Active Electrode for EEG Recording,” *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 1, pp. 162–165, 2007. DOI: 10.1109/TBME.2006.884649.

[25] E. Habibzadeh Tonekabony Shad, M. Molinas, and T. Ytterdal, “Impedance and noise of passive and active dry eeg electrodes: A review,” *IEEE Sensors Journal*, vol. 20, no. 24, pp. 14 565–14 577, 2020. DOI: 10.1109/JSEN.2020.3012394.

[26] V. Jurcak, D. Tsuzuki, and I. Dan, “10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems,” *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, 2007, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2006.09.024.

[27] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution eeg and erp measurements,” *Clinical Neurophysiology*, vol. 112, no. 4,

pp. 713–719, 2001, ISSN: 1388-2457. DOI: https://doi.org/10.1016/S1388-2457(00)00527-7.

[28] J. A. Urigüen and B. Garcia-Zapirain, "Eeg artifact removal—state-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, no. 3, p. 031 001, 2015. DOI: 10.1088/1741-2560/12/3/031001.

[29] M. M. N. Mannan, M. A. Kamran, and M. Y. Jeong, "Identification and removal of physiological artifacts from electroencephalogram signals: A review," *IEEE Access*, vol. 6, pp. 30 630–30 652, 2018. DOI: 10.1109/ACCESS.2018.2842082.

[30] S. Kotte and J. R. K. K. Dabbakuti, "Methods for removal of artifacts from eeg signal: A review," *Journal of Physics: Conference Series*, vol. 1706, no. 1, p. 012 093, 2020. DOI: 10.1088/1742-6596/1706/1/012093.

[31] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from eeg signals: A review," *Sensors*, vol. 19, no. 5, 2019, ISSN: 1424-8220. DOI: 10.3390/s19050987.

[32] S. Leske and S. S. Dalal, "Reducing power line noise in eeg and meg data via spectrum interpolation," *NeuroImage*, vol. 189, pp. 763–776, 2019, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.01.026.

[33] Z. Xue, J. Li, S. Li, and B. Wan, "Using ica to remove eye blink and power line artifacts in eeg," in *First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06)*, vol. 3, 2006, pp. 107–110. DOI: 10.1109/ICICIC.2006.543.

[34] E. S. Kappenman and S. J. Luck, "The effects of electrode impedance on data quality and statistical significance in ERP recordings," *Psychophysiology*, vol. 47, no. 5, pp. 888–904, 2010. DOI: 10.1111/j.1469-8986.2010.01009.x.

[35] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent Component Analysis of Electroencephalographic Data," in *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, 1995.

[36] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004, ISSN: 01650270. DOI: 10.1016/j.jneumeth.2003.10.009.

[37] S. Long, R. Ding, J. Wang, Y. Yu, J. Lu, and D. Yao, "Sleep Quality and Electroencephalogram Delta Power," English, *Frontiers in Neuroscience*, 2021, ISSN: 16624548. DOI: 10.3389/fnins.2021.803507.

[38] C. J. Davis, J. M. Clinton, K. A. Jewett, M. R. Zielinski, and J. M. Krueger, "Delta wave power: An independent sleep phenotype or epiphenomenon?" *Journal of Clinical Sleep Medicine*, vol. 7, no. 5 Suppl, S16–S18, 2011. DOI: 10.5664/JCSM.1346.

[39] G. G. Knyazev, "Eeg delta oscillations as a correlate of basic homeostatic and motivational processes," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 1, pp. 677–695, 2012, ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2011.10.002.

[40] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Research Reviews*, vol. 29, no. 2, pp. 169–195, 1999. DOI: 10.1016/S0165-0173(98)00056-3.

[41] C. S. Herrmann, D. Strüber, R. F. Helfrich, and A. K. Engel, "Eeg oscillations: From correlation to causality," *International Journal of Psychophysiology*, vol. 103, pp. 12–21, 2016, ISSN: 0167-8760. DOI: https://doi.org/10.1016/j.ijpsycho.2015.02.003.

[42] J. O'Keefe and M. L. Recce, "Phase relationship between hippocampal place units and the eeg theta rhythm," *Hippocampus*, vol. 3, no. 3, pp. 317–330, 1993. DOI: https://doi.org/10.1002/hipo.450030307.

[43] S. Palva and J. M. Palva, "New vistas for α-frequency band oscillations," *Trends in Neurosciences*, vol. 30, no. 4, pp. 150–158, 2007. DOI: 10.1016/j.tins.2007.02.001.

[44] O. Bazanova and D. Vernon, "Interpreting eeg alpha activity," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 94–110, 2014, Applied Neuroscience: Models, methods, theories, reviews. A Society of Applied Neuroscience (SAN) special issue., ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2013.05.007.

