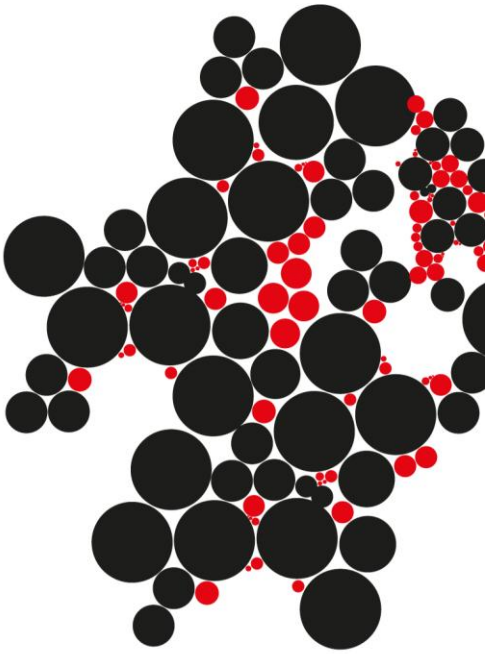
MASTER THESIS



AUTOMATIC CLASSIFICATION BETWEEN ACTIVE BRAIN STATE VS. REST STATE IN HEALTHY SUBJECTS AND STROKE PATIENTS

Victor Mocioiu

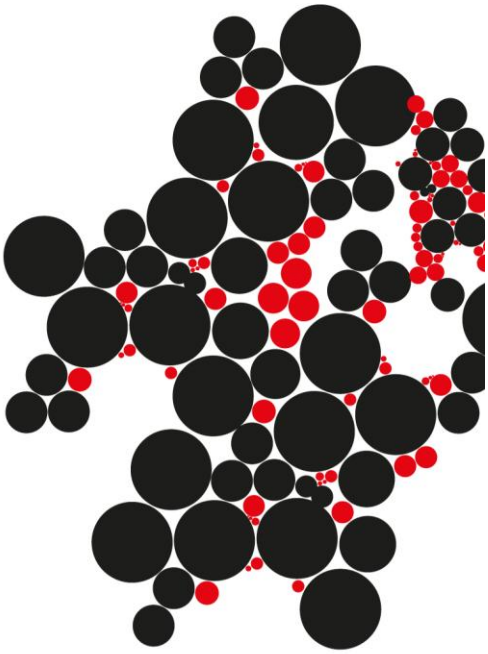
FACULTY OF ELECTRO-ENGINEERING, MATHEMATICS AND COMPUTER SCIENCES
CHAIR BIOMEDICAL SIGNALS AND SYSTEMS

EXAMINATION COMMITTEE
Prof. Dr. W.L.C. Rutten
Prof. Dr. Ir. MJAM Putten
Prof. Dr. Ir. J.R. Buitenweg
C. Tangwiriyasakul

DOCUMENT NUMBER
BSS - 028

UNIVERSITY OF TWENTE.

11/12/2012



UNIVERSITY OF TWENTE.

Abstract

Several methods exist for stroke rehabilitation. One method is the practice of motor imagery. The effect of this approach is improved by neurofeedback. This is done by using electroencephalographic (EEG) signals in a brain computer interface (BCI) setup. The BCI system should give the patient neurofeedback according to his sensorimotor rhythm.

Our goal was to find a way to model the two states associated with the sensorimotor rhythm: synchronized (rest) and desynchronized (active). For this purpose we have investigated four band power features: broad-band (8 - 30 Hz), α-band (8 - 13 Hz), β-band (13 - 30 Hz), and user-defined band and two classification methods: linear discriminant analysis (LDA) and support vector machines (SVM). Furthermore, we have employed a spatial filtering method, namely common spatial patterns (CSP), to see if classification outcomes could be improved. Since the eventual aim is to build a system that can be used at home, we examined several electrode configurations in order to find out the minimum number of electrodes needed to control the system. We extracted the features for different periods (8, 6, 4, and 2 seconds) to see what the influence on all of the above parameters was.

Results show that the highest performances were obtained on average for the broad-band feature, but the other features display good performances as well. We found that the highest classifier performances were obtained for the combination of CSP and SVM, with the general remark that SVM outperforms LDA. The minimum number of electrodes that was needed to ensure reliable control of the system was two. The investigated trial lengths seem not to influence all of the above parameters, good performances being found for all of them.

We consider that CSP is not suited for stroke data because it tends to focus on irrelevant aspects of the data. We deliberate that five channels is the minimum number of channels that can be used in an online system. We have also argued that the results are not influenced by trial length because the features are weakly stationary.

Table of Contents

1. Introduction

The human body is always active, even as we sleep. In order to assure the normality we know as everyday life unconscious activities take place, such as heart beating and regulation of body temperature. We also sense and move around in the external environment, which requires both voluntary and involuntary movements. Daily life also implies taking decisions, going through emotions, exchanging words with fellow humans, etc. The nervous system, the core of which is the brain, mitigates all of these actions.

The brain is divided into three main parts: the cerebrum, the cerebellum, and the brain stem. The cerebellum is responsible for regulating and coordinating movement, posture, and balance. The brain stem is associated with ensuring basic vital functions such as heart beating, blood pressure and breathing. The cerebrum itself may be subdivided into four parts called lobes: the frontal lobe, the parietal lobe, the temporal lobe, and the occipital lobe. Each lobe is "in charge" of certain functions. Broadly speaking, the frontal lobe deals with planning, movement, problem solving, etc. The parietal lobe is associated with movement, orientation, recognition and perception of stimuli. The temporal lobe is involved in memory, speech, and processing auditory stimuli. The occipital lobe mainly deals with visual processing.

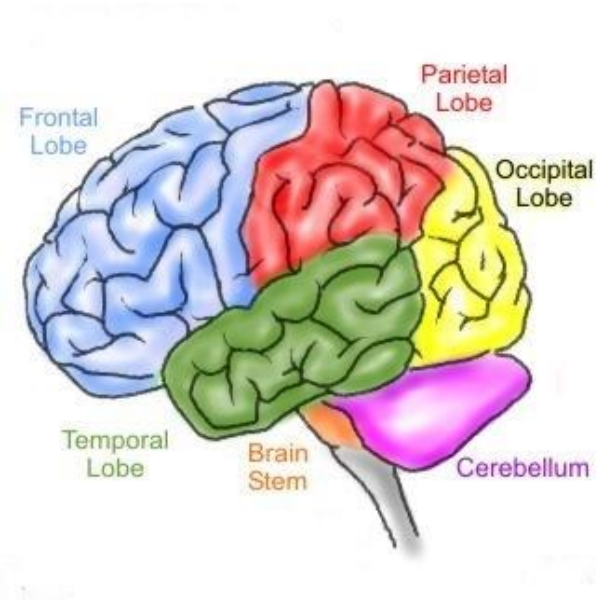


Figure 1.1 Lateral view of the surface anatomy of the brain, showing the brain stem, cerebellum and the four lobes of the cerebrum. Taken from [2].

Unfortunately, the brain is also prone to many neurological impairments that lead to some form of physical and/or mental problem. An affection of the brain may be categorized according to the

dysfunction that it causes: loss of memory – amnesia, impairment of language – aphasia, inability to recognize shapes, persons, etc. – agnosia, some form of speech disorder – dysarthria, and the loss of the ability to carry out learned movements - apraxia [1].

1.1. Stroke

One of the most common affections of the brain is stroke (or cerebrovascular accident - CVA) and it may be the cause responsible for any of the aforementioned dysfunctions [2]. CVA is caused by a sudden limitation of the flow of blood to a part of the brain. The bottleneck happens either due to ischemia (80-90% of all cases) or to hemorrhage (10-20% of all cases). Ischemic stroke can be either thrombotic or embolic. Thrombotic CVA is the result of a blood clot in a vein or artery of the brain; embolic CVA happens due to an embolus that adheres to the wall of an artery thus blocking the blood flow.

Depending on the quantity of tissue affected and the location of the stroke the symptoms can be: right side - paralysis on the left side of the body, vision problems, etc., left side - paralysis on the right side of the body, speech/language problems, etc. [1,2]. In this study we will focus on movement impairments caused by stroke. This means that stroke has occurred somewhere in the motor cortex.

Stroke proves to be a heavy burden on the affected and on society [3,4,5]. Heavily affected stroke survivors cannot be integrated fast and easily back into daily life and need the help of others to lead a close-to-normal life. This, also negatively impacts the wellbeing of stroke caregivers (both professionals and family) who often end up being predisposed to depression [4,5,6]. This again leads to aggravating the psychological status of the stroke patient.

The issue that arises is what to do in order to accelerate the reintegration into daily life of stroke patients? Given that the tissue area affected by stroke is no longer functional, it would be desirable that adjacent areas take over its activity. In other words, induce plasticity thus restoring normal activity. The usual way of achieving this is by stroke rehabilitation methods.

1.1.1. Stroke rehabilitation

Most common post-stroke rehabilitation protocols imply that the patient comes to the hospital for regular training sessions. A normal session implies diverse physical exercises: from active movement, when the patient tries to complete a task by himself using his affected side, to passive movement where a caregiver helps the patient perform the movement.

There are also alternatives to the usual rehabilitation procedures. One such procedure is via biofeedback: a process during which subjects are given information about subconscious physiological processes. This information is then used by the subject to learn to control the process. For example, in one 12-week study by Crow et al. [7] the electromyogram (EMG) activity is used to relay biofeedback. Crow uses a voltmeter, connected to the EMG electrodes, for visual feedback and a speaker, connected to the same system, for sending click sounds as audio feedback. Forty subjects were recruited and divided in two groups – experimental group and control group. For the experimental group the voltmeter was placed within visual range and the auditory feedback was turned on. The electrodes were positioned on a target muscle selected according to the subject. Electrodes were also placed on the subjects from the control group, as a placebo, but they did not receive any visual or audio feedback. The exact tasks that the subjects had to perform are not reported in the article. The outcome of the experiment was assessed using the Action Research Arm test and the Fugl-Meyer assessment. Results show greater improvement in the experimental group than in the control group. The authors conclude that this method of biofeedback *"has more potential as a component of physiotherapy then some previous studies".*

Another procedure that has been shown to improve rehabilitation outcomes is mental practice (or motor imagery; from here on referred to as MI). Moreover, MI can be used in the case where a stroke patient cannot move his hand at all. MI is defined as "the process of imaging and rehearsing the performance of a skill with no related overt actions" [8]. In a series of experiments conducted by Page et al. [9, 10, 11, 12] the integration of MI in a stroke rehabilitation protocol is researched. In one of these studies [12] 32 subjects underwent a protocol designed to compare between two groups: one that only did physical practice and relaxation, and one that combined physical practice with MI; the study lasted six weeks. The mean age of the subjects was 58.69 (SD 12.89) and the time since stroke was between 12 and 174 months. For the motor task, the subjects were asked to reach for and grasp an object (week 1 and 2), turn the pages of a book (week 3 and 4) and try to write with a pen (week 5 and 6). The MI group also performed mental practice of the motor task. Action Research Arm test and Fugl-Meyer assessment were used to evaluate the subject's evolution. The results of this study are shown in Table 1.1. They show that at the end of the study the MI group could perform the motor task better than the other group, presenting significant improvements. It is also concluded that "*a traditional rehabilitation program that includes mental practice of tasks practiced during therapy increases outcomes significantly*".

Table 1.1: Action Research Arm and Fugl-Meyer results after six weeks of therapy protocol. Results show that the group which also performed MI has considerably better scores. (P =0.0001 Wilcoxon test comparing the 2 groups for FM, and P<0.0001 for ARA; taken from [12])

	Action Research Arm			Fugl-Meyer		
	Pre Mean (SD)	Post Mean (SD)	Mean Change (SD)	Pre Mean (SD)	Post Mean (SD)	Mean Change (SD)
Physical Practice Only Group	17.25 (14.29)	17.69 (13.75)	+0.44 (2.03)	35.75 (9.51)	36.75 (10.74)	+1.0 (3.68)
Physical Practice + MI Group	18.00 (10.99)	25.81 (11.29)	+7.81 (0.3)	33.03 (9.37)	39.75 (6.86)	+6.72 (3.68)

A different study conducted by Crosbie et al. [13] was done on ten stroke subjects and showed the positive outcome of MI. Improvements were measured using the Upper Limb Motricity Index method. The mean age of the subjects was 63.9 (SD 10.94) and the time since stroke was between 10 days and 176 day. None of them could perform physical actions with the most affected arm without assistance. The task consisted of imagining reaching for a cup placed on the table, bringing the cup to the mouth and putting it back on the table. Sessions lasted between 25 and 45 minutes and were carried out for a period of two weeks for each subject. According to the Upper Limb Motricity Index, 8 out of 10 subjects showed improvements at the end of the 14 days of training. No control group was present in this study. Results indicate that, even without physical practice, MI may lead to an improvement of the stroke subjects' condition.

Dijkerman et al. [14] fortifies the assumption that MI may be used for stroke rehabilitation. The methods that were used to assess the outcome of this study were the Barthel Index (BI), Hospital Anxiety and Depression Scale (HADS), Modified Functional Limitations Profile (FLP), Recovery Locus of Control Scale (RLOC) and Test of Everyday Attention (TOEA). In this study the mean age of the subjects was 69 (SD = 9) and they had suffered a stroke between 12 months and 48 months earlier. The 20 subjects participating in the study were split into three groups: motor imagery (10 subjects), visual imagery (5 subjects) and no imagery (5 subjects). The last two groups were considered as a control group.

All three groups started the protocol with a common motor task called in this study the *"training task"* (real movement). The task consists of sequentially moving a row of 10 independent 2 cm^3 blocks set up in a line to another line situated 25 cm away from the initial one. After the motor task the MI group performed mentally the same task. The visual imagery group rehearsed imagining a set of pictures that were presented after the motor task. The images that were shown to the visual imagery group were static; i.e. did not contain movement. Results are summarized in Table 1.2. This study advocated that *"there was a greater improvement on the training task (motor task) in the motor imagery group as compared with the control group".*

Table 1.2: Results before and after four weeks of training reveal that the improvements shown by the MI group are higher than the ones of the control group. The higher the values the better. These suggest that MI is a valid approach to maximize the results of a stroke rehabilitation protocol. Taken from [14]

	BI		HADS		FLP		RLOC		TOEA	
	Pre Mean (SD)	Post Mean (SD)	Pre Mean (SD)	Post Mean (SD)	Pre Mean (SD)	Post Mean (SD)	Pre Mean (SD)	Post Mean (SD)	Pre Mean (SD)	Post Mean (SD)
Control Group	95.56 (9.84)	95.56 (9.84)	53.44 (11.80)	57.25 (8.40)	13.89 (7.85)	13.78 (6.80)	35.44 (2.74)	35.33 (3.32)	12.89 (3.41)	14.22 (2.86)
MI Group	95.56 (6.36)	96.11 (6.51)	52.76 (14.95)	50.02 (13.75)	17 (5.27)	16.22 (3.90)	36.67 (5.39)	35.78 (4.27)	12.44 (14.75)	13.56 (3.71)

The three aforementioned studies show that MI improves rehabilitation outcomes but some issues remain. First, there is no reliable measure of the mental implication of the stroke patient. Second, the patient himself does not know how well he is performing the MI task. A possible solution for these problems is to combine the method of biofeedback with MI. As this type of biofeedback uses brain signals it is called neurofeedback.

Brain signals may be acquired with various methods, but the only methods that have the temporal resolution necessary to transmit fast feedback are magnetoencephalography (MEG, usually not used in studies because of the high cost of the equipment) and electroencephalography (EEG). Since the affected area is the motor cortex, neurofeedback should target this area; this suggests that what is read by either MEG or EEG should be the sensorimotor rhythm. This rhythm, also called the μ rhythm, represents a synchronized activity usually between 8 and 12 Hz. **The rhythm is known to desynchronize when movement, passive movement or MI is employed** [15]. In order to relay neurofeedback some processing of the acquired signals needs to be done via a computer. The neurofeedback loop that involves the subject, the MEG/EEG, the computer and the feedback itself may be called a brain computer interface.

1.2. Brain Computer Interface

Many definitions of Brain Computer Interface (BCI) exist. A broad definition might be that the *BCI is a system that decodes the user's intent, via his brain signals, to perform a task*. Millan et al. [16] split the field of BCI into four categories: Communication and Control, Motor Substitution, Entertainment, and Motor Recovery. Most paradigms used in all of the four categories are based on extracting a certain type of information from the electric activity of the brain.

BCIs that fall into the first class of Communication and Control are based on a paradigm that enable, for example, amyotrophic lateral sclerosis (ALS) patients to type using a virtual keyboard or browse the internet. BCIs in the Motor Substitution category usually make use of a similar paradigm aimed at controlling a wheelchair or a telepresence robot. The main purpose of Entertainment BCIs is to offer the user a more immersive game experience. Because this category deals with healthy people it makes use of all common paradigms such as steady state visually evoked potentials (SSVEP), event related desynchronization (ERD), etc. Lastly, the Motor Recovery group of BCI focuses on the rehabilitation of stroke patients. It is based on improving the patient's sensorimotor rhythm and boost plasticity with the aid of MI and feedback. In order to go into the finer details of BCI it is useful to first gain insight on the most common method used for acquiring neural signals – electroencephalography.

1.2.1. Electroencephalography (EEG)

Electroencephalography (EEG) is a noninvasive method that measures the electric activity of the brain. Hans Berger did the first human EEG recording in 1924 [17]. He believed that the EEG waves are directly related to the ongoing cognitive processes. Brainwaves are separated into five categories based on their frequencies:

- Delta δ (0.5-4 Hz)
- Theta θ (4-8 Hz)
- Alpha α (8-13 Hz)
- Beta β (13-30 Hz)
- Gamma γ (30-100+ Hz)

Different brainwaves can be associated to different activities. For example, δ and θ activity is specific to infants and sleeping adults. An increase in α activity can be read in an awake person with his eyes closed. This rise in α activity can be easier seen in the frequency domain over the occipital region. Figure 1.2 shows such an example.

EEG has the advantage of having a high temporal resolution and it is relatively cheap compared to the other functional measurements. On the other hand, the two main disadvantages that EEG holds are poor spatial resolution and its sensitivity to artefacts. The latter are commonly distinguished as technical artefacts or patient related artefacts.

Technical artefacts are usually avoidable with proper experimental design and equipment maintenance. Such artefacts are mainly due to broken wire contacts, gel drying up, gel bridging and not keeping a low electrode/skin impedance (usually it is good to keep this impedance under 5 kΩ). The most common technical artefact is the 50/60 Hz power line hum-noise, due to

capacitive coupling. Fortunately, most frequencies that are investigated with EEG are below 50 Hz and this component may be easily removed by filtering the data. Nevertheless, it is desirable to acquire EEG data as far as possible from power lines.

The two most common patient related artefacts are muscle activity (EMG) and eye blinking. During the EEG recording, the subject might raise an eyebrow, swallow, frown or clench his jaw, etc. All of these lead to EMG contamination of the signal. The blinking artefact is due to the difference in potential between the retina and the cornea that makes the eye behave like a dipole. When one blinks, the eyeball moves upward, resulting in a different projection of the field on the recording electrodes and this can be clearly seen in the raw EEG. Figure 1.3 shows an example of EEG activity with EMG artefacts due to clenching of the jaw and an example of the blinking artefact. With proper subject instruction, the occurrence of the above artefacts may be kept to a minimum.

As one may note both Figure 1.2 and Figure 1.3 contain two noisy channels, Fz and Cz. This is an artefact that has occurred due to either faulty wiring on the cap or problems with the amplifier itself. Such a case is to be avoided.

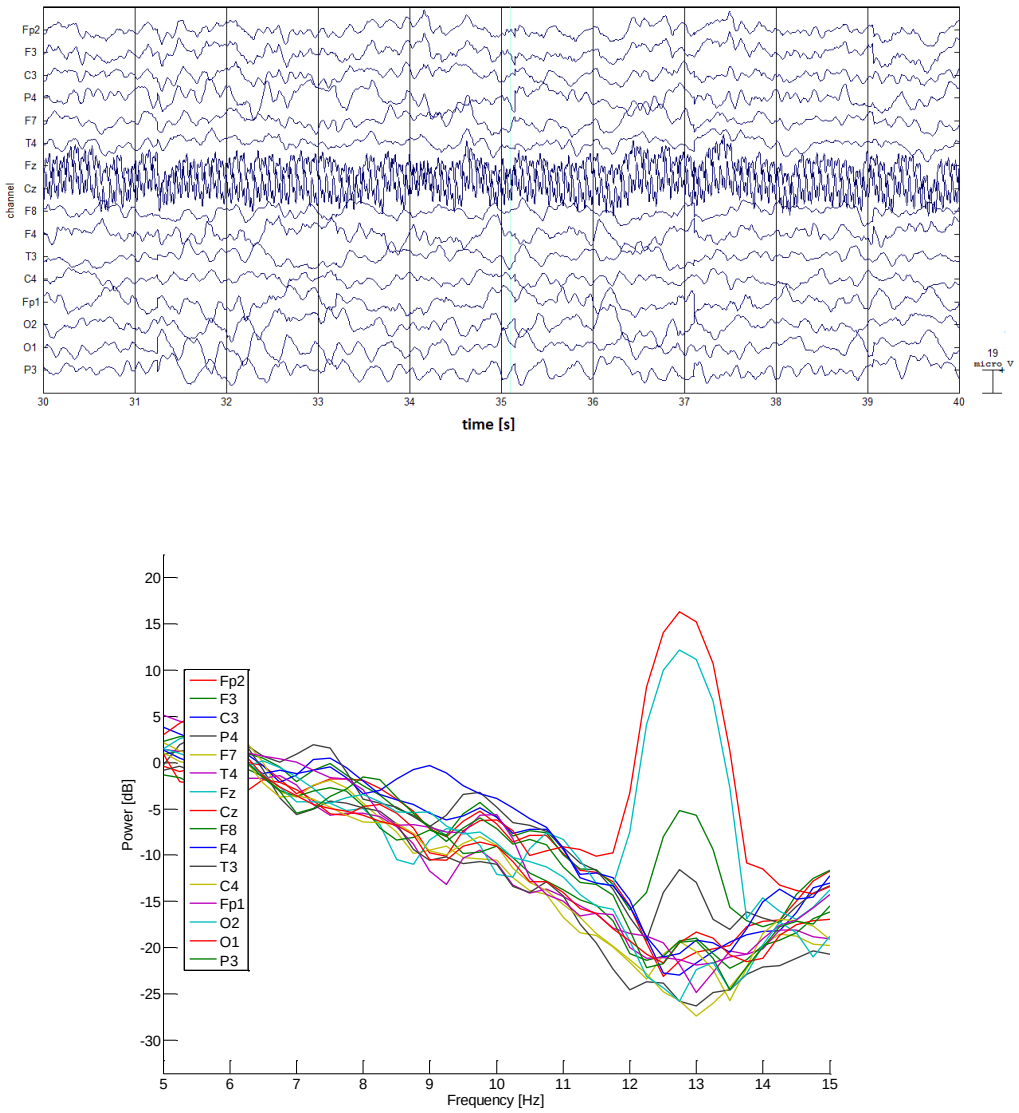


Figure 1.2: Raw EEG -EEG - The top figure represents EEG acquired with a 16-electrode cap, sampled at 256 Hz from a subject with his eyes closed. The α activity cannot be distinguished with the naked eye from the raw data. EEG in the frequency domain - The figure on the bottom shows the frequency domain of the above EEG, for several electrodes. We observe that for four of the electrodes a peak occurs at 12-13 Hz. The two biggest correspond to the O1 (red) and O2 (light blue) electrodes which are placed at the occipital area.

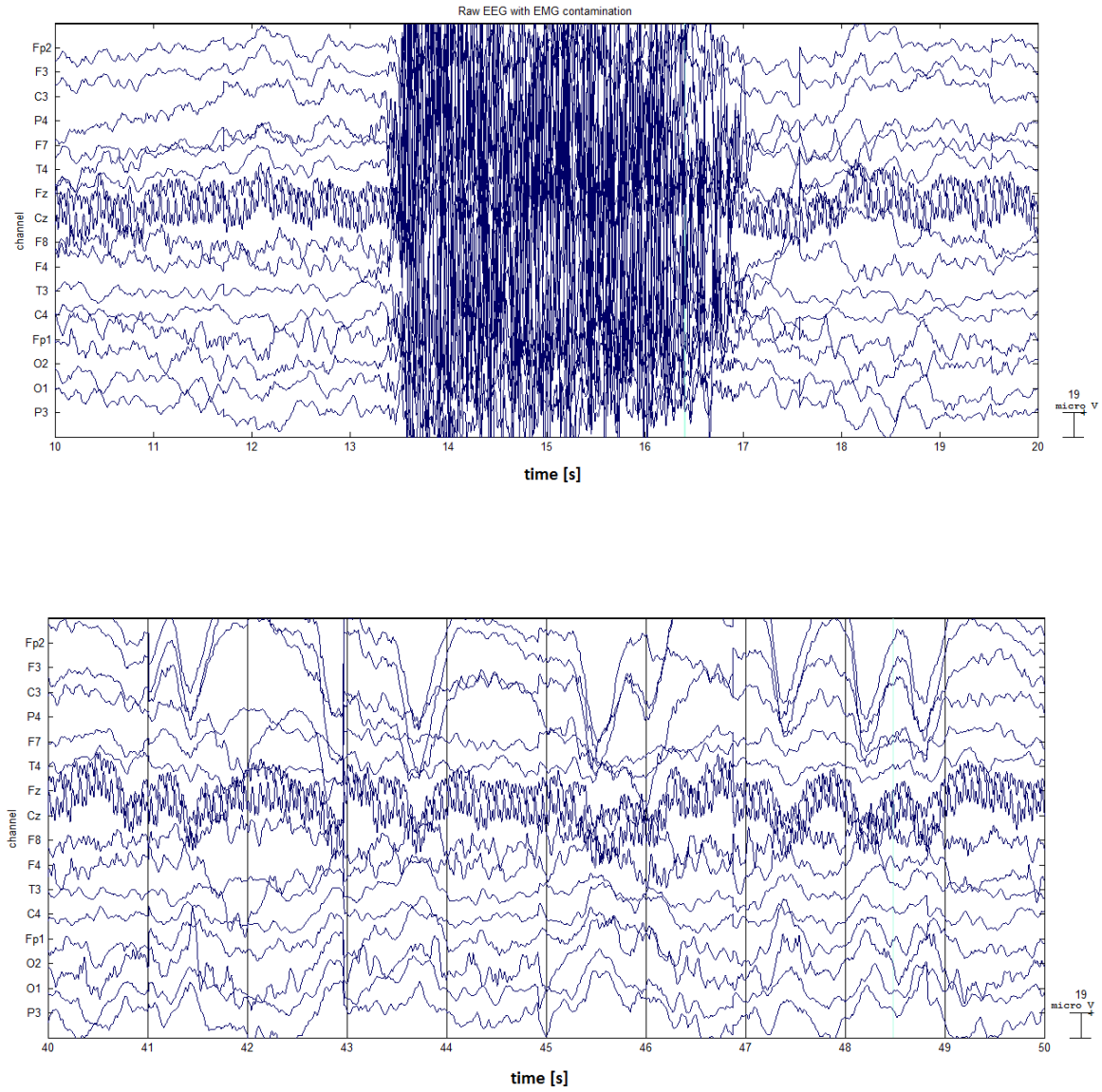


Figure 1.3:EEG with EMG artefacts - When clenching one's jaw the EMG generated has higher amplitude than the EEG thus resulting in noise (top). EEG with blinking artefact - Blinking can be seen in the EEG as a swift change in the polarity of the signal (bottom). Both sets of data were recorded using a 16-electrode cap and sampled at 256 Hz.

1.2.2. BCI terminology

So by the definition used in the beginning we now get to the ingredients that make up a BCI. First, we need to acquire the user's brain signals - this will be referred to as *Signal Acquisition*. Of course, recordings should be as free from noise and artefacts as possible. This requires careful experimental design plus additional filtering of the data. This and other manipulations may be called *Preprocessing*. Thirdly, the intent of the user might not be clearly seen from the preprocessed time series, similar to the eyes closed example provided earlier. As such, *Feature Extraction* is performed to obtain information relevant to decoding the user's intent. Features may fall into different classes, for example, amplitude ranges. The *Feature Classification* part of a BCI gives out signals that are translated via another part into the necessary commands to perform a task. The last three parts can be seen as the components of a bigger block that we will generically call the *Signal Processing* block.

The output of the *Signal Processing* block passes through an *Application Interface*, which translates it into commands and controls for a device, for example a monitor that shows performance information to the patient. This way the user gets feedback so he can learn to modulate his brain patterns to perform the desired action better. Figure 1.4 shows a general layout for the above-described BCI.

In this study, we focus on the fourth category of BCI - Motor Recovery; aspects of which will be discussed in more detail in subsequent sections. Before moving on, it is useful to define some terminology that is commonly used in the BCI community.

Firstly, a *trial* (or *epoch*) is defined as the period during which the subject performs one task, for example movement or relaxation task. A predefined number of *trials* make out a *run*. The number of *trials* that are present in a *run* are defined in the experimental protocol. It may be the case that several *runs* are recorded on the same occasion. In this case, the total number of *runs* recorded on one occasion is called a *session*; if only one *run* is recorded then run is synonymous to *session*.

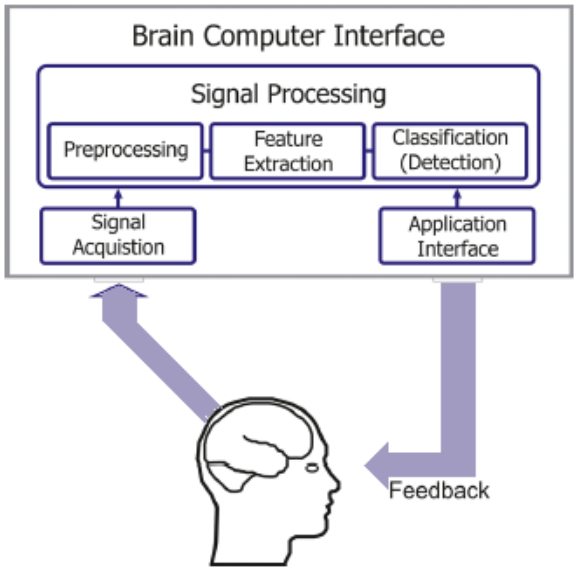


Figure 1.4: General structure of a BCI. Adapted from [47].

1.3. BCI in Motor Recovery, focusing on signal processing, feature extraction, feedback and decision/classification aspects

BCI in Motor Recovery is aimed at aiding the rehabilitation of stroke patients. It is desired that, by the use of the BCI system, plasticity be induced in the affected brain area, so that normal modulation returns [18, 19, 20]. By return of modulation, it is meant that normal event related potentials are produced: the imagination or execution of a movement causes a decrease of EEG amplitudes/ power in certain frequency bands. This is called event related desynchronization (ERD) of the synchronized activity in the μ rhythm (8-12 Hz) and/or in the β rhythm (13-30 Hz) that occurs on the contralateral side of the sensorimotor cortex [21, 22, 23, 24]. ERD is defined as:

$$ERD\ [\%] = \frac{A-R}{R} * 100 \tag{1.1}$$

where A is the power over the frequency band of interest during an movement or MI (*active trial*). R is the power, in the same frequency band, over a time period of relaxation before the beginning of movement or MI (*rest trial*). ERD is usually present on the contralateral side; Figure 1.5 shows the topoplot for performing right motor imagery and the ERD time curve for electrode FC3.

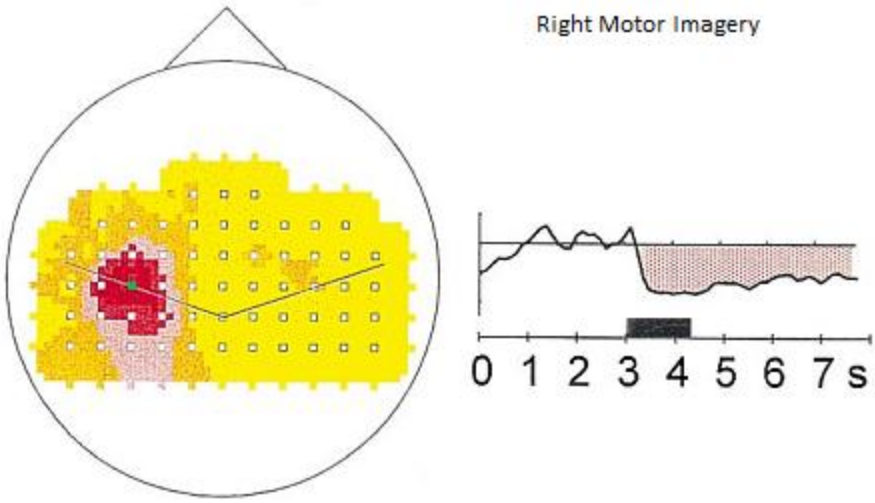


Figure 1.5: Left - topoplot of power distribution during right motor imagery. Activity is present on the contralateral side and concentrated around FC3 electrode (shaded in green). Right - Time curve for FC3; the shaded part represents an ERD. The horizontal bar between 3 and 4.25 seconds represents the cue to start MI. Adapted from [24].

An example will clarify the ERD phenomenon in more detail. For example, the stroke subject in a BCI loop is asked to perform MI of the affected hand. The signal is acquired and then passed through the *Signal Processing* block. If the patient performed MI correctly then it is expected that he will elicit an ERD. This in turn can be translated by the *Application Interface* into a positive feedback, i.e. an encouraging text appears on the screen. In this way the subject knows he performed MI ‘’correctly’’ and needs to keep doing the same thing. If in turn, the system outputs a negative feedback then the subject knows he has to try again, or use a different approach (imagine another movement, for example). In this case, **the goal for the system is to detect and grade, classify the strength of the ERD and give feedback.** This is expected to cause the desired plasticity for speeding up motor recovery.

1.3.1. Approaches to building a BCI for rehabilitation

Literature on BCI and stroke rehabilitation is rather scarce in comparison to the extensive number of articles dealing with MI and healthy subjects. Nevertheless, some studies on the topic exist. In a study by Daly et al. [25], a 43-year old woman who was 10 months after stroke underwent a BCI + FES (functional electric stimulation) protocol. At the beginning of the study, the subject could not voluntarily move her index finger. The FES device was placed so that, when active, it extended the index finger. The FES parameters were pulse width of 255 μs, frequency of 83.3 Hz and the amplitude of the signal was set to a comfort level for the subject. FES was activated with a control signal provided by the BCI. The EEG signal was recorded

using a 58-electrode cap, with a sampling frequency of 250 Hz. Next, the signal was preprocessed with a bandpass filter (0.1-60 Hz).

Power vectors between 5 and 30 Hz were computed for each channel. Each component in the power vector represented the power estimated over a 3 Hz bin. The estimation method was the maximum entropy method. The feature extracted was the frequency band that had the highest explained variance between attempted movement and attempted relaxation over the CP3 electrode. A successful attempt meant to lower the power under a certain threshold for movement and raise it above for relaxation. A threshold was computed for each condition (movement/relaxation) as the feature average on three previously acquired trials. This average was updated at the end of each trial.

The first task was to attempt real movement of the index finger or relax the finger according to a specific cue on a screen. The second part consisted of attempting MI of the index finger or relaxation. One trial for movement (*active trial*) is as follows: a red rectangle appeared on the top of the screen cueing the patient to try to extend the index finger (or perform MI of the same action). If the subject achieved and maintained a signal below the previously identified threshold, then she would be provided with a visual feedback (rectangle changes color from red to green) and FES was triggered. Similarly, for relaxation (*rest trial*), but in the case of a successful trial no FES was applied. Figure 1.6 shows a schematic of the paradigm. In the case of real movement, the subject achieved performances between 82% and 100% and in the case of imaginary movement, the performances ranged from 59% to 97%. For relaxation, the performances were between 65% and 83%. Results show that after three weeks of BCI+FES therapy the subject was able to execute 26 degrees of isolated voluntary movement of the index finger as compared to 0 degrees at the beginning of the study. One strong point that shown in this study is that improvement is possible using a BCI+ FES paradigm. Another important conclusion is that control of the BCI set up can be achieved by using only one electrode.

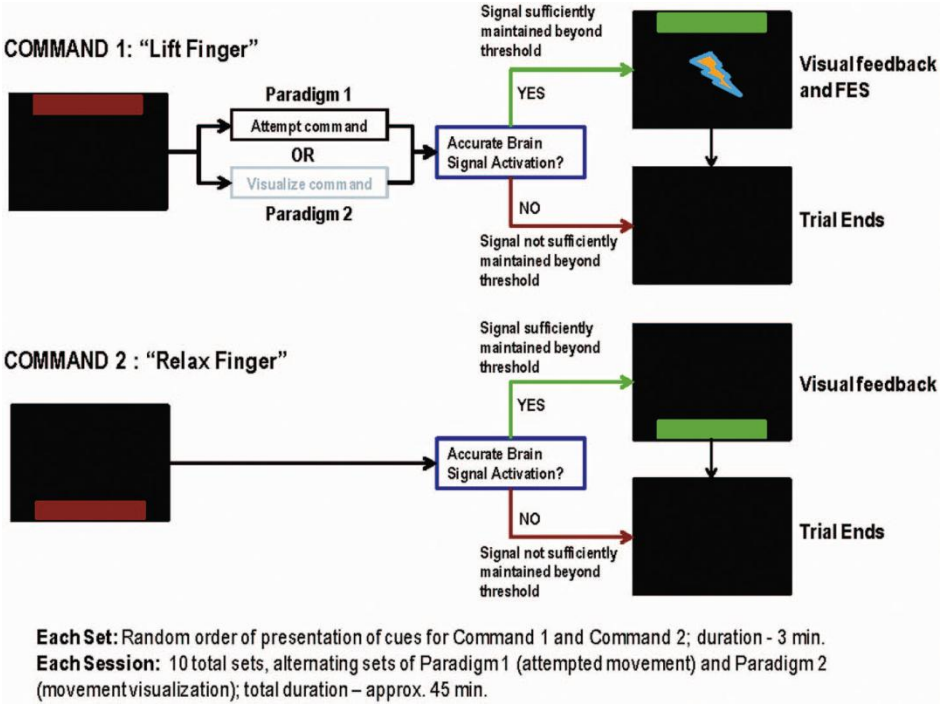


Figure 1.6: BCI+FES paradigm -- The figure on the top represents the task for movement(real or MI) . If the subject achieved and maintained a signal below the threshold, then she would be provided with a visual feedback (rectangle on the top changes color from red to green) and FES was triggered. Otherwise, the screen would turn black. Similarly in the bottom figure, task of relaxation, if the signal could be maintained above the threshold then the rectangle on the bottom would change color. Taken from [25]

In another study, Prasad et al. [26] assessed the feasibility of using solely BCI in upper limb recovery for stroke subjects. Five subjects with ages between 47 and 71 (mean 58.6, SD 8.98), with 15 to 48 months after stroke participated in the study. The paradigm used was the basket paradigm: a ball falls at a constant speed from the top towards the bottom of the screen. At the bottom, there are two "baskets", represented by rectangles. One of them changes its color into green, signaling the fact that it is the target "basket". The subject has to move the ball using real or imagined movement of the left or right hand towards and "into" the target "basket. Figure 1.7 shows a representation for one trial. A trial lasted between 8 and 10 seconds followed by a period between 1 and 3 seconds of rest.

EEG signals were sampled at 500 Hz with a 10-20 system cap using two bipolar channels. The corresponding electrodes were placed 2.5 cm anterior and posterior to the locations of C3 and C4. The signal was then bandpass filtered (0.5 - 30 Hz) and a notch filter was applied on 50 Hz. The proposed features in this study are the powers over the two bipolar channels around C3 and C4 locations for α and β bands. They were estimated from an autoregressive (AR) model with the autocorrelation method. Features were extracted each second and fed into a type-2 fuzzy classifier. The magnitude and sign of the classifiers' output were used as a control signal for the ball's movement to either left or right. It is not stated how often the ball's position is updated according to the control signal. Performances overall subjects for MI ranged from 60 to 75%. The protocol implied 40 trials of real movement followed by 40 trials of imagined movement;

left and right trials being presented in a random order. This was repeated for four runs each session.

Motricity Index (McI), Action Research Arm Test (ARAT) and Grip Strength (GS) were the methods used to assess improvement of the subjects. Only two subjects showed improvement in McI. Out of 5 subjects only 3 could complete the ARAT test and all shoved improvements of 4.0, 6.0 and 10.0 respectively. All 5 subjects showed better dynamometer GS throughout the study. At the end of the study the mean change was 4.4 (20%) when compared to the mean score (22.2) recorded at baseline. The paper concludes that "…*BCI supported MI practice is a feasible rehabilitation protocol combining both PP (*physical practice*) and MI practice of rehabilitation tasks*".

It is worth noting that, in Daly's study, the subject started with very high performances from the first session, 97% for real movement and 83% for MI. These values are unusually high for a naïve BCI subject. The obtained performances over time these, do not indicate any clear trend that may be attributed to plasticity. It is also not clear as how the thresholds were initially computed; i.e. what was the threshold for the first three trials? Furthermore, it is not mentioned whether the threshold given by the last three trials of a session was the initial value for the next session. A clearer and simpler method is desirable.

The power estimation methods used in both studies is known to depend on the order of the autoregressive model (AR). Estimating the optimal order of an AR model depends on the chosen model error criterion (not given in any of these studies), the length of the data used (not given in [25]), and sampling frequency. As such, it might be more favorable to choose a non-parametric method. The methodology would be easier to reproduce and verify by other parties. As previously mentioned, the performance obtained in [25] is overoptimistic and might not extend to other users. The performances obtained in [26] are more realistic. One has to wonder as well what is the optimum number of channels that provides the best classification. In addition, is there a minimum/maximum number of electrodes for which the performance is stable? A natural question that follows is if the information provided by a high number of electrodes can be used somehow to improve the signal-to-noise ratio.

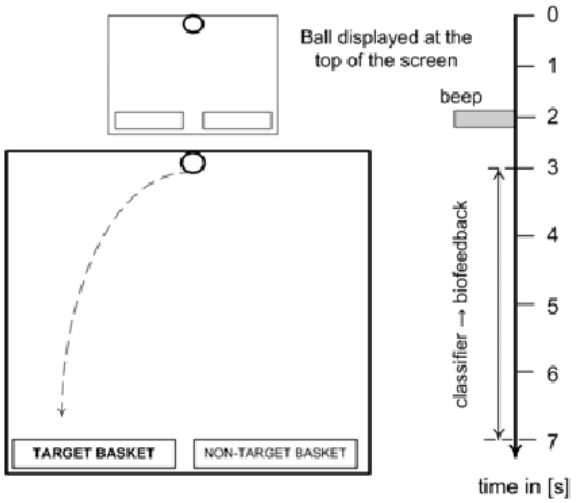


Figure 1.7: The Basket Paradigm - the trial starts with the ball on the top of the screen. After the audio cue, one of the "baskets" on the bottom turns green signaling that it is now the target "basket". At the meantime, the ball starts falling. Now the subject is supposed to perform MI with the hand that is on the same side as the target basket. The user's aim is to get the ball into the basket by actively modulating their EEG. Taken from [26].

1.3.2. Improving classification outcomes in BCI

Commonly there are three techniques that are used in BCI for building spatial filters that in principle boost classification performance: Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Common Spatial Patterns (CSP). All three methods have in common the fact that they operate on the variance of the signals in some aspect in order to remove redundancy and noise [28].

PCA transforms the data using single value decomposition and condenses as much variance as possible into the first extracted components. PCA constructs a spatial filter that forces a maximum amount of variance to be present in the first transformed waveforms. Unfortunately, PCA works on the assumption that the scalp maps are orthogonal. Another downfall of PCA is that useful information for the investigated phenomena might be encoded in the components that will not be taken into account.

ICA separates the EEG data into statistically independent components by using higher order statistics (kurtosis). In other words, ICA is an optimization algorithm that extracts the direction with the least-Gaussian probability density function (PDF), removes the data explained by this variable from the signal, and then iterates. Basically, ICA "rotates" the subspace with the linear mixtures in which the variance of the two axes (mixtures) is equal and the correlation matrix 0,to the original space. PCA and ICA are both unsupervised learning methods meaning that they do not take into account from what class (in the case of EEG - conditions, for example eyes open/closed, or EEG during MI or during rest) the data comes from.

CSP is closely related to PCA but it takes into account EEG signals coming from two different conditions (active state/rest state). After applying CSP, we obtain spatial filters that will maximize the variance for one condition and minimize the variance for the other condition. Given the problem at hand, CSP is a better choice because CSP outperforms ICA in terms of classification performance and because CSP will consider the fact that data will be from two different conditions [28, 29].

A recent study by Ortner et al. [30] propose two paradigms that use CSP on healthy subjects. The algorithm behind CSP will be explained in detail in subsequent sections. This study involves assessment of the proposed algorithm on three healthy users (mean = 28, SD = 1.73). Data was sampled at 256 Hz with 63 channels for two of the subjects and 27 channels for the remaining subject. The data was bandpass filtered (Butterworth, 5^{th} order) between 8 and 30 Hz. After the CSP was applied 4 band powers were chosen corresponding to the first and last 2 newly obtained time series. Band power was computed with the variance method. These four were chosen as features after being normalized and log-transformed.

Linear Discriminant Analysis (LDA) was then used to classify the data. The output of the classifier was used as a control signal for the *Application Interface*. One trial lasted a maximum 8 seconds and started with an audio beep at second 2. Then at second 3 a visual cue appeared, either instructing the user to perform left or right MI. Cue disappeared at 4.25seconds, also accompanied by a beep. The feedback phase started from here and lasted until the end of the trial. A random interval between 0.5 and 1.5 seconds was kept between trials. One session was comprised out of seven runs and each run had 20 trials for left and 20 trials for right.

The approach to build a reliable CSP and LDA was start by recording trials with no feedback. Then, the data from this run was used to build up an initial CSP and LDA. These were used through runs 2 to 5, in which the user was provided with feedback. Then using this four runs new CSP and LDA were built and used for the last two runs of the session. This strategy and the structure of a trial are shown in Figure 1.8. The *Application Interface* could either be comprised of a bar feedback or of virtual reality (VR) feedback. In the case of the bar feedback a bar beginning in the middle of the screen would expand to the left or right depending on the LDA score. In the VR case, the hand movement of a first-person avatar served as feedback. It is not stated what was the time window for feature extraction neither how often the output of the classifier was updated.

Performances were tested on the merged data of runs 6 and 7. The mean performances for the 3.5 to 8 seconds period are given in Table 1.3. For the first two subjects, the performance for 27 channels was computed by discarding channels out of the original 63. This study does not involve stroke patients but it gives an idea on how to achieve good performances needed for feedback. What is not clear in neither of these two studies [26, 30] is what happens, in terms of feedback, if the person tries relaxation, i.e. does not perform MI of any of the hands.

Table 1.3: Performances for the merged data of the sixth and seventh run. The values represent the mean performance obtained in the interval 3.5 – 8 of a trial over all trials. Taken from [30]

	Bar Feedback		VR Feedback	
Subject	27 Channels	64 Channels	27 Channels	64 Channels
S1	87.20%	87.25%	85.20%	80.20%
S2	79.20%	80.10%	75%	80.80%
S3	75%	-	78.20%	-
mean	80.47%	83.68%	79.47%	80.50%

Summarizing, in Daly's study, classification is done by using one threshold for each condition, active/relaxation. Because we are dealing with two conditions, movement and relaxation, only one threshold would suffice. Prasad's study reports only one threshold but uses an unstable (high variance) classifier [31]. The established condition for discriminating between two conditions will be from now on called a separating hyperplane. Stable classifiers such as Linear Discriminant Analysis or Support Vector Machines that are commonly used in the field of BCI should be investigated. Furthermore, these two studies use parametric methods to extract relevant features. These estimation methods are unstable [27]; as such, a non-parametric method is a more viable approach.

In addition, none of the studies studied the influence that the number of channels might have on classification performance. To this end, several sets of electrode configurations and performances obtained should be investigated. The performances obtained using these sets should also be compared to a chosen standard, like the performances obtained after applying CSP.

The main shortcoming in [25] and especially in [26] is that there is no healthy control group to build parameters. Since the main objective is to make stroke patients to regain normal modulation, it is natural to develop a BCI system that has at least some of its parameters tuned on normal subjects. With all these in mind, we proceed to define the objective of the present work.

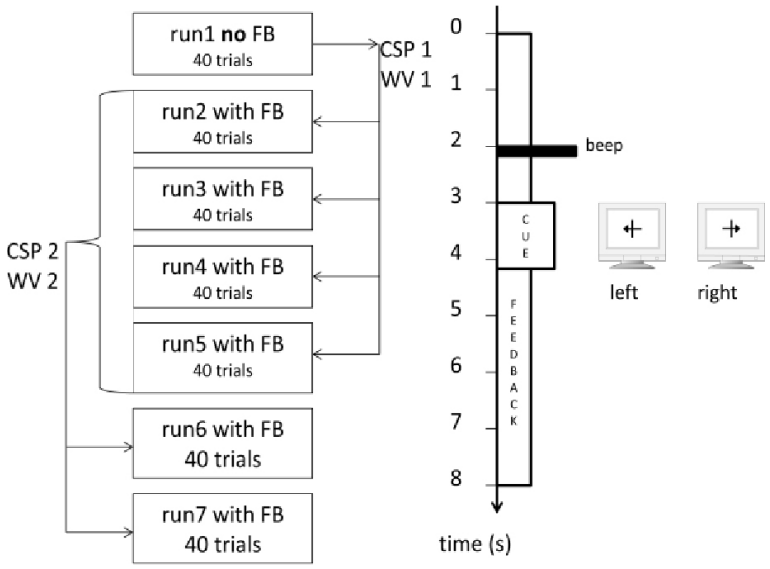


Figure 1.8: Structure of a session– (left) The first CSP and LDA (WV1) are build up from the first no-feedback run. Next, they are used in runs 2 through 5. The second CSP and LDA are build up on these merged runs and used in the last two runs. Trial structure – (right) at second 2 a beep is given to capture the attention of the user. A visual cue starting at second 3 announces the user about the imagery to perform. At second 4.25, the cue ends and the feedback stage begins. Taken from [30].

1.4. Objective and Research Questions

The main objective is to devise a system that can distinguish between resting state and an active state (real movement or MI). As formula (1.1) indicates, the ERD is a ratio given by two consecutive trials, relaxation and MI or execution so we cannot use it to build up such a system. We want our system to classify every newly acquired trial (single trial classification). In other words, we are going to compute a power threshold that will allow us to label new trials.

To this end, we investigate what feature, between broad (8-30 Hz), α (8-13 Hz), β (13-30 Hz) and *user defined* band power, will provide better discrimination between the two classes. We are going to test the first three features to see if we will obtain similar performances as the user defined band. If performances are close to each other it will mean that, in an online setting the specific user defined band does not need to be computed in order to relay appropriate feedback.

Given the fact that the eventual aim is single trial classification, we examine two classification methods that will automatically detect the user's condition (relaxing or active) based on previous labeled trials. The investigated classifiers are Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). We have chosen these two classifiers because they are stable, meaning that their decision is dependent on the input data and not on initialization parameters, and because they have a good ability of generalizing the data. Another question we want to answer is how many trials the classifier needs in order to achieve reasonable performance. The answer to

this question will help to keep the time from when the system is set up to the time it can be used at a minimum.

We would also like to see the minimum number of electrodes required for the system to work properly. We will test out different numbers of electrodes. In addition, we will compare the performances, obtained by using the power over all channels to the performances obtained after applying CSP. This will tell us if the introduction of CSP into the online pipeline is worth the extra computational time that it implies. We will also see what is the minimum number of electrodes needed to give adequate feedback. This information will help reduce the time needed to set up the system.

Another question we want to answer is what is the minimum duration of a trial that can be used while still maintaining high performance? We will test if the results found for using the full trials (1 trial = 8 seconds) are similar when taking only 75%, 50%, and 25% of trial length.

All of these questions will be answered by doing offline analysis on EEG data coming from healthy subjects and stroke subjects. The data from healthy subjects will serve to find out the parameters for LDA, SVM and CSP. The other questions will be answered by using stroke data.

2. Subjects and Methods

2.1. Subjects

Data was recorded from ten acute hemispheric stroke subjects, 5 females and 5 males (mean age = 64.9, SD = 13.14, nine left handed) with conditions ranging from mild to severe when the study started (T0). Sessions were recorded at 2 weeks (T0), 1 month (T1), 2 months (T2) and 4 months (T3) after stroke. There were 2 withdrawals after T0 and one withdrawal after T2; these subjects were not included in the study. We have also excluded subject S10 from the study because sufficient trials could not be recorded during the sessions; also, the first session of S09 was not analyzed for the same reason. The stroke subjects were recruited from the stroke unit of the Medisch Spectrum Twente (MST) hospital within 7 – 14 days after stroke onset. The local ethical committee of the MST approved the study. Magnetic resonance or computed tomography imaging was performed in every stroke patient to confirm the diagnosis and detect the infarct location (see Figure 2.1). Table 2.1 shows the demographics, the clinical condition at T0 and the results of a Fugl-Meyer test for all sessions, of the stroke subjects.

Table 2.1:Demographics, clinical condition at T0, side of the lesion and Fugl-Meyer tests for all stroke subjects

Subjects	*Sex/Age*	*Affected hand*	*Site/type of lesion*	*T0*	*T1*	*T2*	*T3*	*Clinical Condition at T0*
				up-FM	*up-FM*	*up-FM*	*up-FM*	
S01	F/58	Left	R- subcor	54	66	66	66	Mild
S03	F/51	Left	R- subcor	50	58	62	64	Mild
S04	M/68	Right	L-subcor	48	60	66	66	Moderate
S08	M/58	Left	R- subcor	52	64	66	66	Mild
S05	M/84	Left	R- subcor	4	-	-	-	Severe
S02	F/56	Left	R-Cor	65	65	66	66	Mild
S06	F/81	Left	R-Cor	49	58	63	65	Moderate
S09	F/62	Left	R-Cor	24	49	57	64	Severe
S10	F/49	Left	R-Cor	4	4	7	-	Severe
S07	F/82	Left	R-cor	41	-	-	-	Moderate

Data was also recorded from 11 healthy subjects were also recruited, 9 females and 2 males. All participants signed a consent form and received a small gift at the end of the experiment.

Data from 5 healthy subjects were acquired by the author (all male). The mean age of the healthy group is 47 with SD = 5.85, out of whom fifteen were right handed. The healthy group will henceforth be called the control group.. The data from the stroke subjects and eleven of the healthy subjects were already available at the beginning of this project.

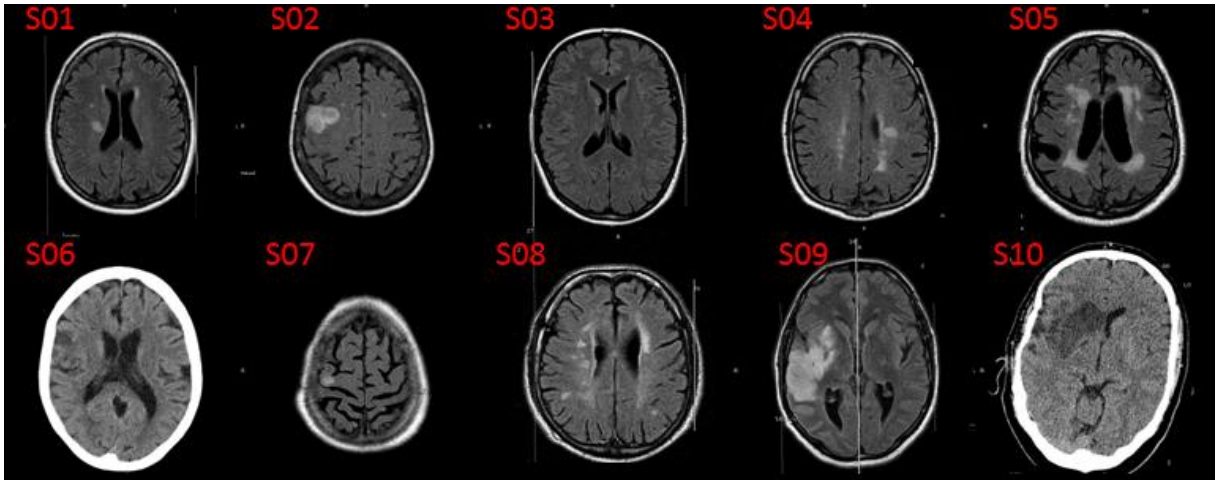


Figure 2.1: T1-weighted MRI (S01-S05 and S07-S09) or CT (S06 and S10) images at the level of maximum infarct volume for each stroke subject.S01, S03, S04 S05 and S08 show subcortical infarcts; other scans show cortical infarcts.

2.2. Methods

2.2.1. Paradigm

The protocol used to acquire the data from the 5 healthy subjects, mentioned earlier, will be described in the this section[1]. The subject signed a consent form and was instructed on the tasks that had to be performed. One session was comprised of four runs, one run in which the subjects had to perform real movement and relaxation and one run with MI and relaxation, and again real movement followed by MI. During the session, the subject was seated in a comfortable armchair placed 1 meter away from a 21-inch LCD monitor.

Before starting the actual session, a calibration run was performed in order to choose the optimal baseline movie (movie that was presented during the resting task). There were three such baseline movies: a static grid, two balls moving, and flowers. The one that induced the most suppressive rhythm for the subject was selected and used throughout the rest of the session. In order to select the proper baseline movie the subject was asked to perform real movement of his dominant hand during the active movie (the subject was only presented with active movies of his dominant hand) and relax when one of the baseline movies was presented. For a detailed explanation of how the baseline movie was selected please refer to Appendix A. After the baseline movie was chosen, the actual session began.

[1] The same protocol was also used for the data that was made available at the beginning of this project

One run consisted of 32 trials – 16 rest and 8 left/8 right. A total of 32 active trials and 32 rest trials were acquired per session for both real and imaginary movement. We shall describe a sequence of four trials to explain the flow of one run better (see Figure 2.2). In the first trial a ~10-second movie, referred to as baseline movie, was presented. During this movie, the subject was asked to relax and not perform any kind of motor action. The baseline movie was followed by an active movie of ~10 seconds. There was no pause between movies. The active movie consisted of five repetitions of an either right or left hand opening and closing. During an active movie, the subject was asked to imitate/imagine the movement in synchrony with the action presented on the screen. A baseline movie followed; after the baseline again an active movie and so on. The succession of left/right active movies was done randomly. This paradigm was used for both the stroke group and the control group. In this study, we will only use the data that was acquired during MI.