[45] R. Schmidt, M. H. Ruiz, B. E. Kilavik, M. Lundqvist, P. A. Starr, and A. R. Aron, "Beta Oscillations in Working Memory, Executive Control of Movement and Thought, and Sensorimotor Function," *Journal of Neuroscience*, vol. 39, no. 42, pp. 8231–8238, 2019. DOI: 10.1523/JNEUROSCI.1163-19.2019.

[46] E. Başar, "A review of gamma oscillations in healthy subjects and in cognitive impairment," *International Journal of Psychophysiology*, vol. 90, no. 2, pp. 99–117, 2013. DOI: 10.1016/j.ijpsycho.2013.07.005.

[47] C. Herrmann and T. Demiralp, "Human eeg gamma oscillations in neuropsychiatric disorders," *Clinical Neurophysiology*, vol. 116, no. 12, pp. 2719–2733, 2005, ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2005.07.007.

[48] F. Lotte *et al.*, "Areview of classification algorithms for eeg-based brain–computer interfaces: A 10 year update," eng, *Journal of Neural Engineering*, vol. 15, no. 3, p. 031 005, 2018, ISSN: 1741-2560.

[49] R. Jenke, A. Peer, and M. Buss, “Feature extraction and selection for emotion recognition from eeg,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014. DOI: 10.1109/TAFFC.2014.2339834.

[50] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, “A review of classification algorithms for EEG-based brain–computer interfaces,” *Journal of Neural Engineering*, vol. 4, no. 2, R1–R13, Jan. 2007. DOI: 10.1088/1741-2560/4/2/r01.

[51] A. Knoll *et al.*, “Measuring Cognitive Workload with Low-Cost Electroencephalograph,” in *Human-Computer Interaction – INTERACT 2011*, D. Hutchison *et al.*, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 568–571.

[52] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, “Differential entropy feature for EEG-based emotion classification,” in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, Nov. 2013, pp. 81–84. DOI: 10.1109/NER.2013.6695876.

[53] W.-L. Zheng and B.-L. Lu, “Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks,” *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, Sep. 2015, Conference Name: IEEE Transactions on Autonomous Mental Development, ISSN: 1943-0612. DOI: 10.1109/TAMD.2015.2431497.

[54] A. R. Hassan and M. I. Hassan Bhuiyan, “Automatic sleep scoring using statistical features in the emd domain and ensemble methods,” *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 248–255, 2016, ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2015.11.001.

[55] I. Iguyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005. DOI: 10.1109/TPAMI.2005.159.

[57] A. R. Subhani, W. Mumtaz, N. Kamil, N. M. Saad, N. Nandagopal, and A. S. Malik, "MRMR based feature selection for the classification of stress using EEG," in *2017 Eleventh International Conference on Sensing Technology (ICST)*, 2017, pp. 1–4. DOI: 10.1109/ICSensT.2017.8304499.

[58] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021. DOI: 10.1007/s42979-021-00592-x.

[59] W. Klonowski, "Everything you wanted to ask about EEG but were afraid to get the right answer," *Nonlinear Biomedical Physics*, vol. 3, 2009. DOI: 10.1186/1753-4631-3-2.

[60] N. Brodu, F. Lotte, and A. Lecuyer, "Comparative study of band-power extraction techniques for motor imagery classification," eng, in *2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, IEEE, 2011, pp. 1–6, ISBN: 9781424498901.

[61] Agor153, This work is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/. [Online]. Available: https://commons.wikimedia.org/wiki/File:Map1NN.png.

[62] R. Li *et al.*, "Training on the test set? an analysis of spampinato et al. [31]," *CoRR*, vol. abs/1812.07697, 2018. eprint: 1812.07697.

[63] K. Kingphai and Y. Moshfeghi, "On Time Series Cross-Validation for Deep Learning Classification Model of Mental Workload Levels Based on EEG Signals," en, in *Machine Learning, Optimization, and Data Science*, G. Nicosia *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2023, pp. 402–416, ISBN: 978-3-031-25891-6. DOI: 10.1007/978-3-031-25891-6_30.

[64] E. De Filippi *et al.*, "Classification of complex emotions using eeg and virtual environment: Proof of concept and therapeutic implication," *bioRxiv*, 2020. DOI: 10.1101/2020.07.27.223370.

[65] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012. DOI: 10.1016/j.ins.2011.12.028.

[66] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012, ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.15.

[67] P. Zarjam, J. Epps, and Fang Chen, "Spectral EEG features for evaluating cognitive load," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, pp. 3841–3844.