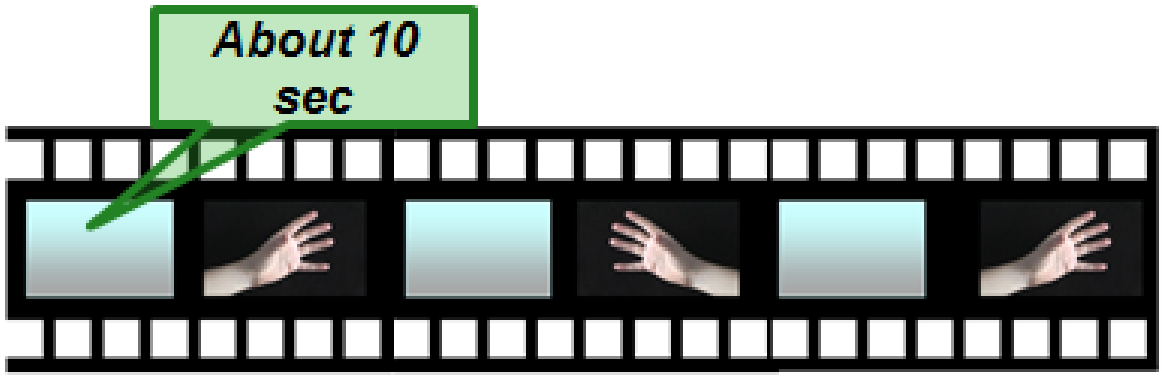


Figure 2.2: Experimental paradigm used for data acquisition. The succession of active movies is presented randomly.

2.2.2. Signal acquisition

Data was recorded using a 10-20 system, Wave Guard 64 electrode cap produced by ANT with Ag/AgCl electrodes and active shielding on the electrode wires in order to reduce the capacitive coupling with the power lines. An electrode placed on the tip of the nose served as ground, and the left mastoid was used as a common reference. Four electrodes were discarded because the connections to the amplifier were broken (P7, P8, TP7, and TP8). The amplifier used was a TMSI system with the sampling frequency of the amplifier set at 5000 Hz, and hardware filter cut-off frequency of 1350 Hz. In the second stage of the amplifier common mode rejection is performed in order to minimize the influence of the power line hum [32]. The recording software

used was ASA-Lab developed by ANT. Electrode impedance was kept under 5 kΩ throughout the whole experiment. Recordings were performed in a shielded room.

For analysis we discarded 16 electrodes(Fp1, Fpz, Fp2, F7, F3, Fz, F4, F8, AF7, AF3, AF4, AF8, F5, F1, F2, F6) from the frontal region in order to avoid EMG and blinking contamination of the data. Upon visual inspection it was noticed that in most subjects the FC3 electrode showed abnormal behaviour so it was also discarded.This left a total of 43 channels availabe. We believe that the abnormal behaviour was probably due to faulty wiring in the cap. All of the parts that will follow have been implemented using MATLAB version 7.14 64-bit on a PC with 6 GB RAM and a core i5 M430 CPU at 2.27 GHz.

2.2.3. Preprocessing

The raw data was band pass filtered (8-30 Hz, Butterworth 6^{th} order). Next the signal was downsampled to 500 Hz. The signal was common average referrenced in order to get a better signal to noise ratio. The following procedure was to make baseline and active trials the same length. By discarding ~1 second from the beginning and the end of a trial we obtained 8 second trials. Our decision considered the absence of breaks between trials. Here we also separate the 64 trials that we get in one sessions in two sets –left/relaxation and right/relaxation. The sets were built by taking an active trial and the relaxation trial that occurred before. Out of the 8 second trials we also computed trials of 6, 4, and 2 seconds. The new trial lengths were computed by taking the first seconds from an 8 second trial. No artefact rejection was performed in order to keep the analysis as close as possible to an online situation.

2.2.4. Feature extraction

Four power features are extracted from the preprocessed data – broadband, α, β and user defined band. We already can extract the broadband power feature but for α, β and user band we obtain three new datasets by filtering the broadband one. The power was computed by taking the variance of the signal in the selected band for each channel. The power vector was computed as:

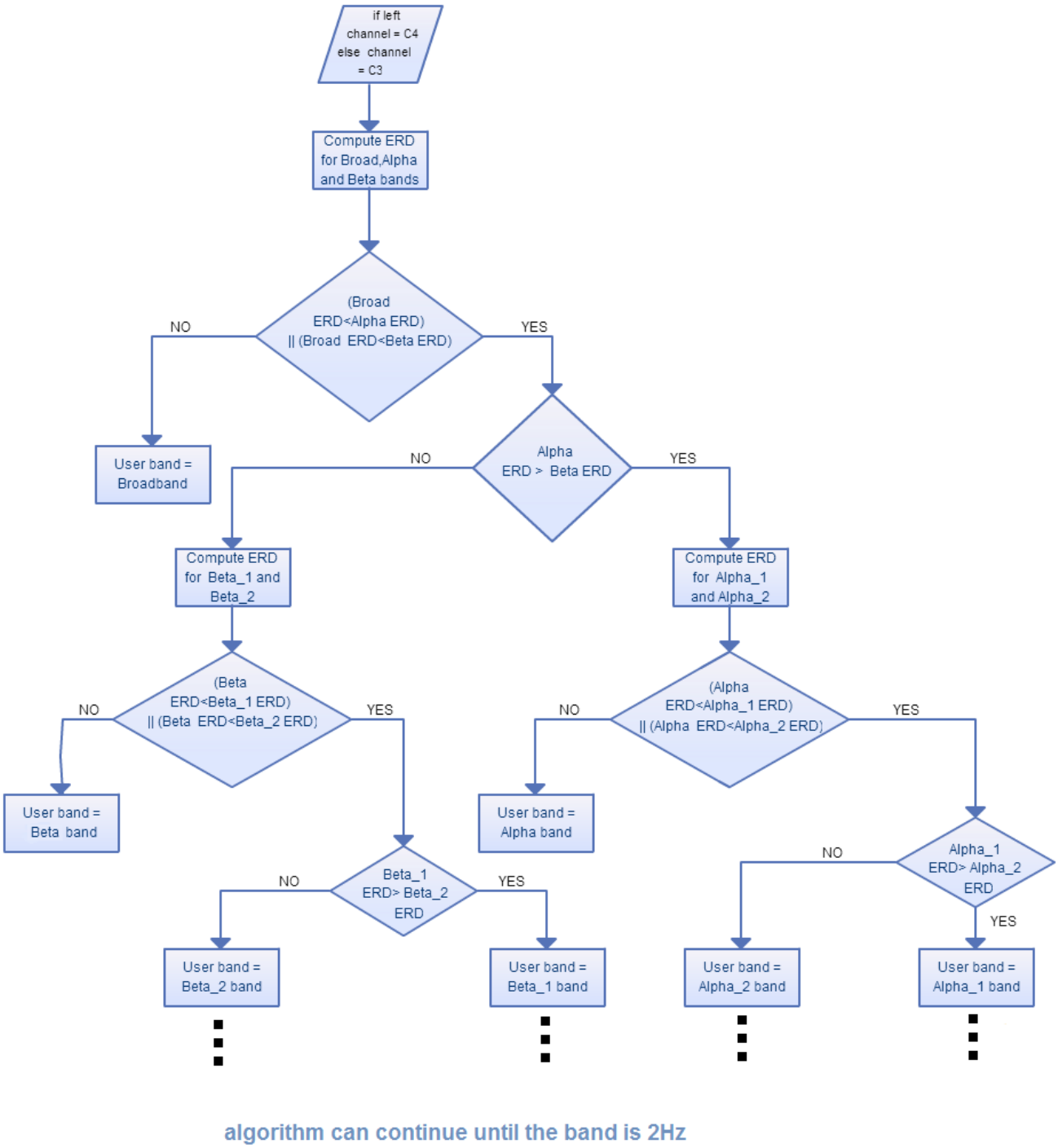
$$\sigma_s^2 \triangleq \frac{1}{M-1}\sum_{m=0}^{M-1} s_m^2 - \frac{M}{M-1}\mu_s^2 \qquad (2.1)$$

where s_m is the m^{th} sample of the signal s.

In the case of the user band a divide et impera search algorihm was employed. The degree of ERD was taken as the objective criterion for this search; C3 (right movement) and C4 (left

movement) electrodes were chosen because of their physiological relevance. At the beginning of the analysis, the ERD for broad, α and β bands were compared to see which is highest. In case the broadband is the highest then the algorithm stops and takes the user band as broadband. If the ERD for β-band is higher than the one for α-band, the interval is split in two (13-21 Hz and 21-30 Hz) and the ERD is computed and compared for this two intervals. If, for example, the ERD is higher for the 13-21 Hz band then this interval is again split in two (13-17 Hz and 17-21 Hz) and ERD is computed. The algorithm can continue until the bandwidth is as small as 2 Hz; similarly for α-band. The user band will be defined as the frequency interval thatwas found to have the highest ERD.Figure 2.3 shows a diagram of how the algorithm works. This algorithm is only used to compute the user specific frequency and will not be used in future steps, such as classification.

In other words what we do by computing the band power is extracting relevant information from the time series. We now have feature vectors that will be used to describe relaxation and movement of one hand (real or MI) conditions/states. From here on these two conditions/states will be referred to as classes. Also the space that is described by the power over a number of channels will be called feature space. In this space one trial will be represented by a n-dimensional point, the coordinates for which are given by the feature vector. In our case n represents the number of channels used, and the components in the feature vector are the bands' power over the channels.



algorithm can continue until the band is 2Hz

Figure 2.3: Flow chart for detecting the frequency band, which manifests the highest ERD, and assign the user specific frequency band. The algorithm starts with computing the ERD on C3 or C4 for broad band, α band and β band. If the highest ERD is found in the broadband then this becomes the user band. If the highest ERD is found on β band then this band is then split in two and ERD is computed for both resulting bands (β1 and β2). The ERDs are compared and if βband is found to be the highest then the user band will be β band. If the ERD is found to be highest on either β1 or β2 then this band is again split in two parts and the algorithm continues until it finds the highest ERD. The algorithm may continue until the bandwidth is 2 Hz. Similarly for α band.

2.2.4.1. Common Spatial Patterns

It is known that the most relevant information for the task at hand is supposed to arise from the sensorimotor area. This means that the electrodes placed further away from this area might contain unimportant information, i.e. worsen the discrimination power of the feature space. In turn, the high number of electrodes brings two advantages. First correlation between adjacent electrodes can be used to remove noise and second, weights can be assigned to the electrodes according to their relevance in distinguishing between movement (real or MI) and relaxation in order to create virtual channels [33, 34]. The way to accomplish the above is by applying the Common Spatial Patterns (CSP) algorithm to the data [35, 36, 37].

For a formal description of the algorithm, please refer to Appendix B. We consider S as being a matrix representing the recorded signal on N channels over a period defined as T:

$$S = \begin{pmatrix} S_{1,1} & \cdots & S_{1,T} \\ \vdots & \ddots & \vdots \\ S_{N,1} & \cdots & S_{N,T} \end{pmatrix}. \tag{2.2}$$

We take:

$$Z = F'S, \tag{2.3}$$

where F is the CSP matrix. Z has the property that its rows are uncorrelated. The columns of the CSP matrix represent spatial filters. When we apply the CSP matrix to the data, we will get new channels, virtual channels that are linear combinations of the CSP columns. For example, the first virtual channel represents the linear combination of the initial channels given by the first column of the CSP matrix.

In order to describe class **a** (active) and class **b** (rest) we need only to take the first and last virtual channels because the ones in the middle will have an almost equal variance coming from both classes. Figure 2.4 shows the first and last virtual channel obtained after applying the CSP transformation to an EEG with two classes. Now the power over the selected virtual channels is computed and a new, more compact, feature vector is obtained. Note that the first channel exhibits higher variance during the active class and lower variance during the rest class, and vice versa for the last channel.

By applying the CSP not only have we obtained decorrelated channels but also we may now reduce the feature space by choosing only virtual channels that hold the most relevant information. The only questions that remain now are how many trials to use for estimating the covariance matrices and how many filters (virtual channels) to choose in order to discriminate optimally between classes.

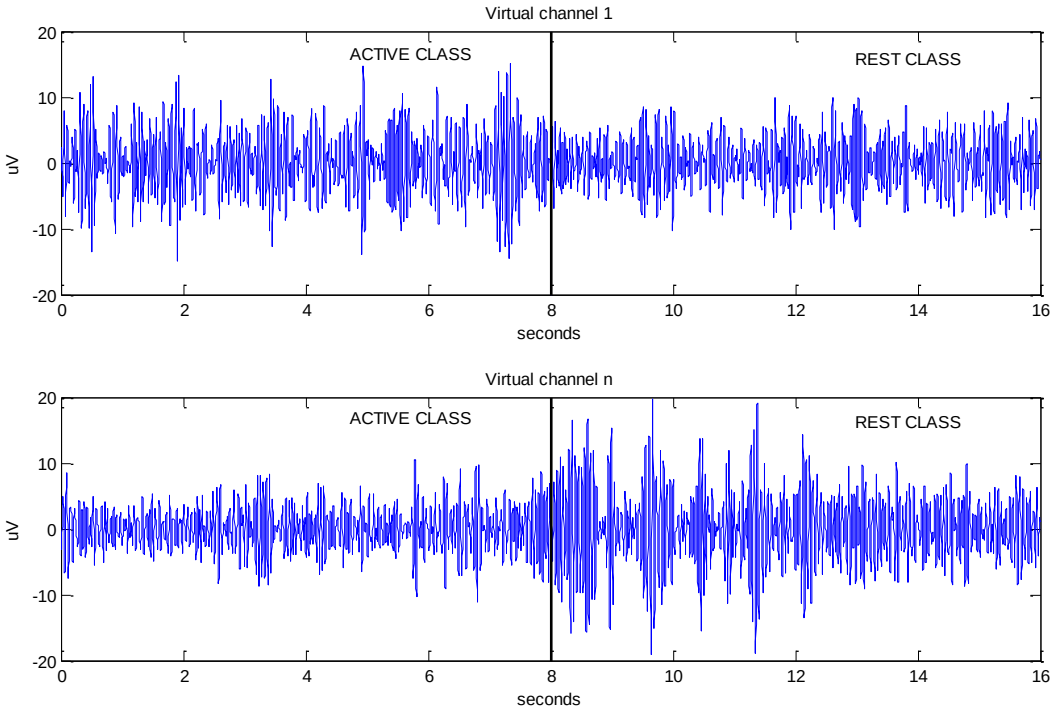


Figure 2.4 Virtual channels obtained after applying the first and last spatial filters given by the CSP matrix to the data. The image contains data coming from an active class followed by a rest class. Note that the first channel exhibits higher variance during the active class and lower variance during the rest class, and vice versa for the last channel.

2.2.5. Classification

As previously mentioned, one instance of a class will be represented in the feature space by a point. Only having one point in the feature space will give us no information to which class it belongs. In order to make a distinction between two classes we need more points (at least one more that has different coordinates than the first one). For example, Figure 2.5 shows how instances belonging to the active class and the baseline class look like in the space described by channels C3 and C4.

Let us assume we have a random number of points belonging to each class; if a new point appears in the feature space to which class does it belong? Given the points that are already present in the feature space and assuming we know from which class they come we can build up rules that will describe in some manner the two classes. Now, based on the rules, we can say to which class the new datapoint belongs. Making up rules from past examples in order to discriminate between classes may be called *classification*. In other words, learning by example

means that with a given set of feature vectors X and a set of labels L that identify each instance of set X it is possible to label a new feature vector as belonging to one class or the other. Most of the classification problems in MI-BCI are nonlinear (as can be seen from Figure 2.5, but also keep in mind that the actual feature space has a higher dimensionality). At first glance, the problem is easily solved by using a nonlinear classifier such as Neural Networks or k-nearest neighbor. The main issues with nonlinear classifiers are instability and the tendency of overfitting the data.

Instability means that given the same dataset, the separating hyperplane depends on some initialization parameters and will therefore not always be the same. Overfitting happens when data is poorly generalized and new datapoints have a higher chance of being misclassified [31, 37]. Linear classifiers have less free parameters to tune and are less prone to overfitting; also, the found separation hyperplane is unique and depends mainly on the input data. The two most commonly encountered classification methods in BCI literature are Linear Discriminant Analysis and Support Vector Machines.

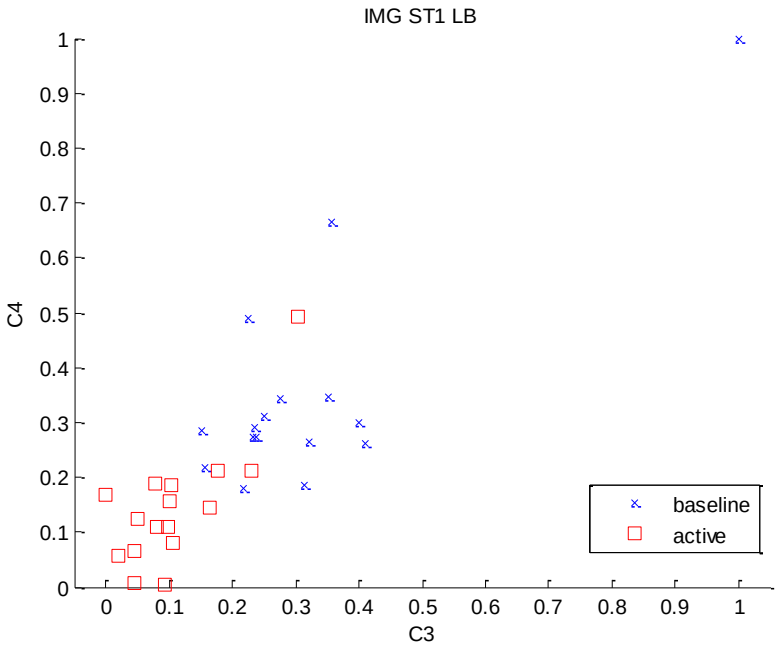


Figure 2.5: Instances belonging to both active and baseline/rest classes represented in the space described by C3 and C4. These instances come from stoke subject 1; the active class is represented by the broadband power for left hand MI.

2.2.5.1. Linear Discriminant Analysis

LDA uses information about the distribution of the already existing datapoints (mean and variance) of the classes to make a decision regarding new datapoints; LDA belongs to the family of parametric classifiers. For a formal description of the algorithm, please refer to Appendix D. Because of the manner in which, LDA builds its decision rule it is very sensitive to outliers – datapoints that have "abnormal" values for one class. The decision rule in our case is a power threshold. Other disadvantages that LDA holds are the assumptions that it makes. First, it assumes that the data is normally distributed and second, that all classes have identical covariance matrices. It is known that MI data in not normally distributed, nevertheless if the feature vectors of the two classes are well separated LDA may perform reasonably [31]. Because of this, after taking the variance of the signals we apply a log transform in order to force the data to obey a Gaussian distribution.

2.2.5.2. Support Vector Machines

A classifier that does not suffer from the same shortcomings as LDA is Support Vector Machines (SVM) [38, 39, 40, 41]. SVM is a nonparametric classifier, meaning it does not take into account the distribution of the data. This classification method builds its decision rule by using datapoints located at the outskirts of the class towards the other class. These special points are called *support vectors*. Even though SVM is a linear classifier, it can be used for classifying non-linearly separable datasets by using kernels. For a formal description of the algorithm, please refer to Appendix D.

The most used kernel in BCI, and the one we will use in this study, is the Gaussian kernel (radial basis function – RBF) [48]

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \gamma > 0 \quad (2.4)$$

In our study, we implement the SVM with the aid of the libSVM toolbox for MATLAB.

2.2.5.3. Performance evaluation

Summarizing, the classifiers will compute power thresholds that will discriminate between an active state and a resting state. In other words if a patient is not performing MI during an active movie he will receive feedback accordingly. For example a text saying "Try again!". Given the classifiers, how can we tell which one is better? The classifier is trained on a portion of this set called training set. After training, the classifier is tested on the remaining part of the initial dataset called test set. In our case we are dealing with two classes (active/rest) so we can say that one is the positive class and the other the negative class. Given this we can evaluate one classifier by counting the number of true positives, true negatives, false positives (true label is negative but classified as positive) and false negatives (true label is negative but classified as positive). In our case the positive class is the active class and the negative class is the rest class. Since we perform offline analysis, the trials are already labeled so we can easily evaluate the training set. Usually in machine learning, the correct classifications and the misclassifications are represented in a confusion matrix (Table 2.2).

Table 2.2: Confusion matrix

	Positive class	Negative class
Predicted Positive Class	True Positive	False Negative
Predicted Negative Class	False Positive	True Negative

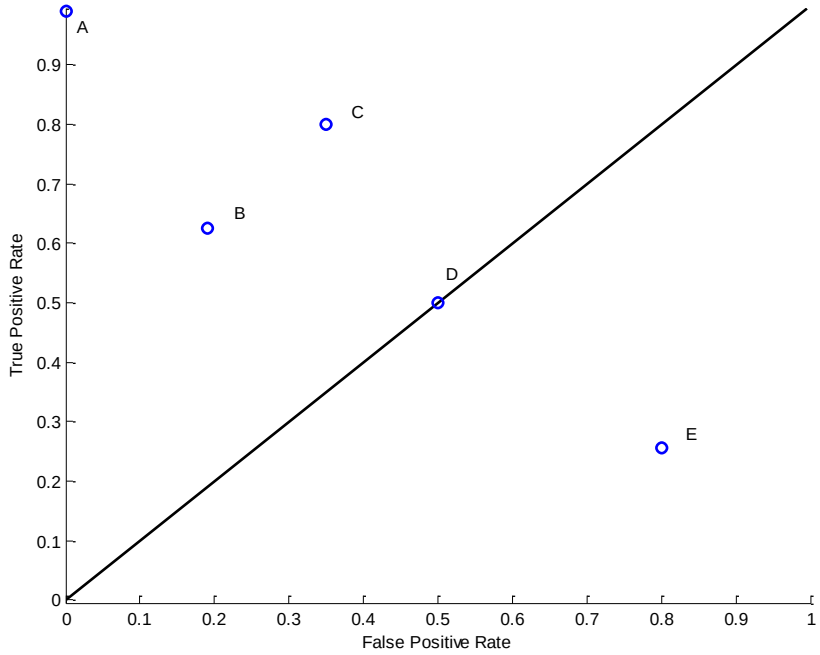


Figure 2.6: Points in the ROC space – A is the point of perfect classification, B and C are equivalent in a sense that both are good operating points but whether to choose one of the points depends on the application. Compared to C,B makes a positive classification harder, meaning that a new instance has to be provide strong evidence that is from the positive class. Point D is equivalent to random guessing and point E is an undesirable operating point.

Usually in the case of judging a classifier by the number of classifications and misclassifications, *Accuracy* is used:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2.16}$$

Several other important measures can be derived from the confusion matrix, such as:

- *True Positive Rate* - $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ (also called hit rate, recall or sensitivity) (2.17)
- *False Positive Rate* - $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ (also called false alarm rate or selectivity/specificity) (2.18)

These two measures are used as the x-axis (FPR) and y-axis (TPR) of the Receiver Operating Characteristic Space. There are specific points in this space that make the understanding of what this space represents easier. For example the point (0,0) means that our classifier never predicts a positive class, meaning that the classifier never commits false positive errors and never gets any

true positives either. Perfect classification is described by the point (0,1). Figure 2.6 shows an example with multiple such points and their interpretation. The line represented by TPR = FPR is the line of no-discrimination and it is equivalent to random guessing. In other words, graphs in this space represent a tradeoff between benefits (true positives) and costs (false positives). By varying this threshold, we get the ROC curve for a specific classifier on a specific dataset. In order to compare multiple classifiers we would like to evaluate their quality using a single scalar value. This value is given by the area under the curve (AUC), value that represents the probability that a randomly chosen positive example is correctly ranked with a higher probability than a randomly chosen negative example. In our case, the AUC indicates what are the chances that trial coming from an active class is correctly classified or, put another way, how good is the power threshold given by the classifier.

2.2.6. Practical Implementation

Now that we have described all the major building blocks that will be used in our study, we proceed on to showing how they connect and how they will help us answer our research questions, or in other words find optimal parameters.

As a reminder the parameters are: four power features (broadband, α band, β band, and user band), training to testing ratios for the two classifiers (LDA/SVM), the number of trials taken to train the CSP, the number of virtual channels after applying CSP, the number of channels, and the trial length (8, 6, 4, and 2 seconds). We start by considering the datasets consisting of data coming from healthy and stroke patients as instances describing the phenomena we are trying to model. It is now useful to consider what we are trying to model – a separating hyperplane, or power threshold, between active and rest conditions/classes. With respect to this, we will divide the aforementioned parameters into three types: independent, primary, and secondary parameters.

We consider the secondary parameters to be the training to testing ratio for the classifiers, number of trials for CSP training, and number of CSP spatial filters (virtual channels). They are dependent on the type of the input data. The features and number of channels are considered as primary parameters and the trial length as an independent parameter. We consider trial length to be the independent parameter of our system whereas the primary parameters depend on the independent parameter and the secondary parameters depend on the primary parameters. This means that we may compare the independent parameters only after we have fixed values for all the other parameters.

It might be that the optimal secondary parameters are different for all four power features. In order to find the dataset that will best model the data we need an objective criterion to compare between them. We have chosen this to be the AUC. In order for the AUC to be a valid criterion,

it needs to be related to the power of the sensorimotor rhythm. We assume that in the case of healthy patients, whom we assume to have a normal sensorimotor rhythm, this is so. AUC is a measure of the performance of the classifier; the performance of the classifier is based on the "quality" of the feature space; the feature space is given by the time signal and the time signal is supposed reflect the presence or lack of desynchronization.

2.2.6.1. First Stage – Healthy Subjects

Now we can implement a first stage, to find out what the best secondary parameters are based on the healthy dataset. In this case, we will vary the training/testing ratio from 20%/80% to 80%/20% for the LDA and SVM. We chose the optimal split according to the AUC. Using the same dataset we will find out what is the minimum number of trials to train the CSP matrix and the minimum number of virtual channels needed to achieve proper discrimination. Again, the criterion we used here is the AUC. We did this for every feature and trial length.

It must be said that the electrode configurations used in this study were (see Figure 2.7): 20 (config. A), 14 (config. B), 10 (config. C), 5 (config. D), and 2 (config. E). Config. A and B should have had 21 and 15 electrodes but this could not be done because FC3 was taken out of the analysis due to the aforementioned technical problems. **This stage is carried out using 20 channels**.

The schematic of this stage can be seen in Figure 2.8. First, we select the hand, and then we fix the number of channels to 20. Afterwards, the broadband data that was filtered in the preprocessing step is filtered into α-band, β-band, and user defined band. The algorithm used for the user defined band is implemented as in Figure 2.3, but it is computed using different number of trials corresponding to different training/ testing ratios. We have done this to avoid bias when computing the secondary parameters, and because it simulates an online calibration pipeline. This means we compute the ERD using 4, 6, 10, 12, 16, 18, 22 and 24 trials corresponding to 10% to 80% training, depending on the pathway. This step is not shown explicitly in the schematic of this stage.

We now have four distinct EEG time series for trial length of 8 seconds, one for each band. Out of this dataset, we compute individual time series for 6, 4, and 2 seconds. This is where the first stage breaks into two distinct pathways. In the first pathway, we extract the power for each time series. After, we split the power datasets into training to testing ratios ranging from 20%/80% to 80%/20%; then we model the threshold between active and rest classes with LDA and SVM for each ratio, feature and trial length. We repeat classification 10 times by randomly selecting data for training and testing; this step is not shown explicitly in the schematic. We repeat this procedure for each healthy subject. Then then we compute the AUC for every subject and dataset,

and store the information in a buffer. In order to find the optimal secondary parameters we take the AUC grand average over all subjects for these secondary parameters.

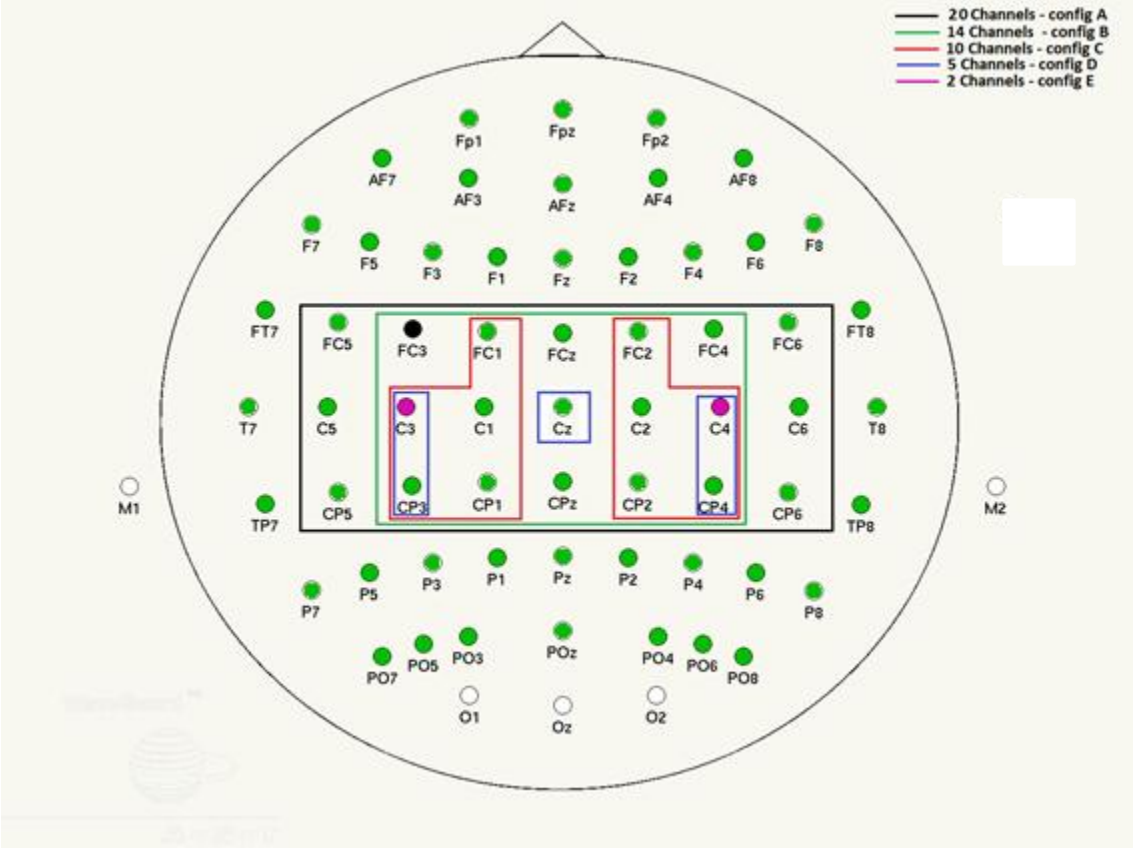


Figure 2.7 Electrode configuration sets– we start by using 20 channels (black rectangle –config. A), 14 channels (green rectangle – config. B), 10 channels (red polygon – config. C), 5 channels (blue rectangles – config. D) , and 2 channels (purple circles, C3/C4 - config. E).

The second pathway is aimed at finding out the minimum number of trials to train the CSP matrix and the minimum number of virtual channels needed to still get reliable performances. In order to achieve this we began by taking from 10% to 50% of the trials from the time series (4, 6, 10, 12, and 16 trials – half from the active class and half from the rest class) in order to train the CSP. We chose trials only from the first run because it is more likely that the subject paid more attention to the task and fatigue did not intervene; also, this simulates better the online case where if we were to choose to implement CSP, the training would be done with the first acquired trials. The CSP transformation for using 4, 6, 10, 12, and 16 filters is then applied to the data. The next step is to compute the power for every resulting time series and build thresholds by using LDA and SVM. In this case, the training/testing ratio will be from 10%/90% to 50%/50%, where the training set will be the same data taken for training the CSP matrix. Finally, we

computed AUCs for each dataset and the minimum number of trials and filters for CSP were selected. The data is then stored in a buffer as in the case of the previous pathway until the AUCs are computed for every healthy subject. The optimal values for all of the secondary parameters are then chosen and kept fixed for the second stage. The output of this stage is used to have the secondary parameters at a **fixed value** for the second stage.

2.2.6.2. Second Stage

In the second stage, we use the data from the stroke subjects. We vary the primary parameters and the independent parameters while keeping the secondary parameters at the fixed values found in the previous stage. By keeping the secondary parameters at a fixed value, we can now answer what power band feature provides the best discrimination and what is the minimum number of channels that can be reached while still maintaining reasonable performance. After finding these out we can argue about what is the minimum trial length that could be used in an online situation.

This stage uses starts by selecting the hand (side) on which to do analysis. The next step is to compute datasets for all of the channel configurations (config. A through config. E), thus five datasets. The stage continues in a similar manner as the first one by computing the datasets for all bands and all trial lengths. Now the pipeline splits into two paths. In the first one, we compute the variances for each dataset and then build power thresholds with LDA and SVM. The training/testing ratios used are the optimal ones found in the previous stage. Then the AUCs are computed and the data is kept in a buffer until the procedure is done for each stroke subject. In the second pathway the CSP matrix is computed and applied to each dataset. The number of trials and virtual channels are fixed to the values found in the previous stage. Then power thresholds are computed with LDA and SVM, and then AUCs are stored in a buffer until the procedure is done for every stroke subject. The AUCs in this case will let us know the optimal primary parameters. Finally judging also on AUCs we can now compare between the optimal performances for the independent parameters. The schematic of this stage can be seen in Figure 2.9.

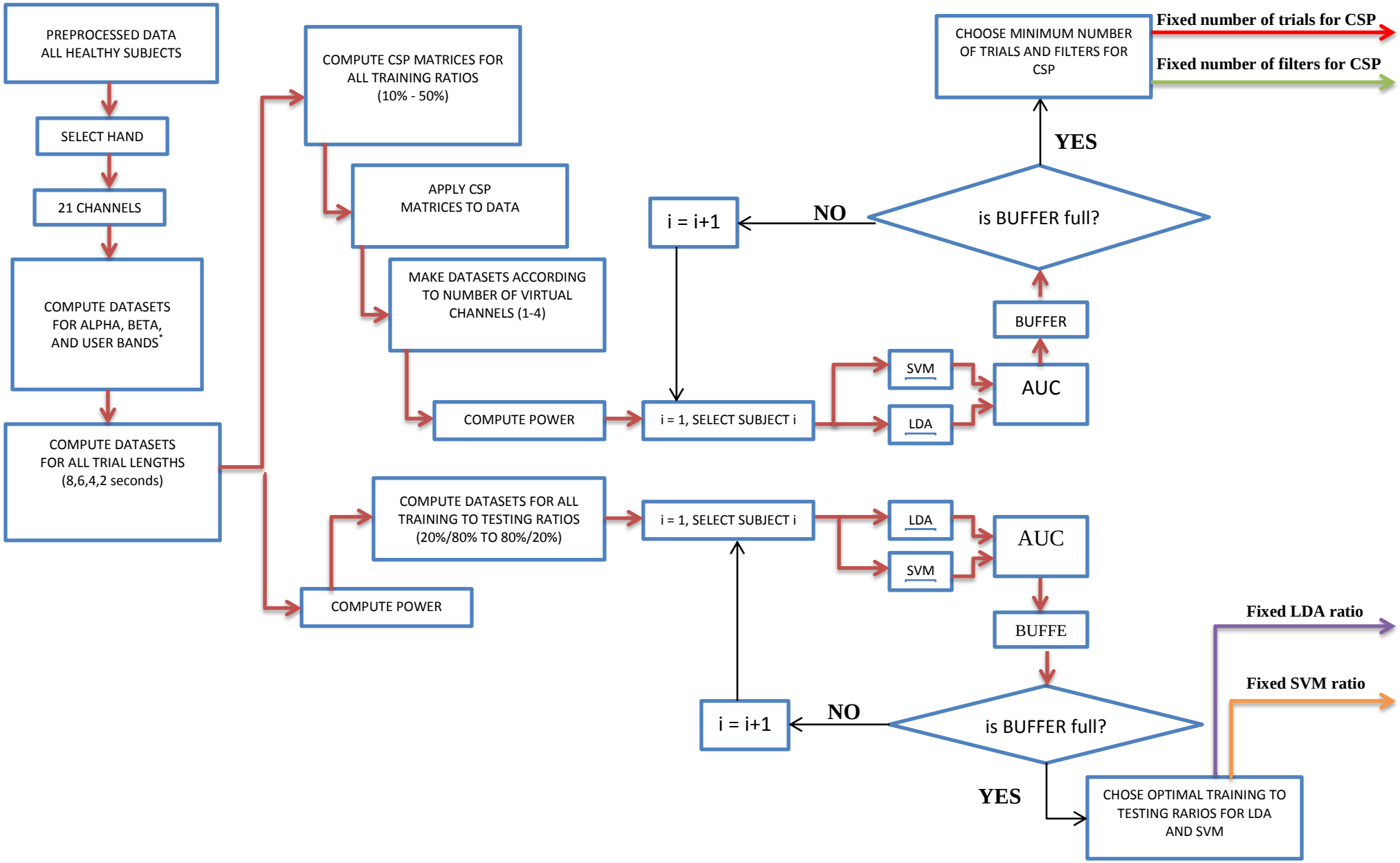


Figure 2.8: First Stage – healthy subjects; will output the minimum training to testing ratio needed in order to achieve reasonable predictions. It will also chose the number of trials used to get the CSP matrix and the number of CSP filters to be used. The pipeline is run for every subject and the results are stored in a buffer. After the data . Parameters are selected individually for each trial length.

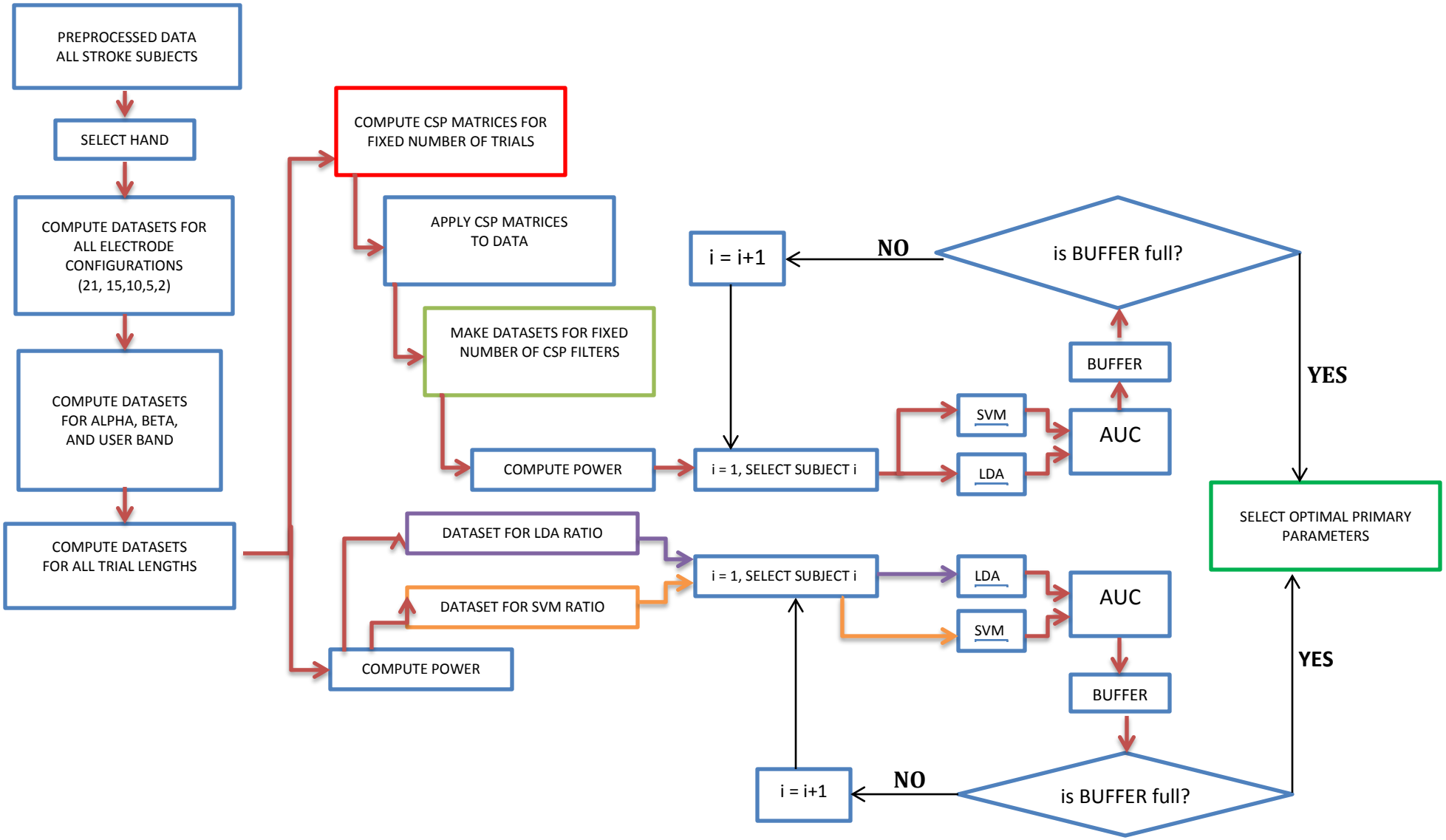


Figure 2.9: Second Stage - stroke subjects; Similar to the previous stage but now we keep the secondary parameters at a fixed value. In this stage, we also vary the number of channels. The outcome of this stage will let us choose the optimal primary parameters and let us compare between the independent parameters.

3. Results

This chapter will focus on the results obtained for one specific task: left MI vs. rest; we chose to present only these results because most of the stroke subjects had the left hand affected.

3.1. First stage

3.1.1. Choosing optimal training/testing ratio

Figure 3.1 and Figure 3.2 present mean and standard deviation AUCs averaged over all healthy subjects for all training/testing ratios investigated for LDA and SVM. We have found that in case of LDA only 70/30% and 80/20% splits could ensure performances above random for everybody in the control group. Even though the 80/20% split had the smaller standard deviation across subjects, we decided to choose the 70/30% split because we wanted to keep as many samples as possible for testing the classifiers performance. In the case of SVM all training/testing ratios showed good AUCs so we chose to use a 20/80% split for the next stage. These ratios were found to be valid for all four frequency bands and all trial lengths.

Table 3.1 summarizes the splits that were chosen and their corresponding AUC averaged across all healthy subjects. The results show that SVM outperforms LDA, even with small training sets. In the case of a 20/80% split, the AUC is between ~0.75 for α-band power and ~0.92 for β-band power. This indicates that SVM can generalize the data based on only a few instances.

Table 3.1: Results for training/testing split – LDA requires a larger amount of data to perform reasonably, whereas SVM can make accurate predictions using only a small training set.

	Broadband		α -band		β - band		User-band	
	Ratio (Tr/Ts)	Mean AUC	Ratio (Tr/Ts)	Mean AUC	Ratio (Tr/Ts)	Mean AUC	Ratio (Tr/Ts)	Mean AUC
LDA	70/30%	0.6720	70/30%	0.5781	70/30%	0.6493	70/30%	0.6318
SVM	20/80%	0.8653	20/80%	0.7580	20/80%	0.9196	20/80%	0.8510

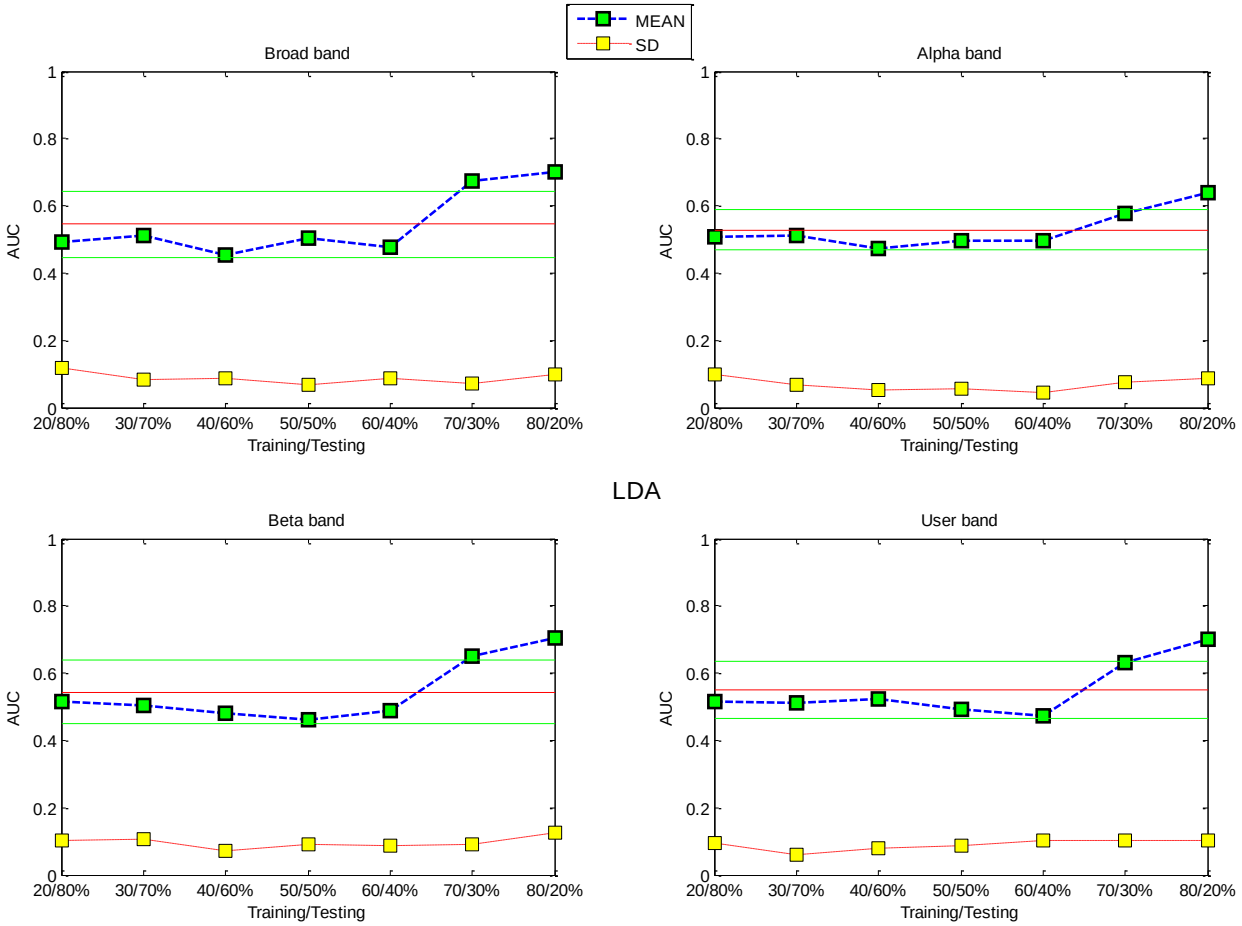


Figure 3.1: LDA - Mean and standard deviation for AUC over all subjects for the four features. The x-axis represents, in ascending order, the ratios between 20%/80% and 80%/20%

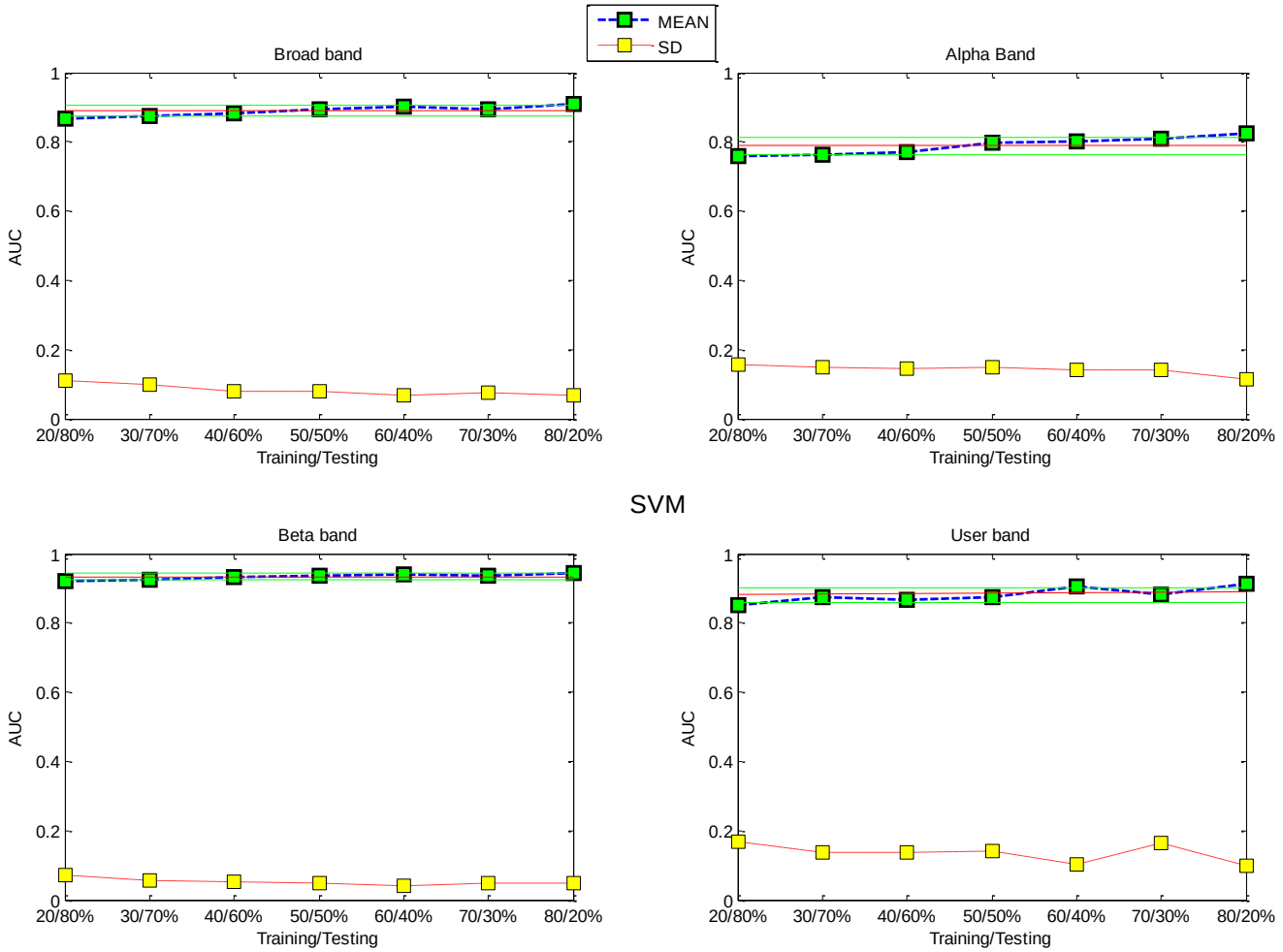


Figure 3.2: SVM - Mean and standard deviation for AUC over all subjects for the four features. The x-axis represents, in ascending order, the ratios between 20%/80 % and 80%/20

3.1.2. Choosing the optimal number of CSP filters and trials

For choosing the minimum number of trials to train the CSP and the number of CSP filters we employed a grid search. This means that for one specific number of trials we have computed the AUC for LDA and SVM for all number of filters mentioned in the previous chapter. We thus obtained 25 AUCs for each subjects; in order to assess the general performance we averaged the results over all subjects. Figure 3.3 shows the results for LDA; for the rest of the trial lengths and classification methods please refer to Appendix CSP. If the tone of the color is very similar across the whole grid it means that it does not matter what number of trials or filters we take. The results are similar regardless of trial length, band or classification method. As a result, we chose the number of trials and number of filters to be 4 for all trial lengths, for both LDA and SVM (minimum AUC ~=0.72). Table 3.2 summarizes the choices that were made after the "training" stage.

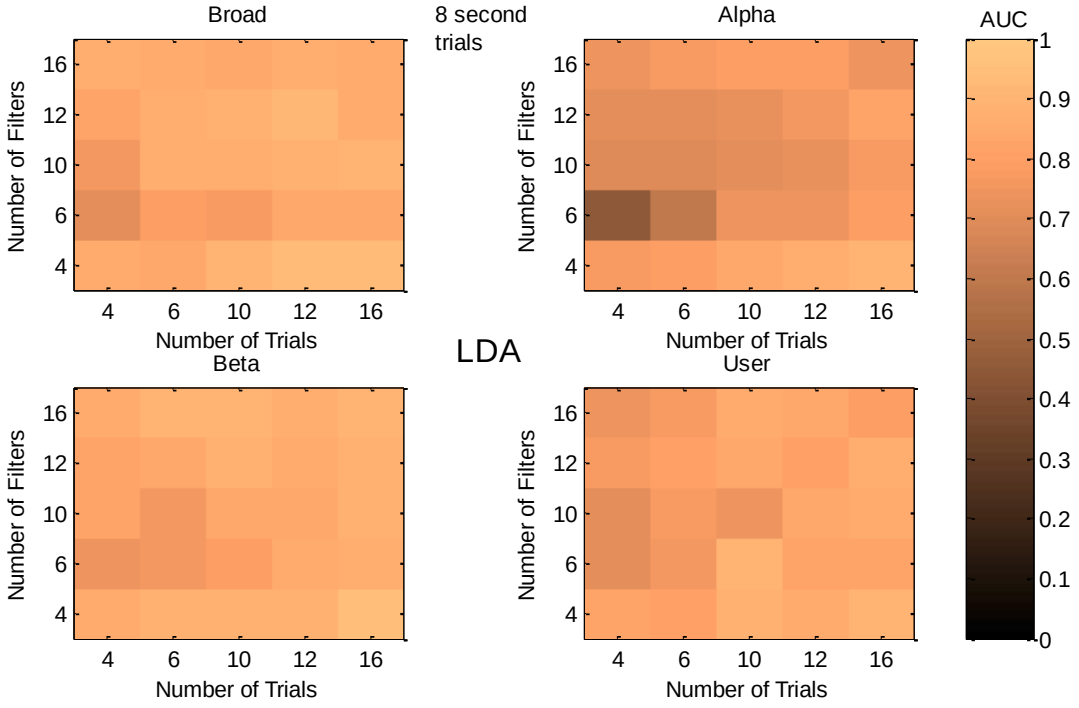


Figure 3.3: Grid search across all number of trials and filters for LDA for 8 second trials. The lighter the color tones the higher the value of the AUC.

Table 3.2: Parameters chosen after system calibration

	Training/Testing ratio	Number of trials for CSP matrix	Number of CSP filters
LDA	70%/30%	4	4
SVM	20%/80%	4	4

3.2. Second stage – Detailed Results

In this section, we will show detailed results for running the "testing" stage a stroke subject that had suffered a subcortical stroke - S03. Appendix F shows the results for a subject with cortical stroke –S09. These two subjects were chosen because they had had suffered different types of stroke and exhibited different clinical conditions at T0 – mild for S03 and severe for S09. No major differences were observed between the two subjects. Detailed results will not be shown for the other stroke subjects, but the results of the "testing" stage (AUCs) will be presented as an average in a following section.

Figure 3.4/Figure 3.5 show the ROC curves (right panel) and AUC (left panel) for S03, for LDA/SVM across all four sessions and frequency bands for 21 channels, for 8 second trials. As expected, SVM outperforms LDA for this case.

We proceed with showing the AUC for the same subject but this time over all channel configurations and trial lengths for T2. We chose this run because it had the worst performance, even for SVM. Figure 3.6/Figure 3.7 show the AUCs (z-axis) for all electrode configurations (x-axis) and trial lengths (y-axis) for LDA/SVM in stroke subject S03. It is worth noting that reducing the number of channels in general improves performance. As an observation, results show that we can go as low as 2 electrodes (C3/C4) and still retain high performance suggesting that computing the user specific frequency band based on either C3 or C4 is a valid approach.

We move on to presenting the results for CSP in the same subject; CSP was applied only on three electrode configurations (A, B, and C) because it would not make sense to apply CSP on 5 channels or less when we have chosen the number of CSP filters to be 4. Figure 3.8/Figure 3.9 show the ROC and AUC for the combination of CSP+LDA/CSP+SVM for 8 second trials across all features and sessions in stroke subject S03. It can be noted that the CSP performance is highly dependent on the feature. As in previous case, we present the AUC for session T3 in S03 for all electrode configurations (x-axis) and trial lengths (y-axis) for LDA/SVM. Figure 3.10/Figure 3.11 illustrate this for S03.

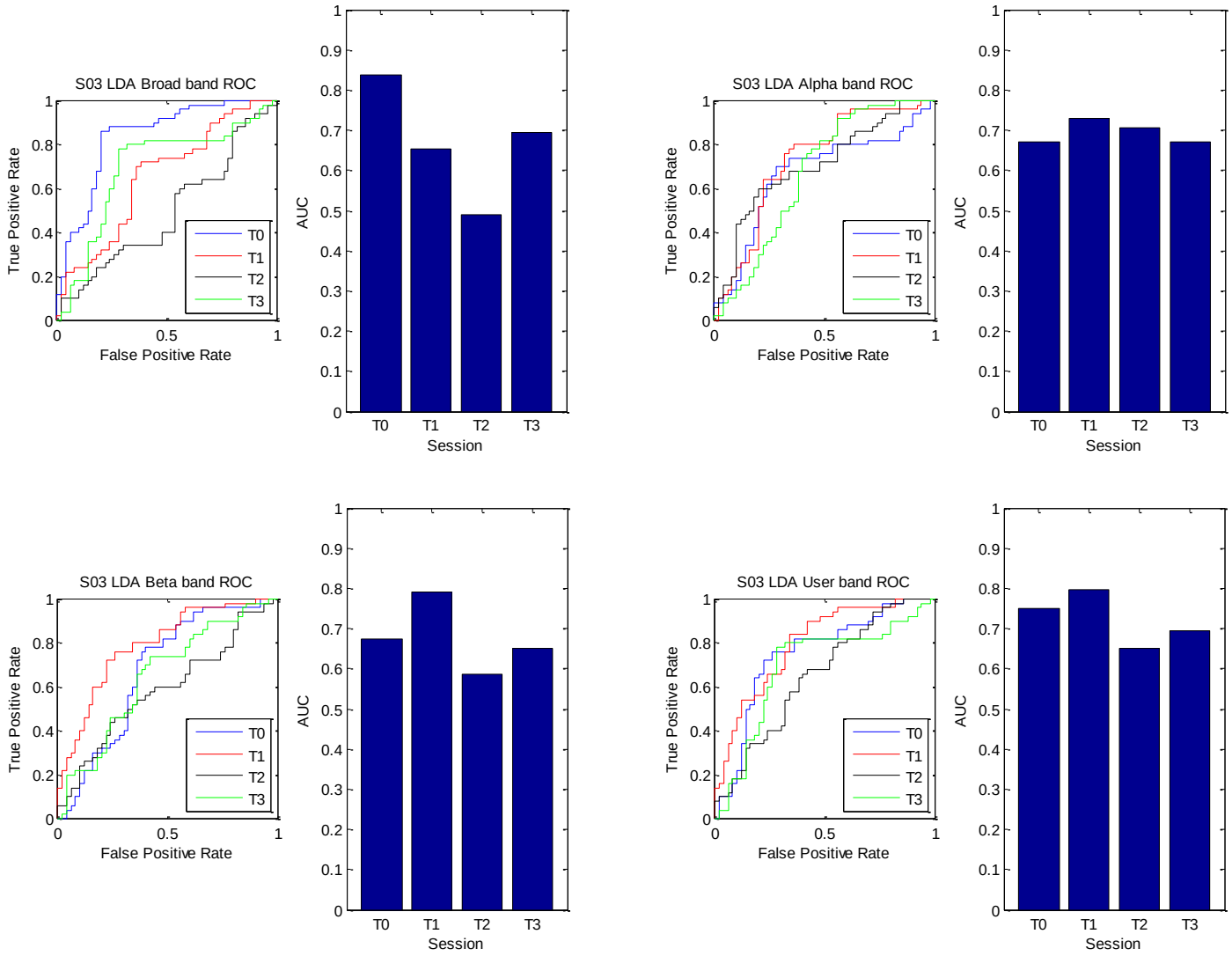


Figure 3.4: LDA – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S03

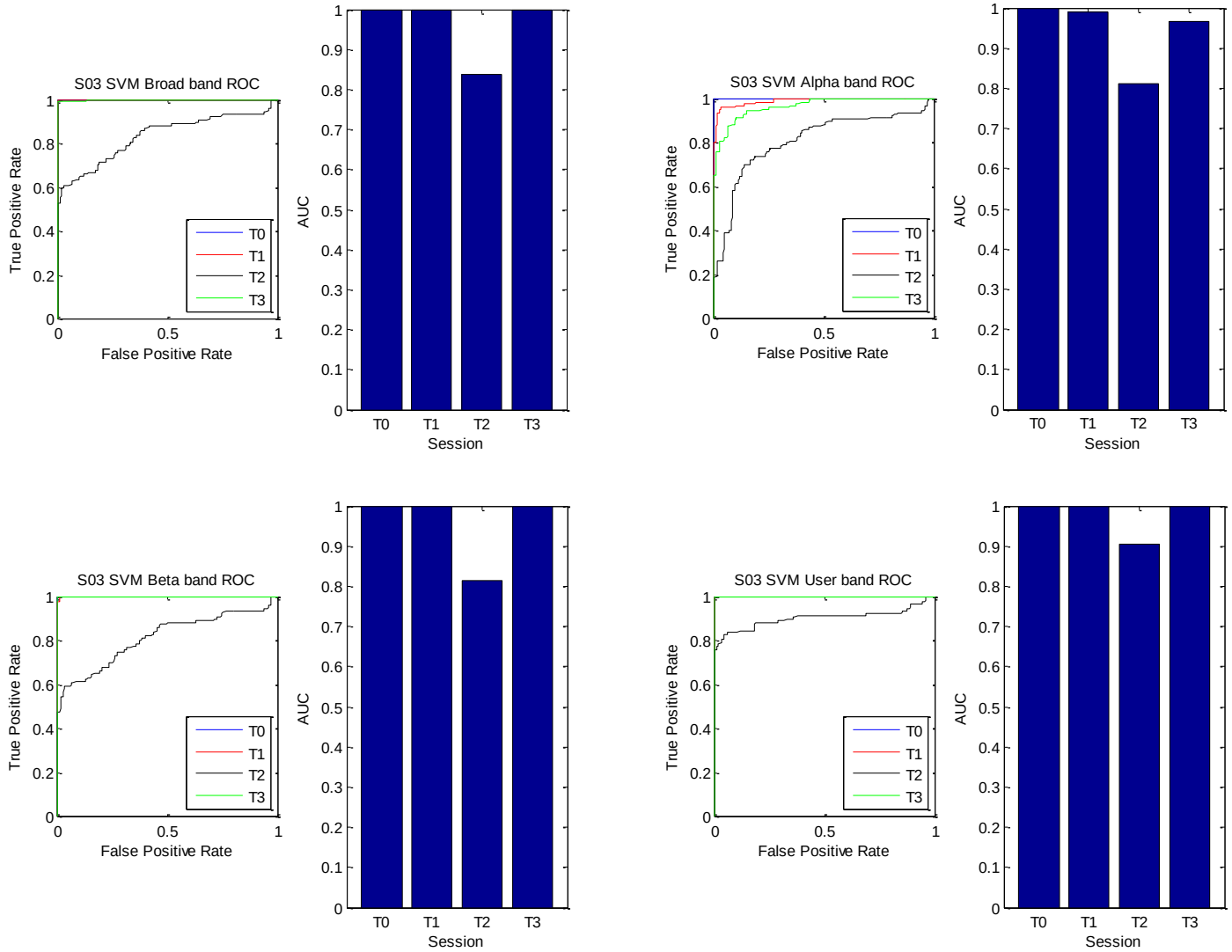


Figure 3.5: SVM – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S03

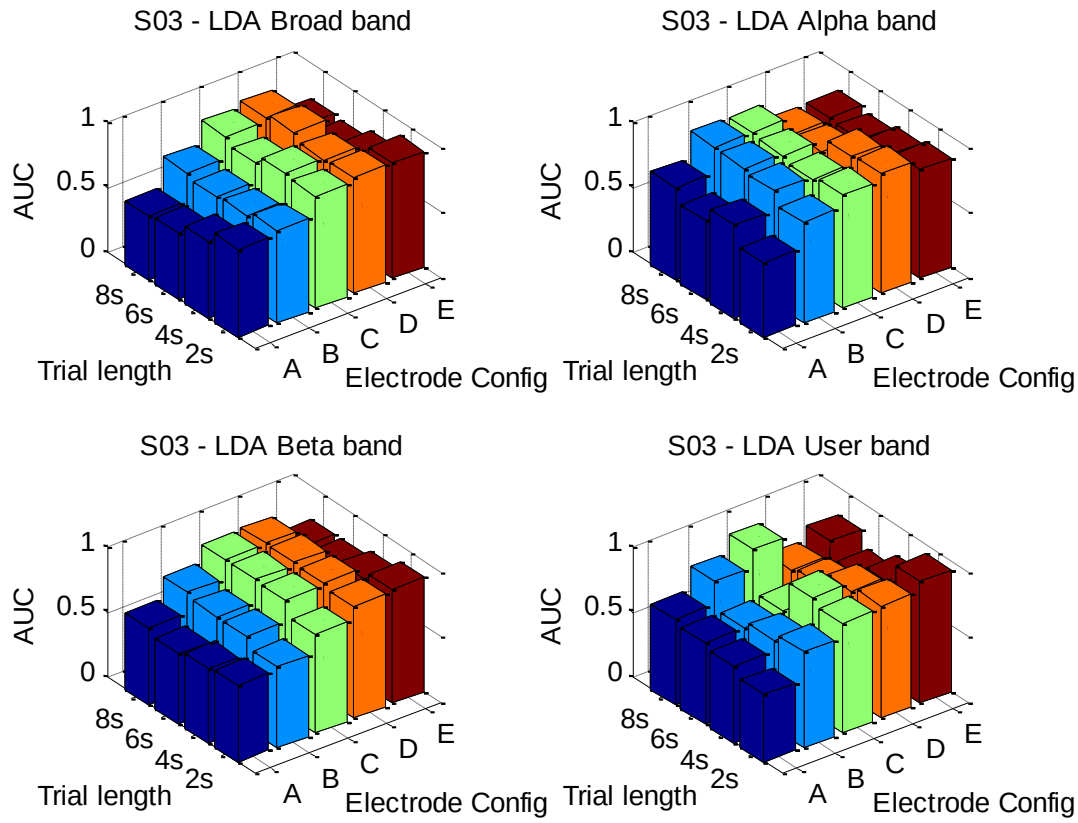


Figure 3.6: LDA AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S03

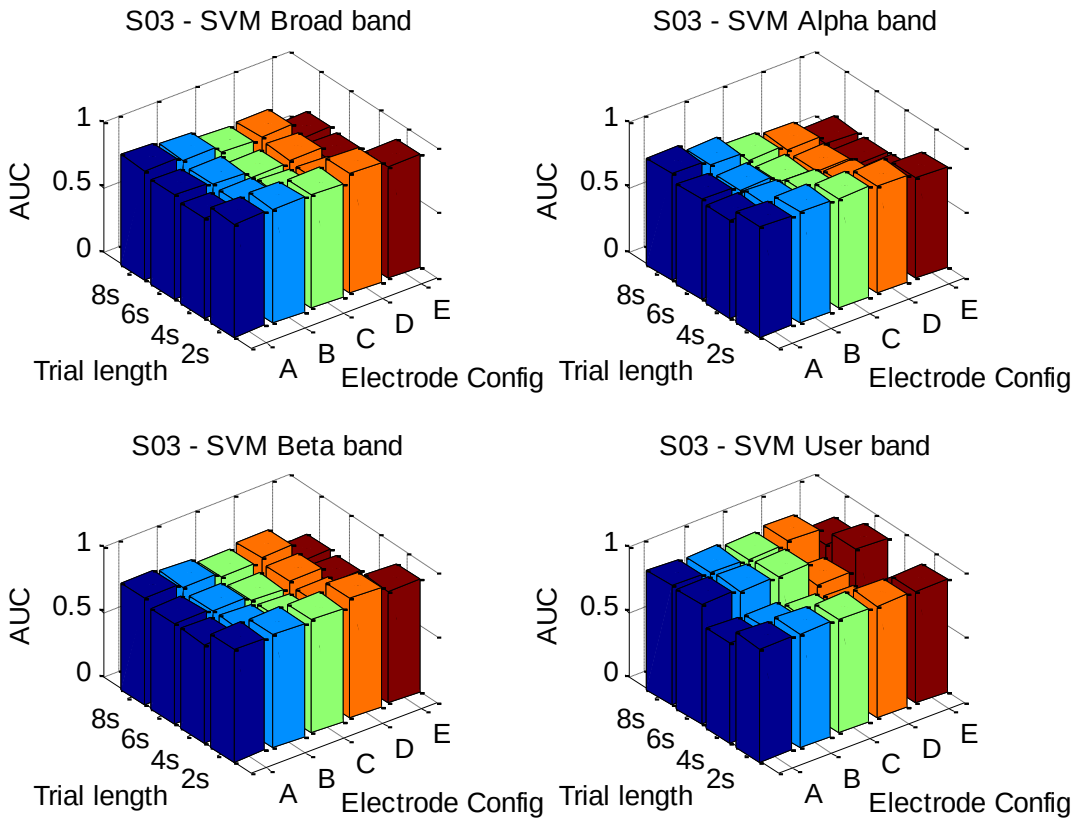


Figure 3.7: SVM AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S03

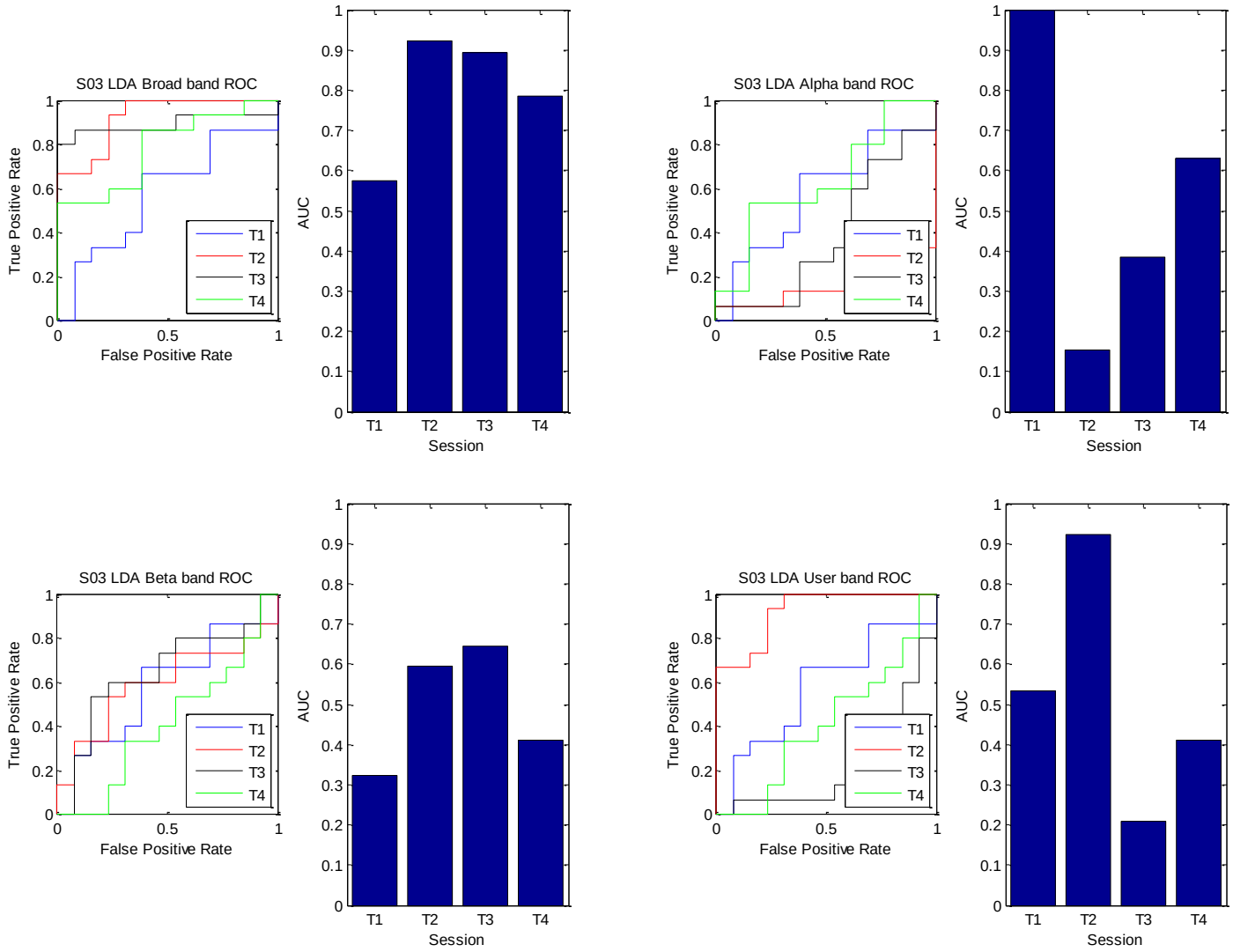


Figure 3.8: CSP + LDA – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S03

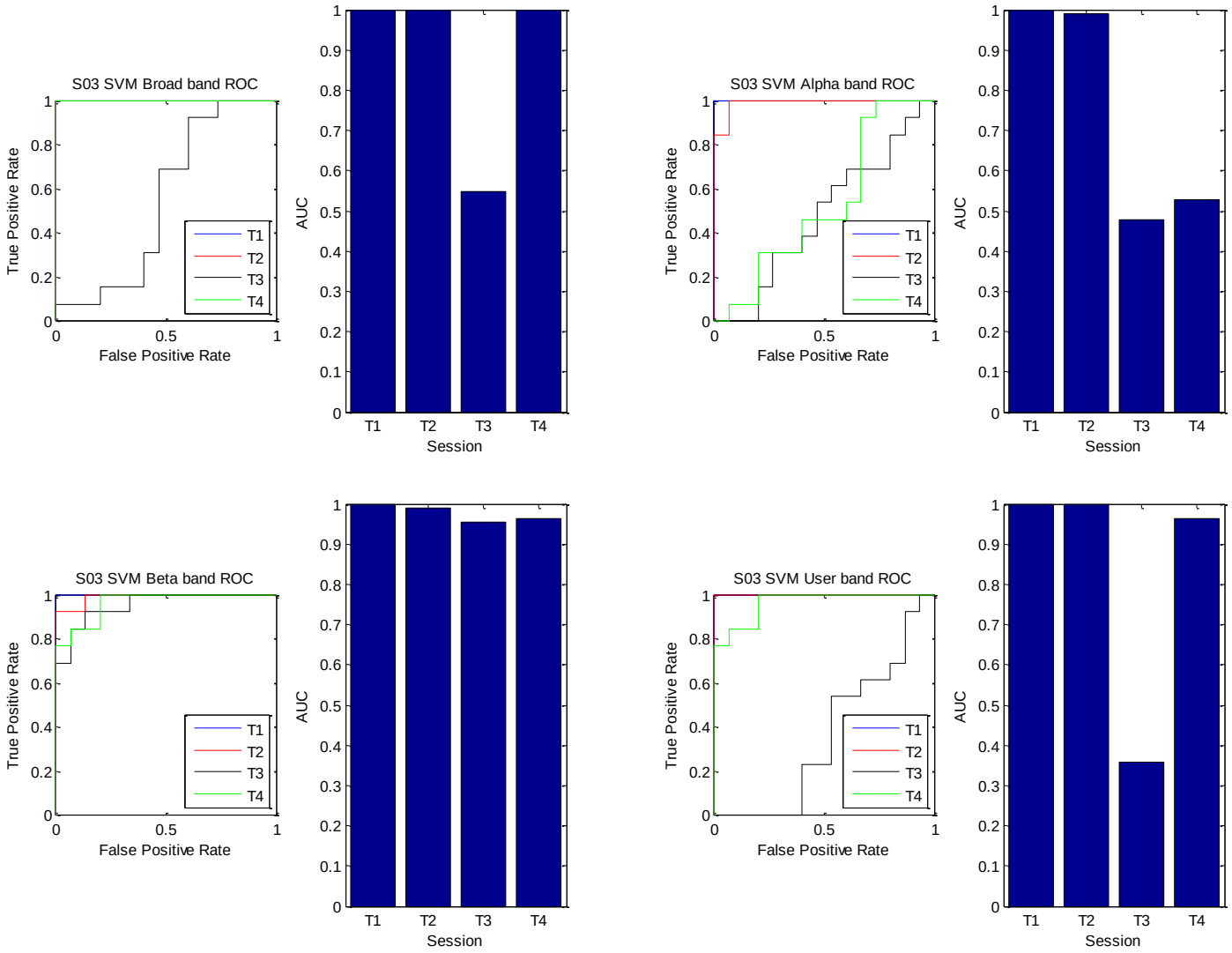


Figure 3.9: CSP + SVM – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S03

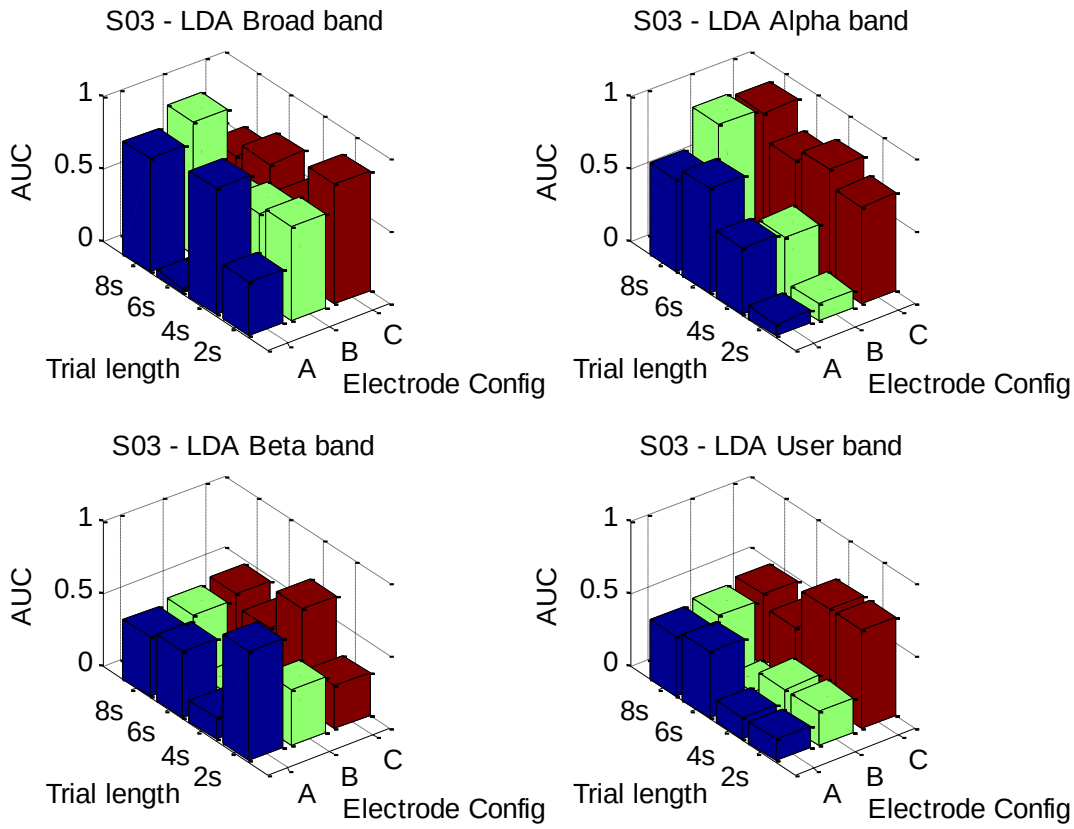


Figure 3.10: CSP + LDA AUCs for all trial lengths and all electrode configurations for run T3 in stroke subject S03

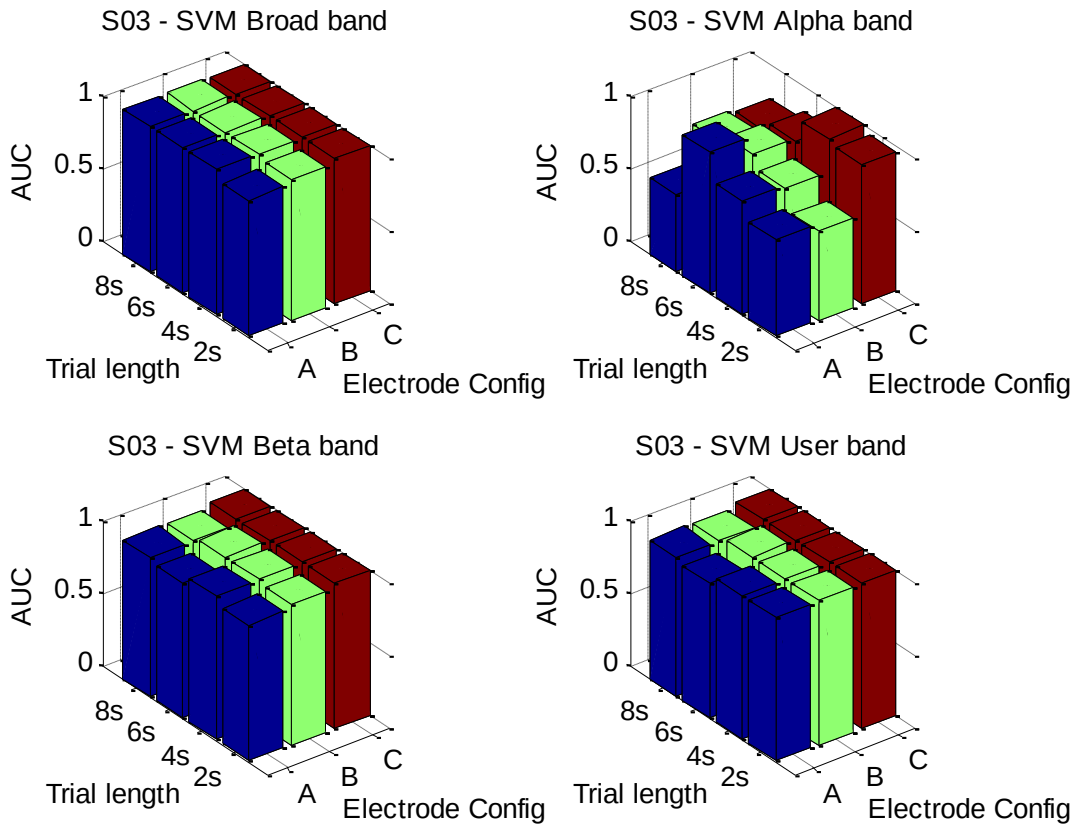


Figure 3.11: CSP + SVM AUCs for all trial lengths and all electrode configurations for run T3 in stroke subject S03

3.3. Overall outcome

Figure 3.12 to Figure 3.15 present the performances for all features, all classification methods, and electrode configurations averaged over all stroke subjects, including S03 and S09, and sessions. We observe that in general all features have high AUCs, thus opting for user band in an online setting would not be worth the extra computational time. The safest choice is the broad band feature because performances for α and β are similar suggesting that useful information is present in both.

In terms of classification methods, SVM and CSP+SVM exhibit the highest performances across electrode configurations A, B, and C (20, 14, and 10 electrodes). In configurations D and E (5 and 2 electrodes) LDA slightly outperforms SVM. Despite the fact that the combination of CSP + SVM shows the highest performances, LDA + configuration, D or E is sound option for an online system. We say this because the difference in performance is not notable and because one of the objectives is to minimize the number of electrodes. Another argument for choosing LDA is that it is less computationally expensive than CSP+SVM.

Results indicate that high performances are maintained across all trial lengths. This means that an online system can be implemented for any of the trial lengths.

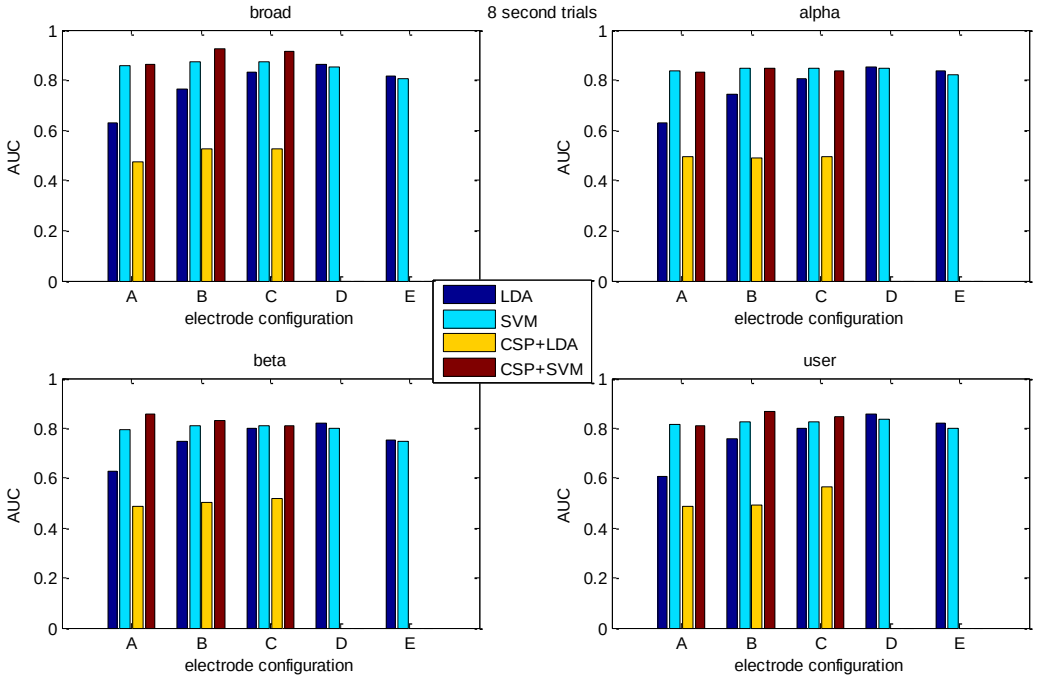


Figure 3.12: AUCs for all features, all classification methods and all electrode configurations averaged across all stroke subjects for 8 second trials.

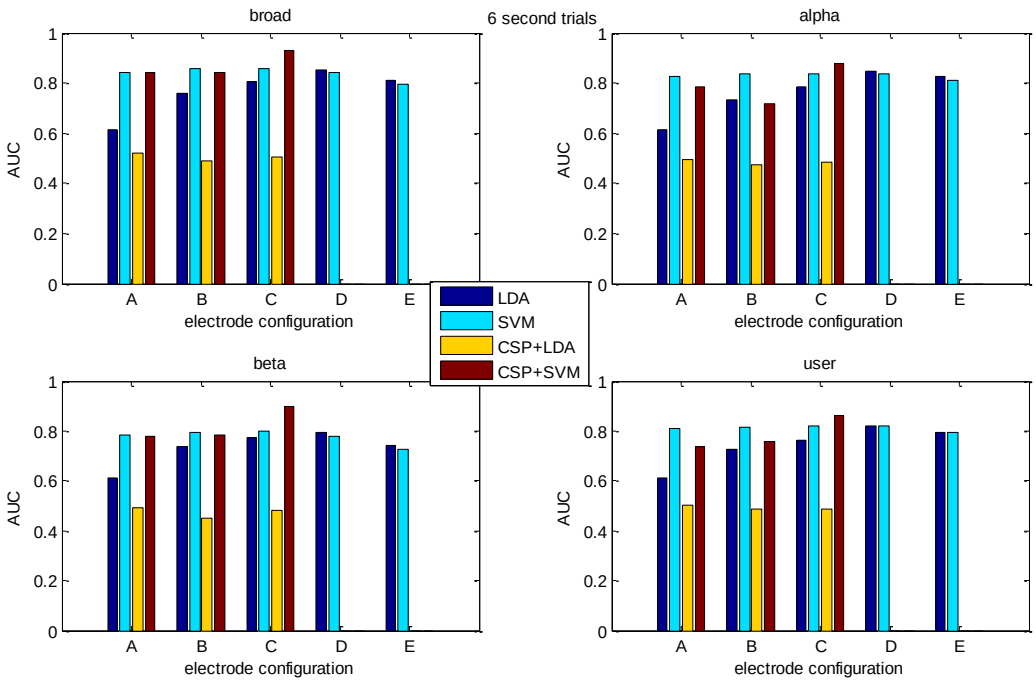


Figure 3.13: AUCs for all features, all classification methods and all electrode configurations averaged across all stroke subjects for 6 second trials.

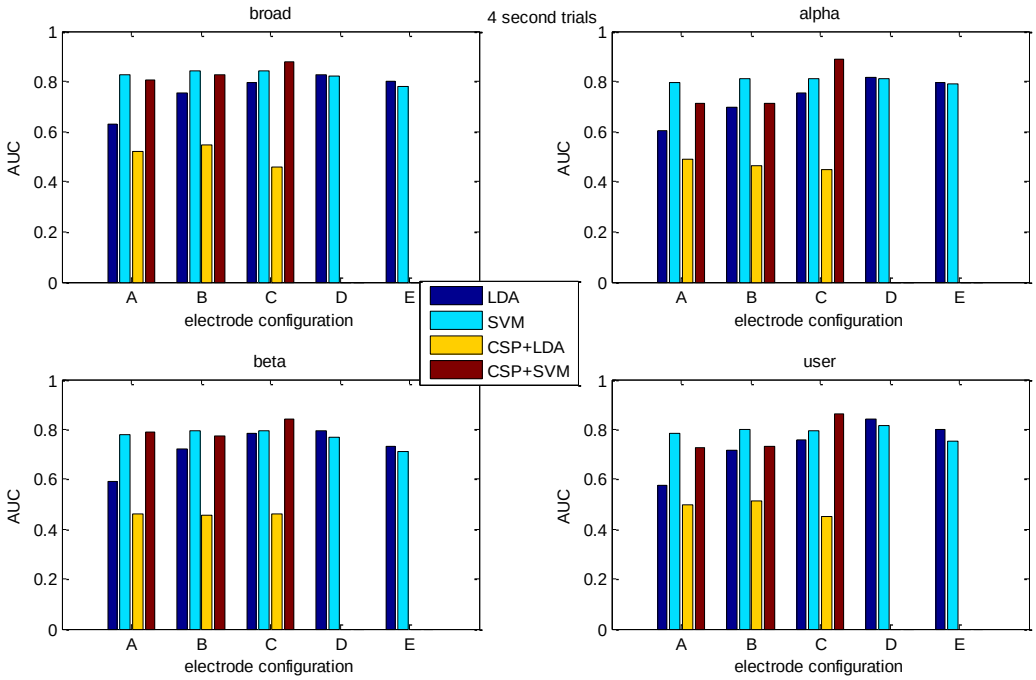


Figure 3.14: AUCs for all features, all classification methods and all electrode configurations averaged across all stroke subjects for 4 second trials.

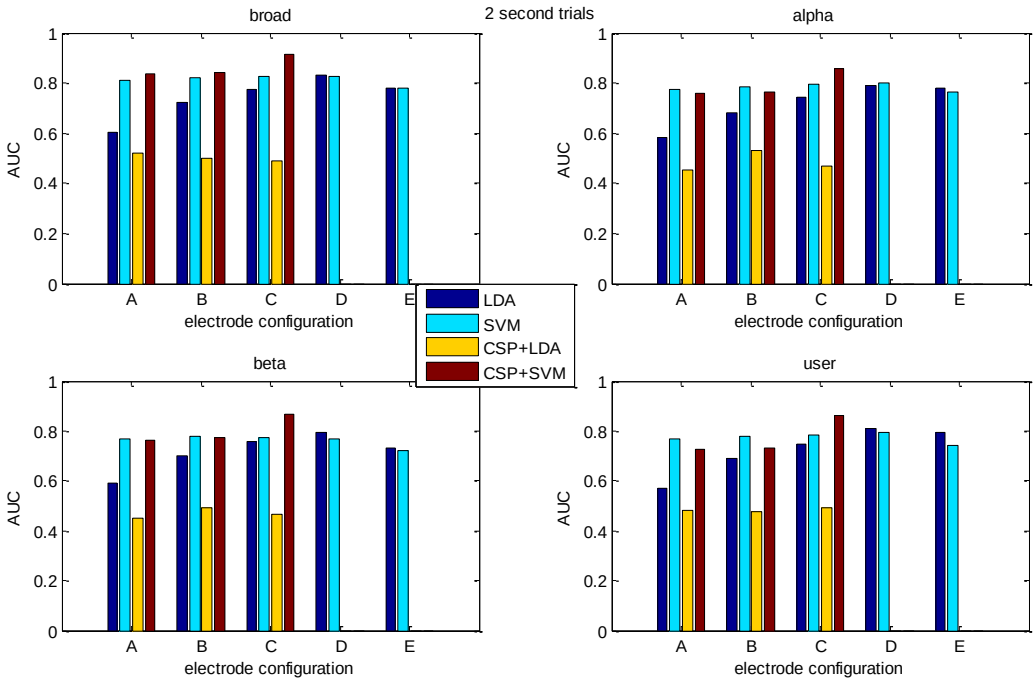


Figure 3.15: AUCs for all features, all classification methods and all electrode configurations averaged across all stroke subjects for 2 second trials.

Summarizing, the training/testing ratio used in this study was 70%/30% for LDA and 20%/80% for SVM. We have shown that using 4 trials to train the CSP and 4 virtual channels are enough to provide good performances. The best classification outcome is given when using 15 electrodes by the CSP+SVM combination. Nevertheless good performances are exhibited for all electrode configurations and classifiers, except for the combination of CSP+LDA. We have shown that results are similar for all trial lengths.

4. Discussion and Conclusions

The research questions that we have addressed at the beginning were: (1) what feature is most suited for discriminating between active and rest state, (2) what classification technique is more suited for achieving single-trial classification, (3) can the number of channels be diminished while still keeping reliable performances, and (4) what is the influence of trial length on the previous points?

4.1. Best candidate feature for online classification

In our study, we have used four bandpower features that were non-parametrically estimated as opposed to the methods presented in Daly's et al. [25] and Parsad's et al. [26] studies. Even though the parametric methods used in [25, 26] provide higher spectral resolution it is known that using a too short length results in an overly smooth estimate. In addition, the model estimate depends on the sampling frequency and model error criterion [27]. Our method does not suffer from any of these shortcomings.

We have shown that in general all four features, i.e. signal power in the broad, α, β, and user bands elicit good performances for all electrode configurations. This suggests that on average meaningful modulation is present in both α and β bands. Daly et al. [25] report using the spectral power estimate between 21 and 24 Hz as their feature, whereas Parsad et al. [26] use the power estimates from the α band and β band.

Broad-band had the highest performances in most cases. This is to be expected, on average, because it is the band that contains the most information. Whereas this is true for the average, it might be that on an individual level it is not the best option. For example, in the case of stroke subject S09 broad-band had closer performances to α, while β performances were worst. This indicates that meaningful modulation is mainly present in α band. Since the aim is to target only the sensorimotor rhythm of the stoke subject, in this case, it is better to use the power in the α band as a feature.

Initially we had expected the user band to have the highest performances. We believe this was not the case because the user band was computed only according to the activity present on C3 or C4. Daly et al. [25] found that in their subject the highest modulation was present in electrode CP3. This suggests that the sensorimotor rhythm may be shifted towards the parietal area. As a result, performances may improve if the user specific frequency algorithm is extended to take into account adjacent electrodes of C3/C4.

4.2. Best classification method for single-trial classification

Although we use a different metric for measuring classification performances, overall our performances are better than the ones presented in Parsad's [26] study. We have shown that LDA ensures good performances with the general observation that its performance increases when the number of electrodes decreases. Because LDA builds its threshold based on the pooled covariance matrix it means that it requires at least as many trials for training as the number of channels. This ensures that the covariance matrix is nonsingular. To our knowledge, there is only one study, by Keiser et al. [43] that uses LDA and deals with actual stroke data. In terms of classification performances, our results are similar to theirs.

We have shown that, in general, SVM outperforms LDA even given the difference in training/testing ratios. This happens due to the maximum-margin hyperplane, allowed misclassification on the training set, and the RBF kernel [31, 54].

One interesting case is that of stroke subject S03 where the AUC is 1 when using SVM for several sessions and features (Figure 3.5). At first glance, one can interpret this this as the result of overfitting. We suspect this is not the case because SVM is known to be relatively insensitive to overtraining [55] and because the number of trials used for training was small (6 trials). Whether this is really a matter of overfitting the data or of near- perfect modeling of the data is a question that is better to answer using an online system.

Our results show that, when combined with SVM, CSP provides the best classification performances. Similar results concerning the combination of CSP and SVM are mentioned in [52]. When combining CSP and LDA, our results are comparable to the ones presented by Ortner [30] for healthy subjects but not for stroke subjects. It is known that CSP is sensitive to outliers and is prone to overfitting when provided with small training sets [37]. This means that CSP focuses on irrelevant data shown in stroke EEG leading to a nonlinearly separable feature space. In light of these facts it is not surprising that CSP+LDA performs badly in the case of stroke subjects. Improved CSP algorithms that address these shortcomings of the original CSP are presented in [53]. Nevertheless, given the nature of stroke data, it is uncertain if the improved CSP algorithms will actually reflect the underlying physiological phenomenon.

Despite the fact that CSP+SVM improves classification performances, it is questionable whether it is reliable for online feedback. The high performances are clearly due to the aforementioned advantages of SVM and do not reflect the desynchronization of the sensorimotor rhythm.

4.3. The meaning behind the number of channels

Our third research question was what is the minimum number of channels that can be used to reliably provide feedback. A study by Tam et al. [56] investigates 6 electrode configurations with 31 (2 configurations) an 10 electrodes (4 configurations). Their results show that the best classification performances are obtained for the lower number of electrodes with close performances for all 4 configurations.

Our results indicate that high performances are displayed even for 2 channels. Nevertheless, the choice of the minimum number of channels should be done with respect to the underlying physiological phenomenon.

As argued earlier it might be that the most representative activity for the sensorimotor rhythm is not necessarily found in C3/C4 electrodes. As such choosing a higher number of channels is a more sound decision. One other observation is that LDA starts outperforming SVM in the case of 5 and 2 channel configurations. This indicates that the feature space becomes linearly separable in this cases. When combining these two pieces of information we can say that the modulation we expect is represented best by the feature space described 5 channels.

4.4. Influence of trial length on the primary and secondary parameters

To our best knowledge, there are no studies that investigate the influence of trial length on feature reliability and classification performances. Our results also indicate that the investigated trial lengths do not influence a BCI system that uses our classification methods.

This happens because the estimated power is almost the same for all trial lengths. This suggests that our feature is weakly stationary. To see if this is so we computed the features for 2 seconds over 7 intervals , for 4 seconds for 3 intervals, and for 6 seconds over 2 intervals for all stroke subjects and compared it to the power estimated on the whole trial. We observed that on average the variances were stable (for example, for broad-band during left MI, C4 electrode the mean was 6.6328, SD 0.0856 for active trials and mean 9.6582, SD 0.4363 for rest trials; grand average over all stroke subjects).

This would suggest that the estimated values are close to the actual variance. The variance is known to be an unbiased estimator given a large number of samples. Even in the case of 2 second trials we use 1000 samples to estimate the power. This implies that the sampling frequency plays a great role in this result.

From a physiological point of view, we can conclude that after ~2 seconds of active state the user can revert to a state. We say this because we have discarded ~1 second from the beginning and the end of the trials.

4.5. Conclusions and future considerations

We have shown that the power over broad, α, β, and user bands are reliable features for discriminating between active and rest state, with slightly better performances in the case of broadband. Furthermore, we have proven that LDA and SVM are good candidates for classification in an online setting and we have argued why the normal CSP algorithm is not a good spatial filter for stoke data.

The minimum number of electrodes needed to provide feedback was found to be 2 (C3/C4), but in consideration of the underlying physiological phenomenon and classifier performance, electrode configuration D (5 channels) is more suited for an online setting. Lastly, the trial length does not have a major impact on the system's performance suggesting that trials as small as 2 seconds can be used in an online setting, provided a cue be given before the task starts.

The findings of this study imply that a MI based BCI system for stroke rehabilitation in a home environment is a feasible possibility. Our results suggest that set up time for such a system can be done faster than in a clinical setting. Overall system speed can be increased and calibration times lowered by using a choosing LDA and using less than 22 trials for training. This can be done because the number of channels is 5 implying we would need a minimum of 5 samples for estimating class covariance matrices.

In light of the knowledge gained by the author during this project several suggestions are given. It is possible, according to [43], for calibration to be performed using data from real movement and achieve good performances for classifying MI data. Unfortunately, this is a possibility only if the subject's affected limb is not completely paralyzed.

A second suggestion is for the online system to have both LDA and SVM combined in a voting system. The drawback of this approach is that it adds computational time. In order to avoid this problem, the system should be implemented in C/C++/C# or another language that ensures high computational speed. A final suggestion is to include an outlier detection and rejection module. This can be done with the aid of SVM; if a trial is found to have a large value for the Lagrangian multiplier, α, then the sample is most probably an outlier.

Acknowledgements

I would like to express my gratitude to my supervisors, Wim Rutten, Michel van Putten and Chin Tangwiriyasakul for their constant constructive attitude towards my project, and making me feel like I belong.

I want to thank Wim for his sincere critic attitude and for making me question my decisions until I was 100% sure of what I was saying/doing.

I would like to thank Michel whose words, “Deep insights requires long contemplation”, echoed daily in my mind.

I would like to show my appreciation to Chin for withstanding my constant questions with a smile on his face, and his constant encouragements.

I would like to thank Irina Stoyanova and Ed Droog who, without their knowing, cheered me up when I was at my lowest ebb.

Special thanks to my parents, who in these two years have always been in my heart and were always there through thick and thin. I want to thank my girlfriend whose smile always brightened my day. I would like to thank my friends here at Twente: Alex, Antonia, Aykan, Dan, Iannis, Ioannis, Hristos, Laura, Matei, Mircea, Rãzvan and all the others for making my free time so enjoyable. I would also like to thank my friends back home: Costel, Dinu, Lavinia, Loredana, Mircea, Oana, Vanda and Vlad for their constant support.

Last, but not least I want to thank Nicoleta Stoica, who laid the foundation for my journey in the field of engineering.

References

[1] Dale Purves et al., “Neuroscience” third edition, 2004, Sinauer Associates

[2] American Stroke Association, http://www.strokeassociation.org/STROKEORG/

[3] Wolfe Charles D. A., “The impact of stroke”, 2000, British Medical Bulletin, 56 (2): 275-286

[4] Adriaansen Jacinthe J.E. et al., “Course of social support and relationships between social support and life satisfaction in spouses of patients with stroke in the chronic phase”, 2010, Patient education and counseling, Nov 85(2):e48-52

[5] Low Joseph T.S et al., “The impact of stroke on informal carers: a literature review”, 1999, Social Science and Medicine, Sep 49(6) 711-712

[6] Han Beth et al., “Family caregiving for patients with stroke – review and analysis”, 1999, Stroke, 30: 1478-1485

[7] Crow J.L. et al., “The effectiveness of EMG biofeedback in the treatment of arm function after stroke”, 1989, Disability and Rehabilitation, 11(4) 155-156

[8] Cohn P.J. “Pre-performance routines in sport: theoretical support and practice”, 1990, Sport Psychology, Sep 4(3):301-302

[9] Page Stephen J., “Imagery improves motor function in chronic stroke patients with hemiplegia: a pilot study”, 2000, Occupational Therapy Journal of Research, 20 200 –215

[10] Page Stephen J. et al., “Imagery combined with physical practice for upper limb motor deficit in sub-acute stroke: a case study”, 2001, Physical Therapy, Aug 81(8):1455–1462.

[11] Page Stephen J. et al, “A randomized, efficacy and feasibility study of imagery in acute stroke”, 2001, Clinical Rehabilitation, Mar 15(3):233-240

[12] Page Stephen J.et al, “Mental practice in chronic stroke – Results of a randomized, Placebo-Controlled Trial”, 2007, Stroke, Apr (4)38: 1293-1297

[13] Crosbie Jacqueline H. et al., “The adjunctive role of mental practice in the rehabilitation of the upper limb after hemiplegic stroke: a pilot study”, 2004, Clinical Rehabilitation, Feb 18(1): 60

[14] Dijkerman HC et al., “Does motor imagery training improve hand function in chronic stroke patients? A pilot study”, 2004, Clinical Rehabilitation, Aug 18(5):538-549

[15] Arroyo S. et al.,”Functional significance of the mu rhythm of human cortex: an electrophysiologic study with subdural electrodes”,1989, Electroencephalography Clinical Neurophysiology , Sep 87 (3) 76-87

[16] Millán J. d. R. et al., “Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges”, 2010, Frontiers in Neuroscience, Sep 4(161)

[17] Malmivuo Jaakko, Robert Plonsey, “Bioelectromagnetism”, 1995, Oxford University Press

[18] Pfurtscheller G. et al., “Rehabilitation with Brain-Computer Interface Systems”,2008, Computer, 41(10) 58-65

[19] Grosse-Wentrup M. et al., “Using brain–computer interfaces to induce neural plasticity and restore function”, 2011, Journal of Neural Engineering, Apr 8(2)

[20] Daly Janis J et al., “Brain–computer interfaces in neurological rehabilitation”, 2008, The Lancet Neurology, Nov 7(11) 1032-1043

[21] Pfurtscheller G., “Event–related synchronization (ERS): an electrophysiological correlate of cortical areas at rest”, 1992, Electroencephalography and clinical Neurophysiology, Jul 83(1):62-69

[22] McFarland Denis J., “Mu and beta rhythm topographies during motor imagery and actual movements”, 2000, Brain Topography, 12(3): 177-186

[23] Pfurtscheller G., “Event-related EEG/MEG synchronization and desynchronization: basic principles”, 1999, Clinical Neurophysiology, Nov 110(11):1842-1857

[24] Purtscheller G. al., “Motor imagery activates primary sensorimotor area in humans”, 1997, Neuroscience Letters, Dec 239 (2-3) 65-68

[25] Daly Janis J et al.,”Feasibility of a new application of noninvasive brain computer interface (BCI): a case study of training for recovery of volitional motor control after stroke”, 2009, Journal of Neurologic Physical Therapy, Dec 33(4):203-211

[26] Prasad Girijesh et al., “Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: a feasibility study”, 2010, Journal of Neuroengineering and Rehabilitation, Dec 7(60)

[27] Krusienski D.J. et al., “An Evaluation of Autoregressive Spectral Estimation Model Order for Brain-Computer Interface Applications”, 2006, Proceedings of the 28th IEEE,EMBS Annual International Conference, New York City, USA, Aug 30-Sept 3, 1324-1326

[28] Naeem M. et al., “Dimensionality Reduction and Channel Selection of Motor Imagery Electroencephalographic Data”, 2009, Computational Intelligence and Neuroscience, Article ID 537504

[29] Naeem M. et al., “Seperability of four-class motor imagery data using independent components analysis”, 2006, Journal of Neural Engineering, Sep 3(3) 208-216

[30] Ortner R. et al., “Brain-computer Interfaces for stroke rehabilitation: evaluation of feedback and classification strategies in healthy users”, 2012, The fourth IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, Roma, Italy, June 24-27.

[31] Lotte F. et al., “A review of classification algorithms for EEG-based brain computer interfaces”, 2007, Journal of Neural Engineering, 4(2) 1-13

[32]TMS International, www.tmsi.com/?id=24

[33] Muller-Gerking J. et al., “Designing optimal spatial filters for single-trial EEG classification in a movement task”, 1999, Clinical Neurophysiology, May 110(5) 787-798

[34] Ramoser H et al., “Optimal Spatial Filtering of Single Trial EEG During Imagined Hand Movement”, 2000, IEEE Transactions on Rehabilitation Engineering, Dec 8(4) 441-446

[35] Koles Z.J., “The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG”, 1991, Electroelcephalography and Clinical Neurophysiology, Dec 79(6) 440-447

[36] Koles Z.J. et al., “Spatial patterns in the background EEG underlying mental disease in man”, 1994, Electroencephalography and clinical neurophysiology, Nov 91(5) 319-328

[37] Blankertz B. et al., “Optimizing Spatial filters for Robust EEG Single-trial Analysis”, 2008, Signal Processing Magazine, IEEE, 25(1) 41-56

[38] Vapnik V., “Statistical Learning Theory”, 1998, Wiley

[39] Cortes C. and Vapnik V,”Support-Vector Networks”, 1995, Machine Learning, 20 273-297

[40] Alpaydin E., “Introduction to Machine Learning”, 2010, The MIT Press

[41] Chapelle O. et al., “Choosing Multiple Parameters for Support Vector Machines”, 2002, Machine learning, 46 (1-3) 131-159

[42] Hakin O.,”Neural Networks and Learning Machines” ,2008, Prentice Hall, 3rd edition, p209

[44] Fukunaga K, “Introduction to Statistical Pattern Recognition”, 1990, Academic Press

[43] Keiser Vera et al., “First steps toward a motor imagery based stroke BCI: new strategy to set up a classifier”, 2011, Frontiers in Neuroscience, Jul 5(86) 1-10

[45] Haeb-Umbach R. et al., “Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition”, 1992, IEEE International Conference on Acoustics, Speech, and Signal Processing, 23-26 March

[46] Krzanowski, W. J., “Principles of Multivariate Analysis: A User's Perspective”, 1988, Oxford University Press

[47] Muller K. R. et al., “Linear and Nonlinear Methods for Brain-Computer Interfaces”, 2003, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Jun 11(2) 165-169

[48] Vickneswaran Jeyabalan, et al.,”Motor Imaginary Signal Classification Using Adaptive Recursive Bandpass Filter and Adaptive Autoregressive Models for Brain Machine Interface Designs”, 2007,World Academy of Science, Engineering and Technology, May (5) 425-432

[49] Lodder Shaun, “Single-Trial Classification of an EEG-Based Brain Computer Interface using Wavelet Packet Decomposition and Cepstral Analysis”, MSc Thesis

[50] Reuderink Boris, “Robust Brain-Computer Interfaces”, 2011, PhD Thesis, SIKS dissertation series no. 2001-44

[51] Poel Mannes, 2012, Personal correspondence

[52] Ang K K et al., 2008, “A clinical evaluation of non-invasive motor imagery-based brain-computer interface in stroke”, IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 4178-4181

[53] Lotte F et al., 2011, “Regularizing Common Spatial Patterns to Improve BCI Designs: Unified Theory and New Algorithms”, IEEE Transactions on Biomedical Engineering, Feb 58(2) 355-362

[54] Bennett K. P. et al. , 2000, “Support vector machines: hype or hallelujah?” ACM SIGKDD Explorations Newsletter, Dec 2(2) 1-13

[55] Jain A.K. et al., 2000, “Statistical pattern recognition: A review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Jan 22(1) 4-37

[56] Wang T et al., 2004, “Classifying EEG-based motor imagery tasks by means of time–frequency synthesized spatial patterns”, Clinical Neurophysiology, Dec 115(12) 2744-2753

Appendix A - *Selection of optimal baseline*

Step 1: Signal partitioning and power density spectrum estimation

We have split the EEG signals into trials corresponding to the visual inputs (10 second baseline movies "resting", or 10 second hand movie "active"). The same hand movie was always used. We split trials into three groups, depending on their previous baseline movie. Afterwards we estimated six power density spectra for each channel, using Welch's method (2 second window with 1 second overlap), for each baseline movie and it's corresponding active movie.

Step 2: SMR selection

Channel C3 or C4 was analyzed depending on the dominant hand of the subject. We paired the six PSDs according to the baseline. We selected the most suppressive rhythm (μ or β) via visual inspection.

Step 3: Computation of ERD and optimal baseline selection

For each pair of PDSs, we computed the power over the selected band by taking the area under the PSD. Next, we computed the ERD for each channel according to formula (1.1). Finally, we selected the optimal baseline based on visual inspection of the topographical-ERDs. We chose the optimal baseline to come from the pair that showed the clearest ERD over the contralateral sensorimotor area.

Appendix B – *Common Spatial Patterns*

Let us start by considering X as being a matrix representing the recorded signal on N channels over a period defined as T:

$$S = \begin{pmatrix} x_{1,1} & \cdots & x_{1,T} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,T} \end{pmatrix}. \tag{B.2}$$

We now estimate the spatial covariance matrix of this trial:

$$C = SS' \tag{B.3}$$

where X' is the transpose of X. C can be normalized by dividing each of its elements with the trace of C.This is done to compensate for the magnitude variations in the EEG that exist between individuals. The normalized covariance matrix can now be decomposed into three matrices by means of eigenanalysis:

$$\bar{C} = U\lambda U' \tag{B.4}$$

with λ (N×N diagonal) being the eigenvalue matrix of $\bar{C}$, and U (N×N) containing as its columns the corresponding eigenvectors and with the property that:

$$U'U = I. \tag{B.5}$$

If we apply the transformation U' to X we get a new signal matrix Y for which the covariance matrix is:

$$C_Y = YY' = U'SS'U = U'\bar{C}U = \lambda, \tag{B.6}$$

thus giving the rows of Y the property of being uncorrelated.

We now follow a similar algorithm but take into consideration the fact that we have two classes. The normalized covariance matrix for the two classes is:

$$C_t = \overline{C_a} + \overline{C_b}. \tag{B.7}$$

Then the same step of eigenanalysis is applied:

$$C_t = U_t\lambda_t U_t'. \tag{B.8}$$

In order to completely decorrelate the normalized covariance matrices we need a whitening transform that has the form:

$$W = \sqrt{\frac{1}{\lambda_t}} U_t'. \tag{B.9}$$

So now, the uncorrelated normalized covariance matrices can be written as:

$$R_a = W\overline{C_a}W' \text{ and } R_b = W\overline{C_b}W', \tag{B.10}$$

with the properties that R_a and R_b share common eigenvectors (basic spatial patterns) and the elementwise sum of the eigenvalues will be 1. R_a and R_b may be rewritten as:

$$R_a = B\psi_a B' \text{ and } R_b = B\psi_b B', \tag{B.11}$$

with

$$\psi_a + \psi_b = I. \tag{B.12}$$

In other words, the eigenvectors optimally describe the variance. Another useful property deriving from the above is that the first m eigenvectors will be maximal for class **a**. Also because of (B.12) this means that the variance explained by these eigenvectors must be minimal for class **b.**

Appendix C - *Linear Discriminant Analysis*

Linear Discriminant Analysis [40, 44, 45, 46] is an algorithm that finds a linear hyperplane H_s that separates instances from distinct classes. If the feature vectors are *q* dimensional then the separating hyperplane will be *q-1* dimensional. To simplify let us consider the case where $q = 2$ and data is linearly separable in the feature space. By using a transformation function *f(x)* the LDA algorithm projects the two dimensional feature vectors on to a one dimensional *decision* space:

$$f(x) = w \cdot x + w_0, \tag{C1}$$

where x is the feature vector, w the projection and w_0 the bias or offset. The offset is the distance from the origin to the hyperplane. This means that, by subtracting this value, *f(x)* can be used as a signed distance function to the hyperplane. As such, the predicted label/class, l_x for a new feature vector id given by:

$$l_x = \begin{cases} \text{label\class a, if } f(x) > 0 \\ \text{label\class b, if } f(x) > 0' \end{cases} \tag{C2}$$

The way that the LDA tries to find the optimal separation between classes is by maximizing the ratio of between class variance to within class variance:

$$\sigma^2_{\text{between}} = \sum_{i=1}^{2}(\mu_i - \mu)(\mu_i - \mu)' \text{ and } \sigma^2_{\text{within}} = \sum_{i=1}^{2}\sum_{k=1}^{n_i}(x^k - \mu_i)(x^k - \mu_i)', \tag{C3}$$

where μ is the mean across all instances, μi is mean for class i n_i is the number of features. Figure C1 shows an example of two classes, with their features being the position in Euclidean space that are separated by LDA.

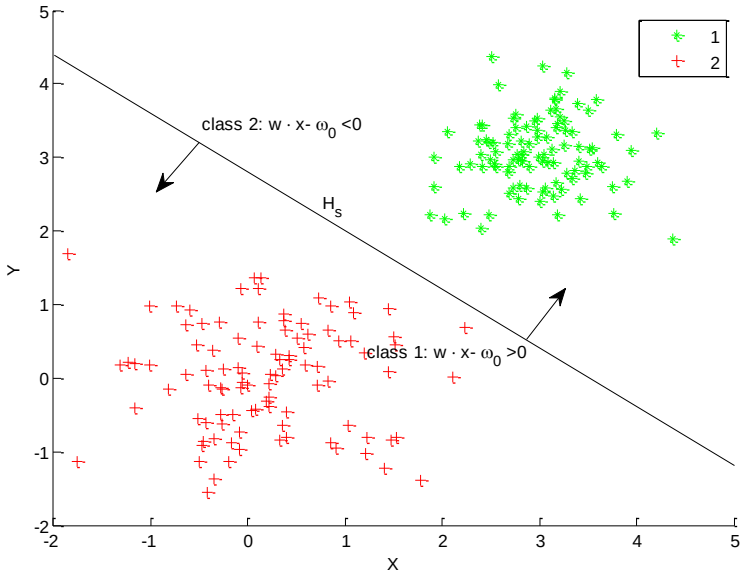


Figure C1: LDA – 2D example; Two classes, 1 and 2, represented in Euclidean space. Hs represents the separating hyperplane found by the LDA. New datapoints that are above this plane will be classified as being from class 1.In our case X and Y represent the power over two channels

Appendix D – *Support Vector Machines*

We start presenting the concept of SVM by using the same example as the one for LDA. The formal description of the SVM algorithm is given after the intuitive presentation of the problem. Figure D1 presents the two classes; because the classes are linearly separable, it means that there is an infinite number of hyperplanes that can do this. The question now is which one to choose? We want to find the one that maximizes the classifiers' performance; intuitively we can say that the best one is at half the distance from the classes. This is called the maximum-margin hyperplane, H_s and it is found by using the aiding hyperplanes H_a and H_b. The first property of these hyperplanes is that they are parallel. The second is that they are as close as possible to the border of one class, in order to still make discrimination possible. The aiding hyperplanes are built by using some datapoints at the border of the classes. These specific data points are called *support vectors*.

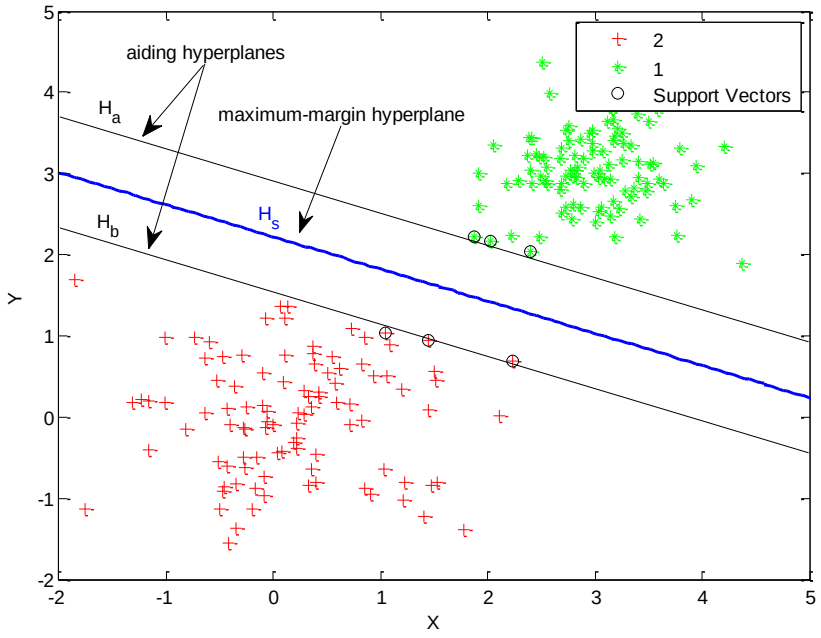


Figure D1: Linear SVM- Two aiding hyperplanes, H_a and H_b, are built with the aid of the support vectors. The optimal hyperplane, H_s, is found at half the distance between H_a and H_b. New datapoints that will be below the maximum-margin hyperplane will be classified as belonging to class2.

Unfortunately, there are cases when the two datasets are overlapping as in Figure D2. This makes it impossible to find a maximum-margin hyperplane. These two instances might be outliers and we could "ignore" them. In other words, we can allow for a degree of misclassification. In this case, the SVM algorithm can build an optimum hyperplane as can be seen form Figure D3. This is called a *soft-margin* SVM.

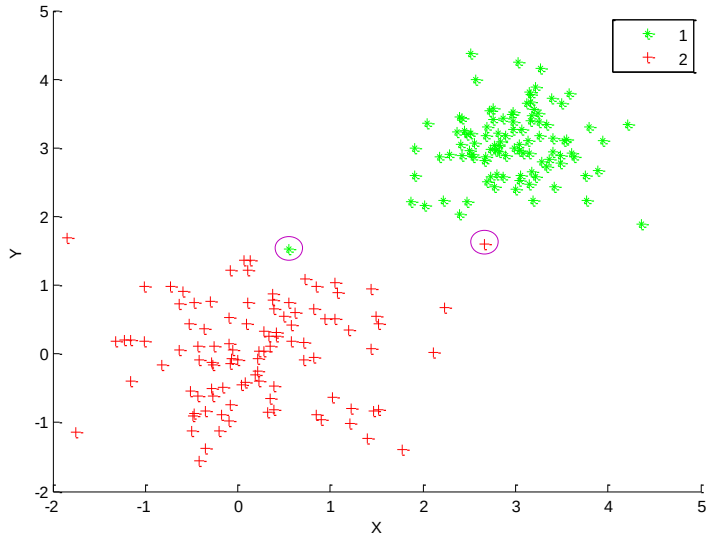


Figure D2: Overlapping datasets - No maximum-margin hyperplane can be found because of the two instances circled in purple

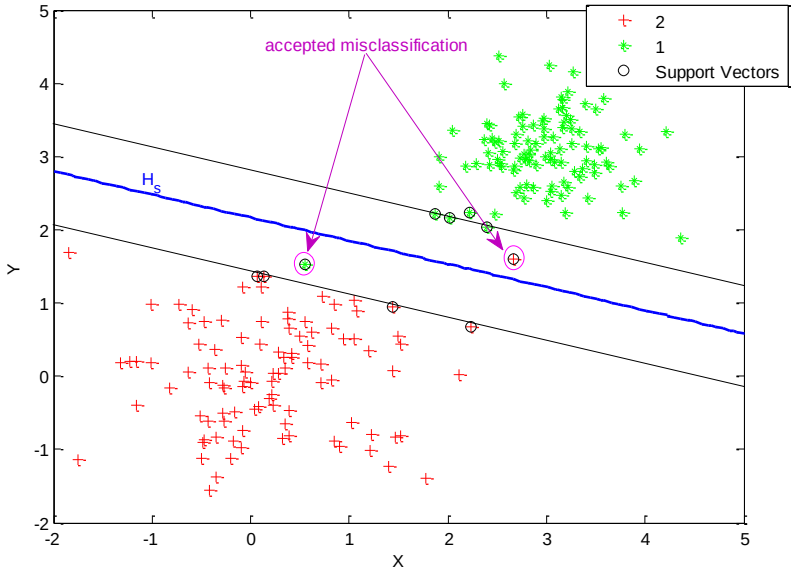


Figure D3 :Soft-margin SVM – By allowing some degree of misclassification the SVM can still find the maximum-margin hyperplane, Hs.

The case portrayed in Figure D2 is a fortunate one, but what is to be done in a case such as one from FigureD4 a)? There is no linear hyperplane that can separate the data and perform better than random. In such a case, it would seem that SVM cannot help us and the only way to classify the data is to go for a non-linear classifier such as artificial neural networks.

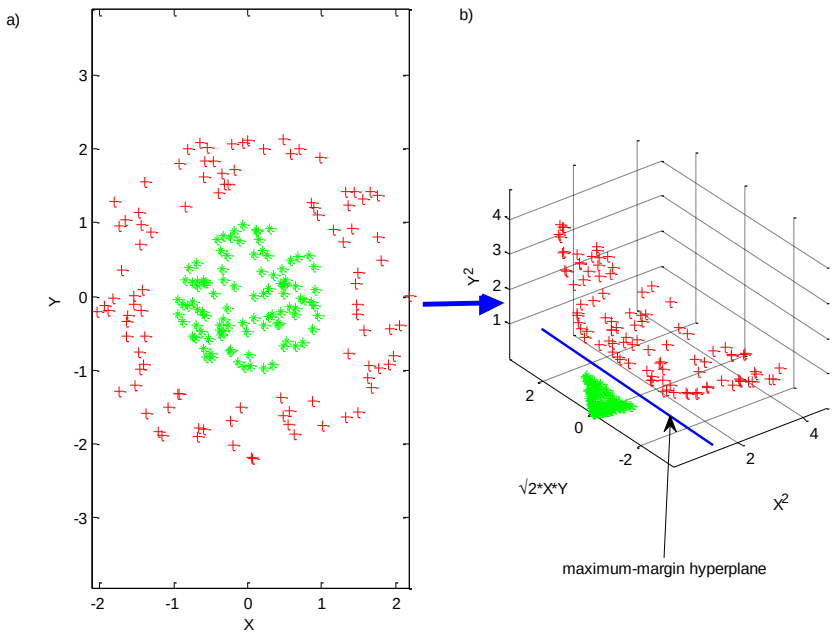


Figure D4: a) Nonlinearly separable dataset in the original 2D feature space b) The same data as in a) but projected into a higher dimensional space. In this space, a linear maximum-margin hyperplane can be found.

Fortunately, the Cover theorem states that [42]: "*A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated*". This means that in formula (C21), from Appendix C we can substitute the term $x_i \bullet x_j$ by $\varphi(x_i) \bullet \varphi(x_j)$, where φ: $\mathbb{R}^q \rightarrow \mathbb{R}^m$ is a non-linear feature map and q being the initial dimensionality. For example let's take our case: φ: $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ with $(X,Y) \rightarrow (Z_1,Z_2,Z_3) = (X^2, \sqrt{2}XY, Y^2)$ (other mappings are possible as well). The data projected into this new space is shown in Figure D4 b). It is clear that in this space exists a linear maximum-margin hyperplane that separates the data. If we take this hyperplane and map it back to the original space, we get the separation plane shown in Figure D5.

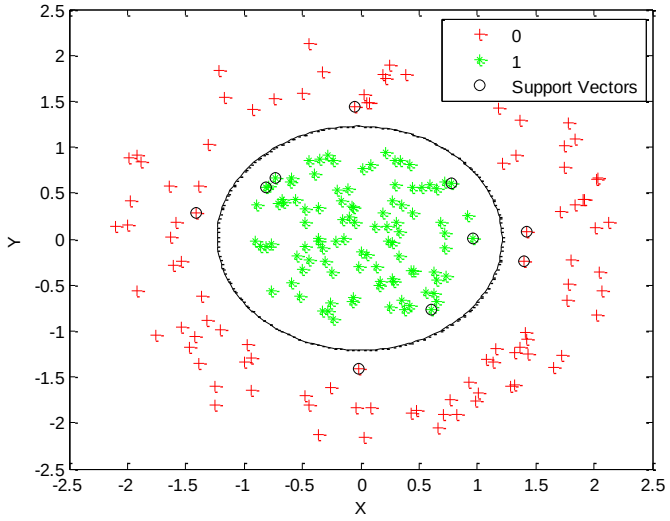


Figure D5: Nonlinear decision boundary that perfectly separates the data.

We can now rewrite $\varphi(x_i)\bullet\varphi(x_j)$ as $K(x_i,x_j)$, called a kernel function. We do this because it enables us to compute formula (D21) without knowing the mapping. We only need to be able to compute the inner product.

Formal Description

Linear SVM

Let us then take *X* to be the total set consisting of *n* feature vectors that are *q*-dimensional:

$$X = \{(x_i, l_i) | \, i = \{1, \dots, n\}, l_i \in \{-1,1\}, \, x_i \in \mathbb{R}^q\} \, . \tag{D1}$$

It was stated in the hypothesis that the classes are linearly separable which means that there are an infinite number of *q-1* dimensional linear hyperplanes that can accomplish this. It is desired to find the hyperplane that maximizes the classifier's performance – the maximum margin hyperplane H_s. This hyperplane is found with the use of two other hyperplanes, H_a and H_b. The first property of these hyperplanes is that they are parallel. The second is that they are as close as possible to the border of one class, in order to still make discrimination possible. Then H_s is located in the middle with equal distance to the aiding hyperplanes.

This can be formally written the following way:

$$w \cdot x - w_0 = 0 \, , \tag{D2}$$

where *w* is a vector that is normal to H_s, *x* a point on the hyperplane, and $\frac{w_0}{\|w\|}$ is the normalized perpendicular distance from the origin to the hyperplane. The feature vectors that are passed by

the aiding hyperplanes H_a and H_b are called *support vectors*. As it is known that the aiding hyperplanes are at an equal distance *d* from H_s then this can be written as:

$$w \cdot x_i - w_0 = +d\,, \tag{D3}$$

for the support vectors that are passed by H_a and

$$w \cdot x_i - w_0 = -d, \tag{D4}$$

for the support vectors passed by H_b. The following constrains are put so that there are no points that fall between the aiding hyperplanes:

$$w \cdot x_i - w_0 \geq +r \text{ for } l_i = +1, \text{ and} \tag{D5}$$

$$w \cdot x_i - w_0 \leq -r \text{ for } l_i = -1. \tag{D6}$$

Equations (B5) and (B6) can be merged to give:

$$l_i(w \cdot x_i - w_0) \geq r \text{ for } i = \{1,\ldots, n\}\,. \tag{D7}$$

Now it easy to observe that the in order to maximize the margin $\frac{2d}{\|w\|}$ between the aiding hyperplanes we need to minimize $\|w\|$ with the constrains imposed by (B7). To do this we scale *w* and *b* with a factor of d^{-1}. This scaling has the advantage that it keeps the distance from any point x_i to H_s is unchanged and now the distance between H_a and H_b is now $\frac{2}{\|w\|}$. This means that equation (B7) can be rewritten as:

$$l_i(w \cdot x_i - w_0) \geq 1 \text{ for } i = \{1,\ldots,n\}\,. \tag{D8}$$

Minimizing $\|w\|$ is equivalent to minimizing $\frac{1}{2}\|w\|^2$ [45] so it now becomes a standard quadratic optimization problem. We can now use Lagrange multipliers α to change the optimizing problem to a form whose complexity does not depend on the dimensionality of the feature vector, *q*, but rather on the number of training instances *n* [46]:

$$\begin{aligned} L_p &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [l_i(w \cdot x_i - w_0) - 1] \\ &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i l_i (w \cdot x_i - w_0) + \sum_{i=1}^{n} \alpha_i \end{aligned} \tag{D9}$$

To solve this optimization problem we then need to minimize with respect to w and w_0, and maximize with respect to $\alpha_i \geq 0$. This can be done by calculating the gradients of L_p with respect to w and w_0.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i l_i x_i \quad \text{(D10)}$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i l_i = 0 \quad \text{(D11)}$$

By substituting (B10) and (B11) in (B9) we get the dual form, L_d, that is only dependent on α:

$$L_d = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j l_i l_j x_i \cdot x_j - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j l_i l_j x_i \cdot x_j + \sum_{i=1}^{n} \alpha_i c_i w_0 + \sum_{i=1}^{n} \alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j l_i l_j x_i \cdot x_j + \sum_{i=1}^{n} \alpha_i \quad \text{(D12)}$$

This form besides only being dependent on α will also allow the use of kernel-based methods. The *n* solutions for α can now be found by means of quadratic optimization methods. We observe that most solutions will be $\alpha_i = 0$ and only a small part of them will have $\alpha_i > 0$. The set of x_i, X_{SV}, with this property are the support vectors. In other words, we can now compute w by using (B10) and all we need now is find w_0. We know from (B8) that:

$$l_i(w \cdot x_i - w_0) = 1\,, (x_i, l_i) \in X_{sv} \quad \text{(D13)}$$

and by putting (B10) into (B11) we get:

$$c_i\left(\sum_{i=1}^{n} \alpha_i c_i x_i - w_0\right) = 1\,, (x_i, l_i) \in X_{SV} \quad \text{(D14)}$$

Now we multiply by c_i, and take into account that $c_i^2 = 1$ we obtain:

$$w_0 = l_i - \sum_{j \in X} \alpha_j l_j x_j \cdot x_i\,, (x_i, l_i) \in X_{SV} \quad \text{(D15)}$$

We observe then that w_0 can be calculated using any support vector; for numerical stability w_0 should be computed for every support vector and the average taken. We can now compute the signed distance function for a new input feature vector $\hat{x}$:

$$y = f(\hat{x}) = w \cdot \hat{x} + w_0 \quad \text{(D16)}$$

with *y* being what is called the *support vector machine*. The predicated class for $\hat{x}$ will be:

$$l_x = \begin{cases} \text{label\class a}, \text{if } y > 0 \\ \text{label\class b}, \text{if } y > 0 \end{cases} \quad \text{(D17)}$$

As it was said in the hypothesis, this method works only for linearly separable datasets. In order to tackle overlapping datasets we need a *soft margin* SVM.

Soft margin SVM

The problem of overlapping datasets is covered in [39]. The solution they propose is to allow the mislabeling of features in the training set and assigning a penalty to them. A *slack variable* $\xi \geq 0$ is introduced to measure the deviation from the margin, i.e. the *degree of misclassification*. This is introduced in equation (B8):

$$l_i(w \cdot x_i - w_0) \geq 1 - \xi_i \text{ for } i = \{1, \ldots, n\}, \xi_i \geq 0 \quad \text{(D18)}$$

We can now introduce (B8) into (B9) and we get :

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[l_i(w \cdot x_i - w_o) - 1 + \xi_i] - \sum_{i=1}^{n}\mu_i\xi_i \quad .$$

$$= \frac{1}{2}\|w\|^2 \; C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i l_i(w \cdot x_i - w_0) + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i \; - \sum_{i=1}^{n}\mu_i\xi_i \quad \text{(D19)}$$

where μ_i are the new Lagrange parameters that constrain $\xi \geq 0$, and C is the cost parameter that trades off the number of support vectors and the number of non-separable points. C may be adjusted to favor one of the two (increased margin vs. decrease of data misfit). The gradients for w and w_0 are the same as before and the gradient for ξ_i is:

$$\frac{\partial L_p}{\partial \xi_i} = C - \xi_i - \mu_i = 0 \Rightarrow C = \xi_i + \mu_i \quad \text{(D20)}$$

Following the same algorithm as before the dual form is:

$$L_d = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j l_i l_j x_i \cdot x_j + \sum_{i=1}^{n-1}\xi_i(\alpha_i + \mu_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j l_i l_j x_i \cdot x_j + \sum_{i=1}^{n}\alpha_i c_i w_0 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\xi_i\alpha_i - \sum_{i=1}^{n}\xi_i\mu_i$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j l_i l_j x_i \cdot x_j + \sum_{i=1}^{n}\alpha_i \quad \text{(D21)}$$

As one can notice, equation (B21) is the same as equation (B12) so the solutions for all the parameters will be the same. Therefore, the signed distance function will also have the same form as in (B16). By using this extension of the SVM, we can now classify overlapping datasets but we still cannot tackle datasets that are not linearly separable at all.

Non-linear SVM

The beauty of classifying nonlinearly separable datasets with SVM is that the algorithm it uses does not try to fit a nonlinear model directly to the feature space [42,43]. What it does instead is map the dataset to a new space where the two classes are linearly separable and discriminates

them with the maximum margin hyperplane. After that, it projects this hyperplane into the original feature space where it is represented by a nonlinear discriminating hyperplane.

This algorithm makes use of kernels to accomplish this task. If we look at the first term of dual form,L_d, we notice that there is a dot product, $x_i \cdot x_j$, that does not perform a projection to a different feature space. This is called a linear kernel:

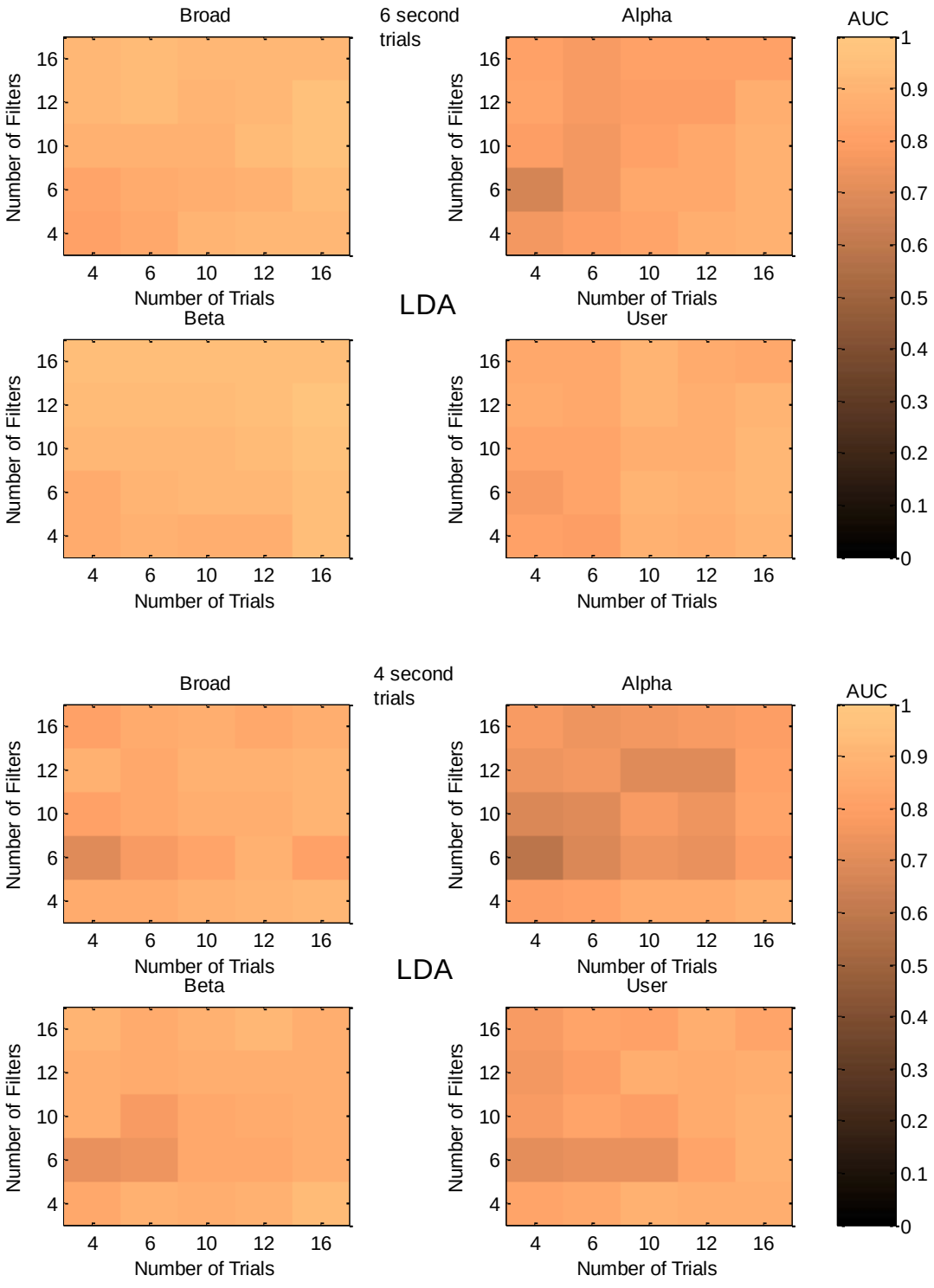
$$K(x_i, x_j) = x_i \cdot x_j \tag{D22}$$

Replacing this kernel with a nonlinear one in the dual form will accomplish the algorithm described before.

The same steps as in the previous sections are taken to reach classification. Except now, we start by projecting each feature vector to the new feature space. This is done with kernel mapping $x \rightarrow \varphi(x)$, where $\varphi(x)$ is the chosen kernel function. In this case the singed distance function will be:

$$y = f(\hat{x}) = w \cdot \varphi(x) + w_0 \tag{D23}$$

Appendix E - *CSP grid search*



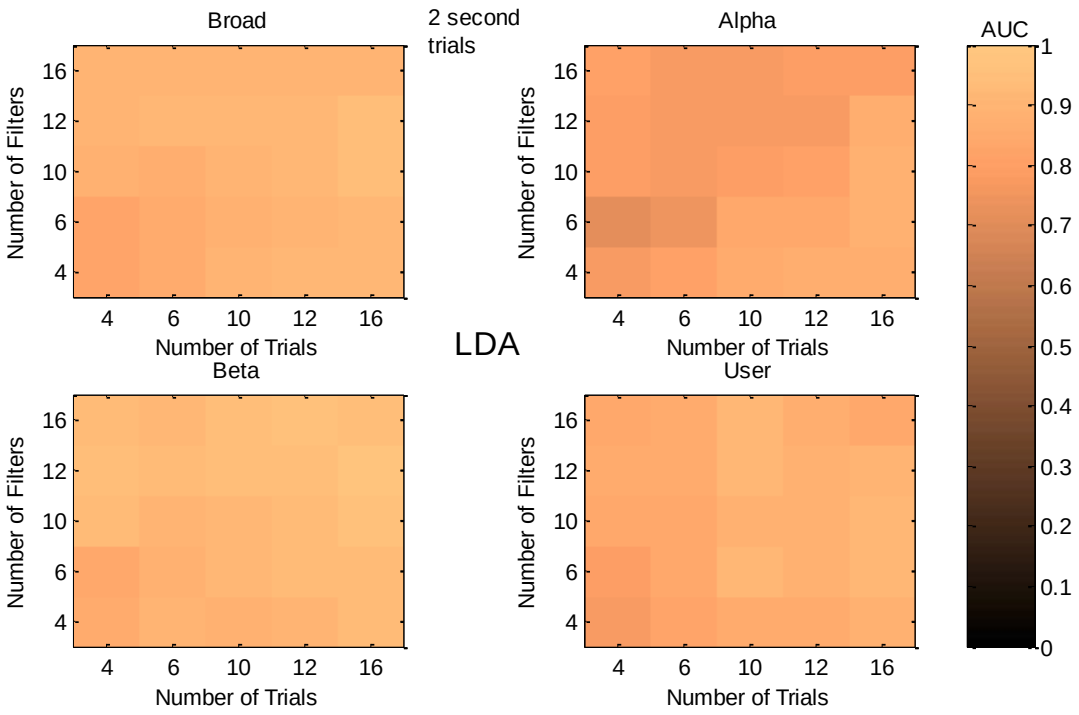
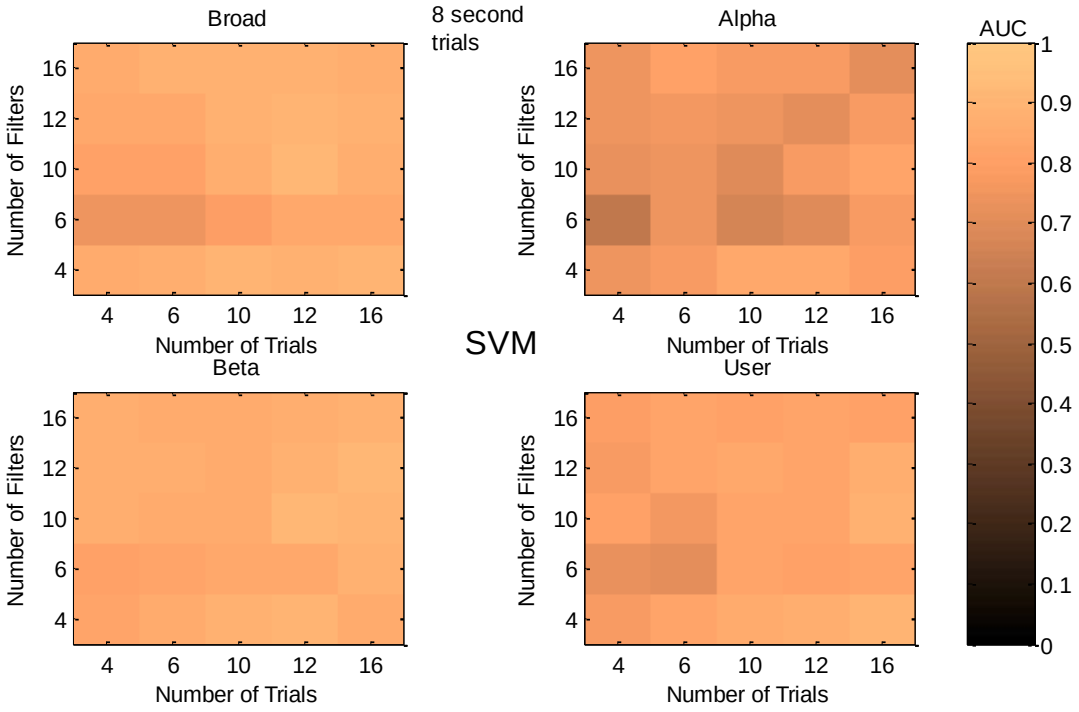
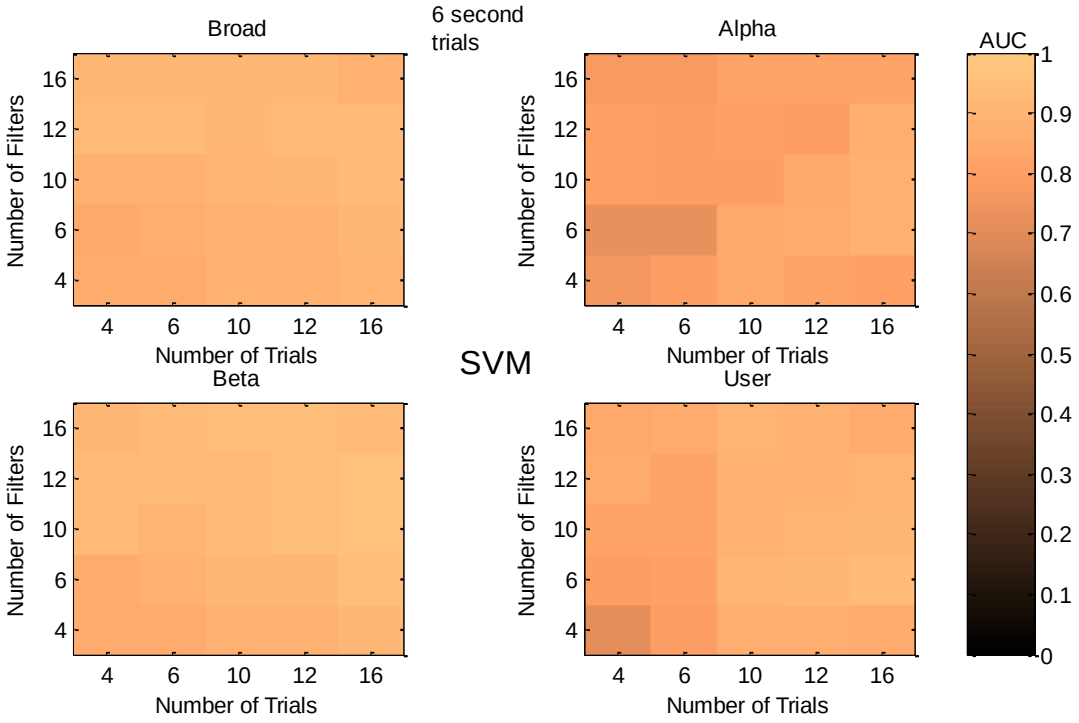
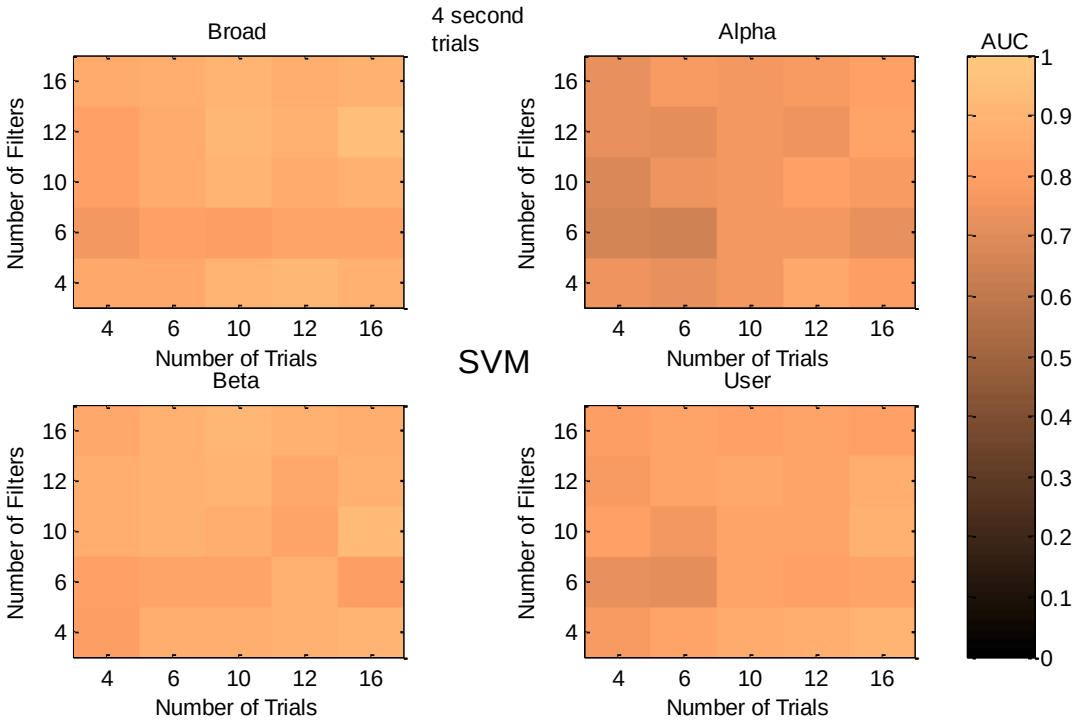


Figure E1: Grid search across all number of trials and filters for all trial lengths(except 8 seconds) for LDA. The lighter the color tones the higher the value of the AUC.




6 second trials
Broad
Alpha
Beta
User
SVM
AUC
Number of Filters
Number of Trials
4
6
10
12
16
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


4 second trials
Broad
Alpha
Beta
User
SVM
AUC
Number of Filters
Number of Trials
4
6
10
12
16
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

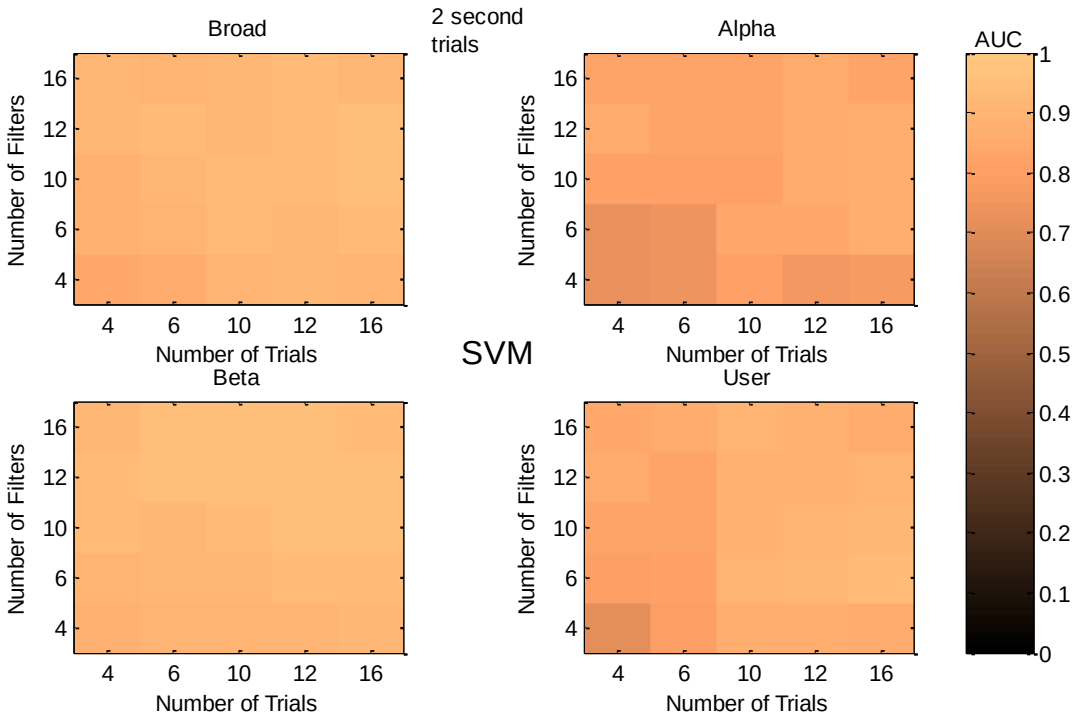


Figure E2: : Grid search across all number of trials and filters for all trial lengths for SVM. If the color tones are very similar across one grid it means that it does not matter what number of trials or filters we choose.

Appendix F - *Detailed results stroke subject S09*

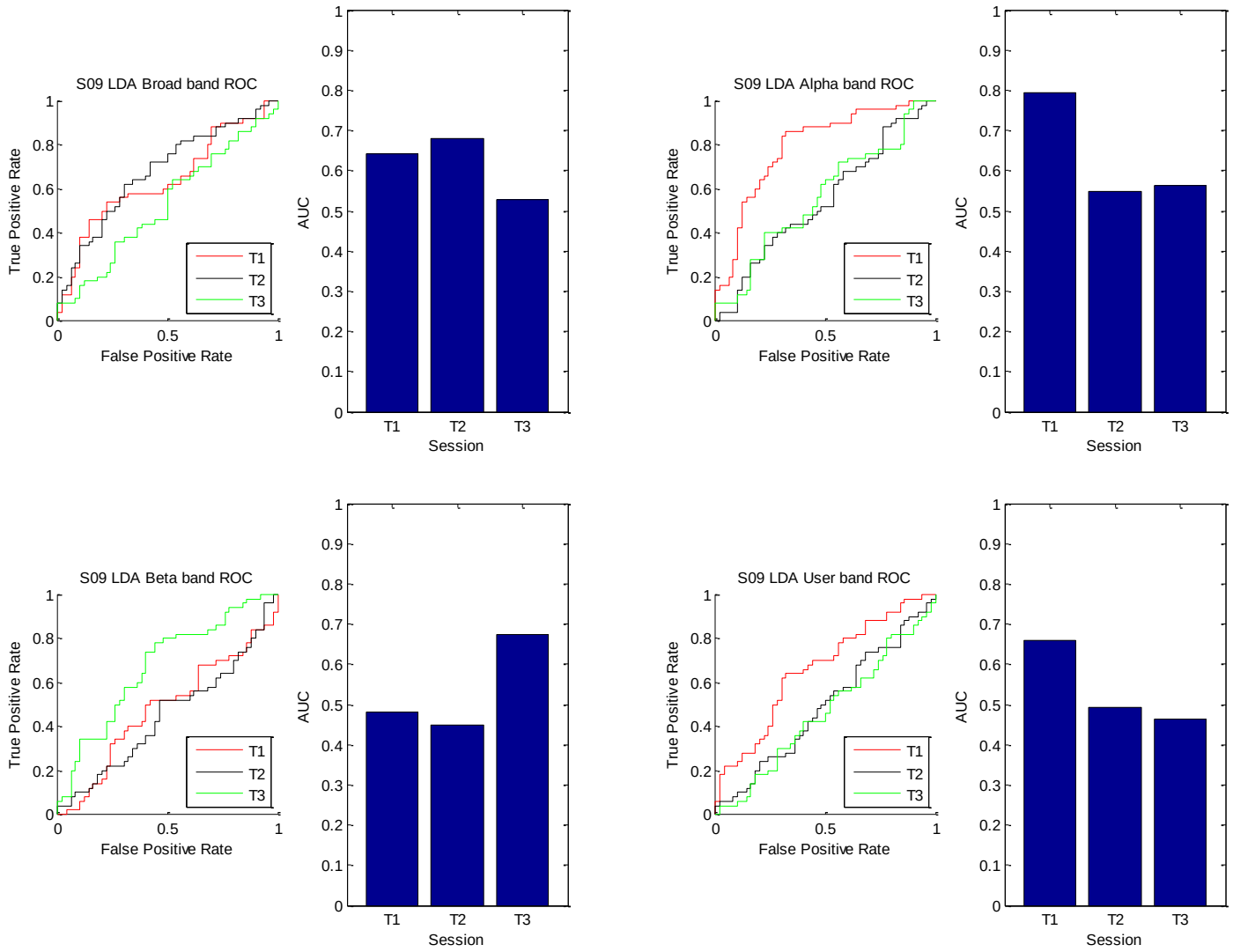


Figure F1: LDA – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S09

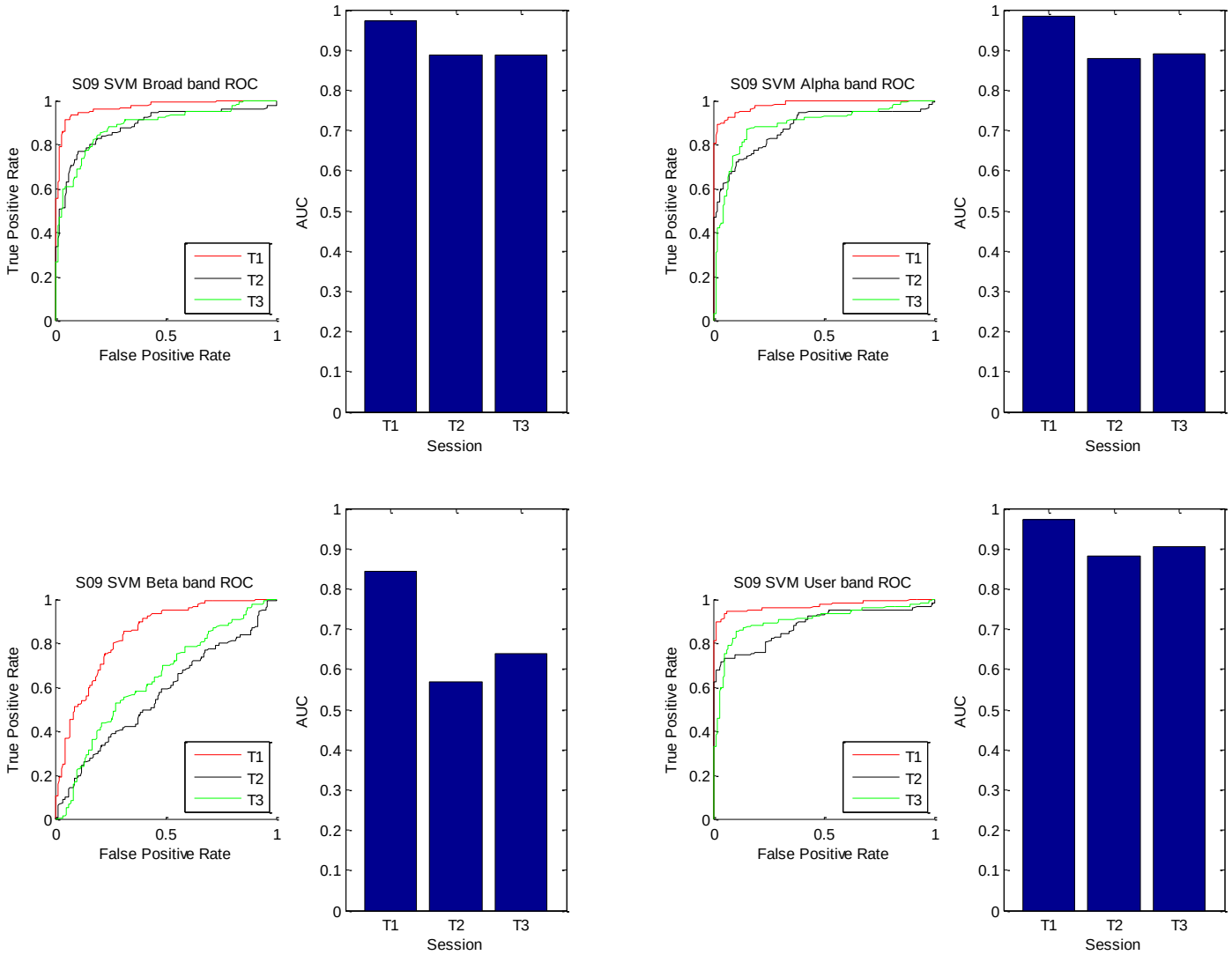


Figure F2: SVM – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S09

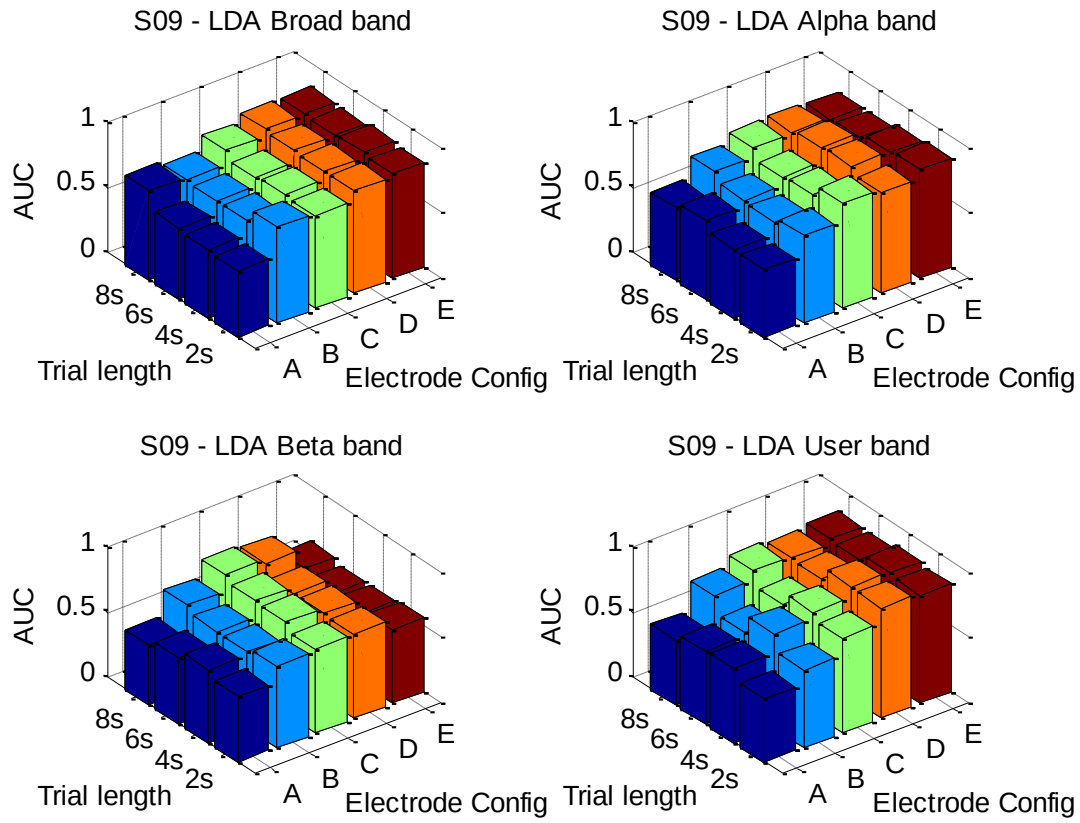


Figure F3: LDA AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S09

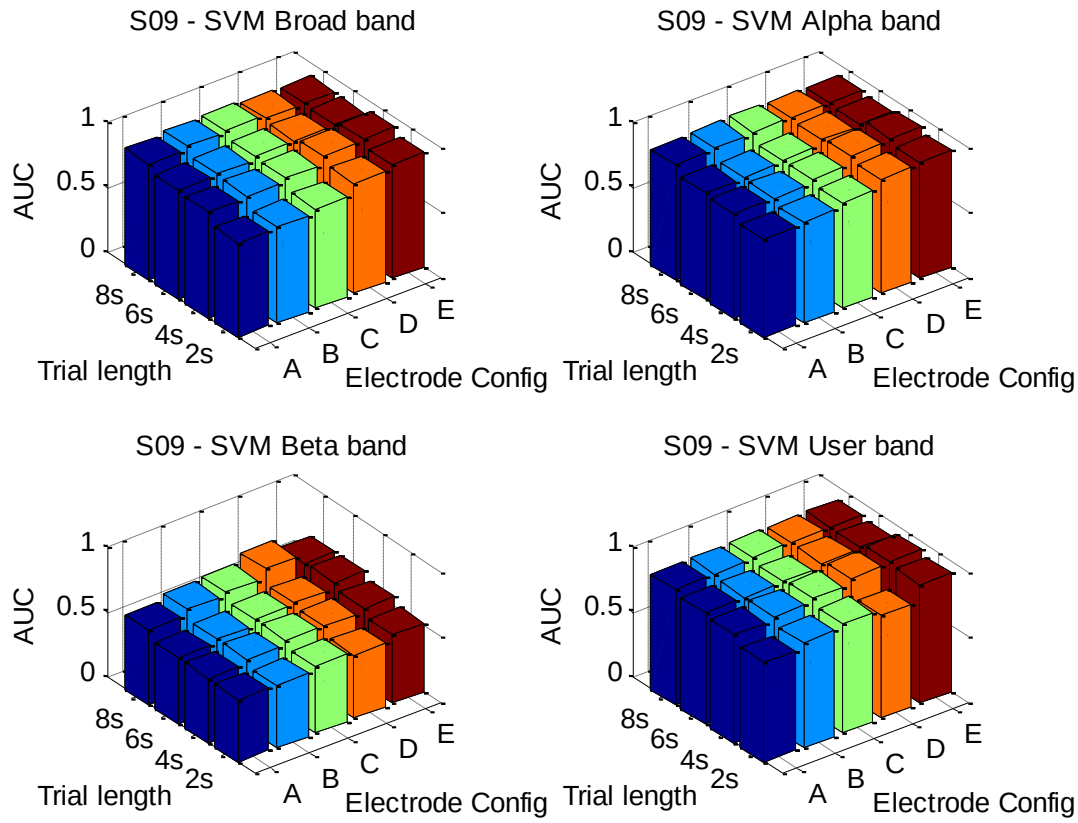


Figure F4: SVM AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S09

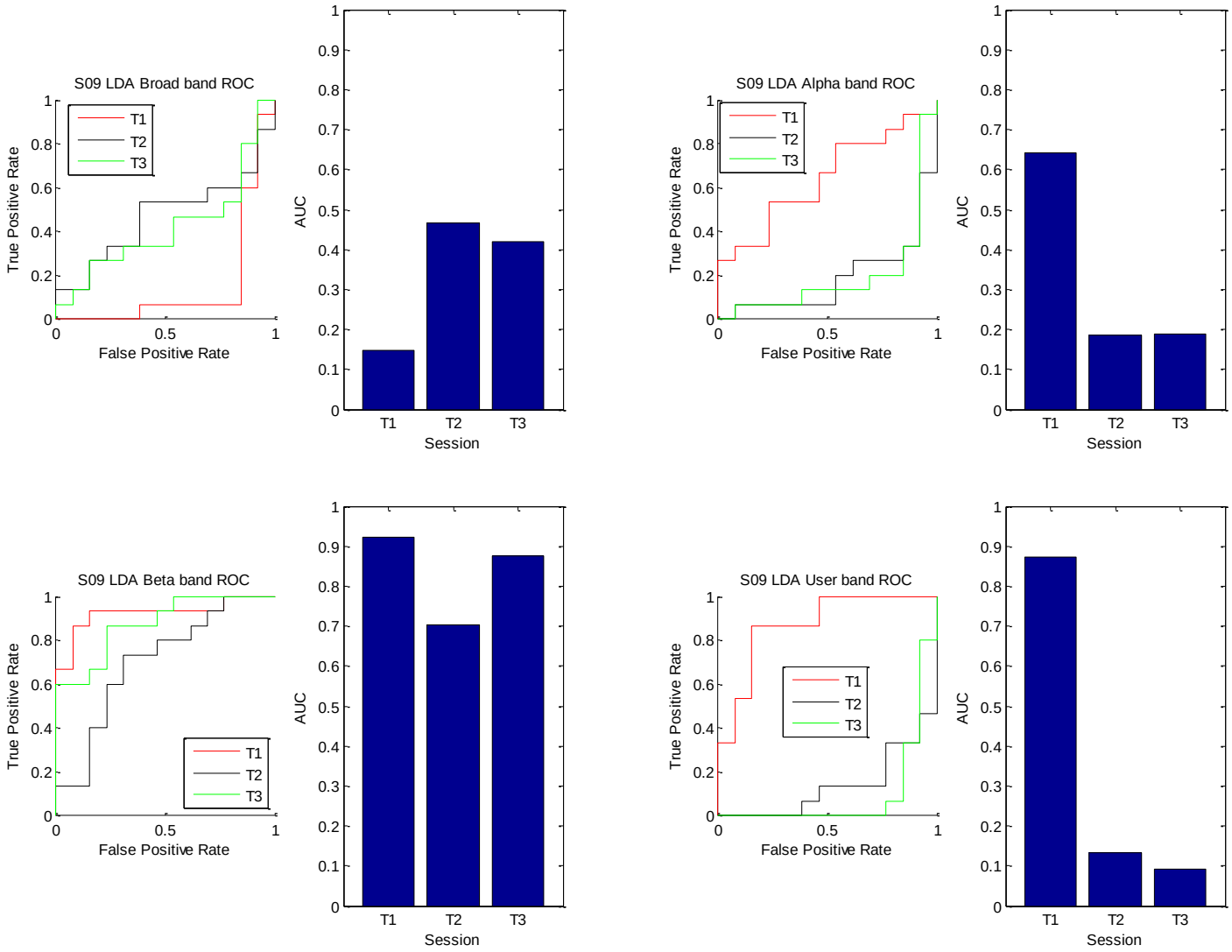


Figure F5: CSP + LDA – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S09

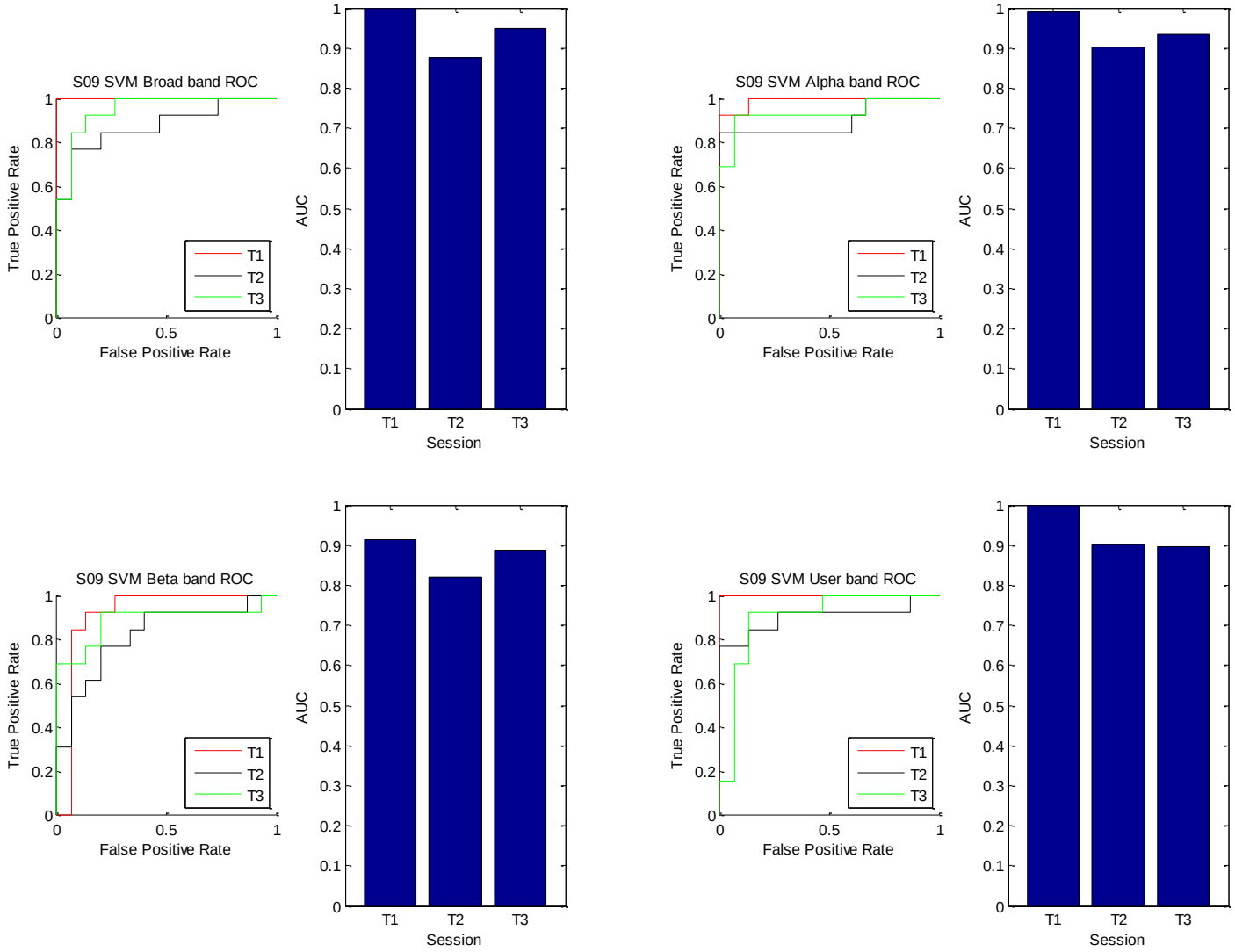


Figure F6: CSP + SVM – ROC and corresponding AUCs for 8 second trials, all runs, and all frequency bands in subject S09

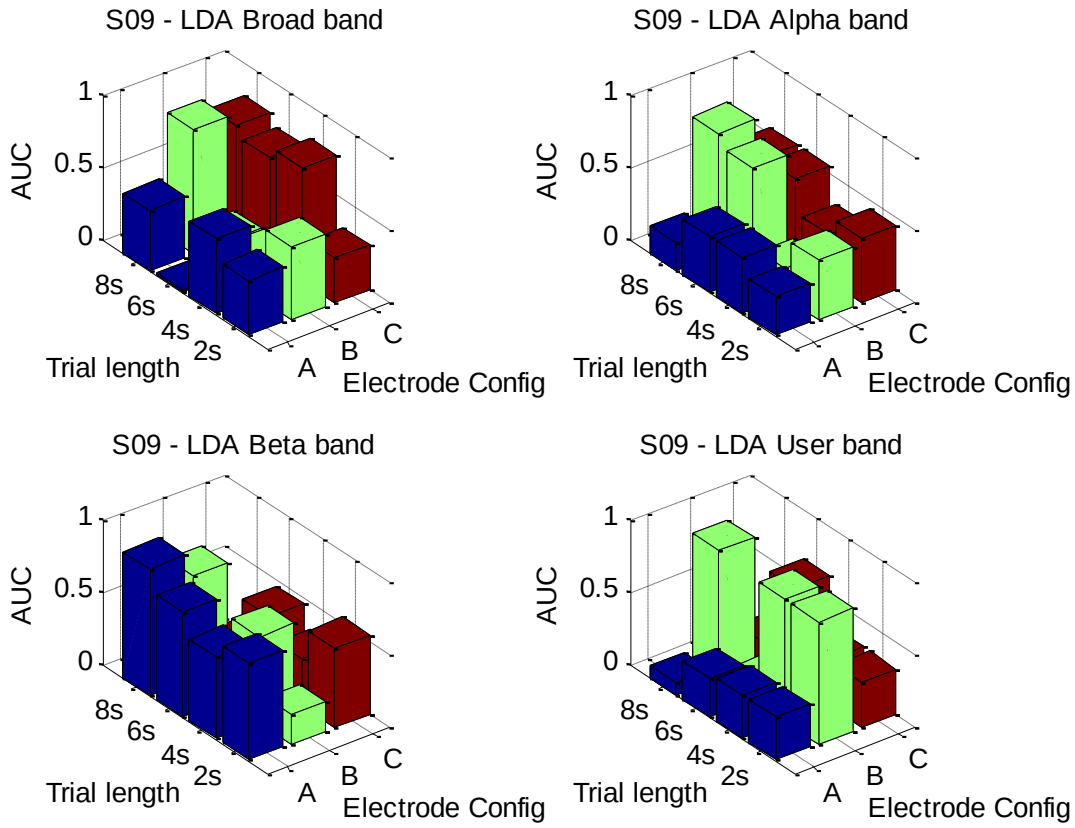


Figure F7: CSP + LDA AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S09

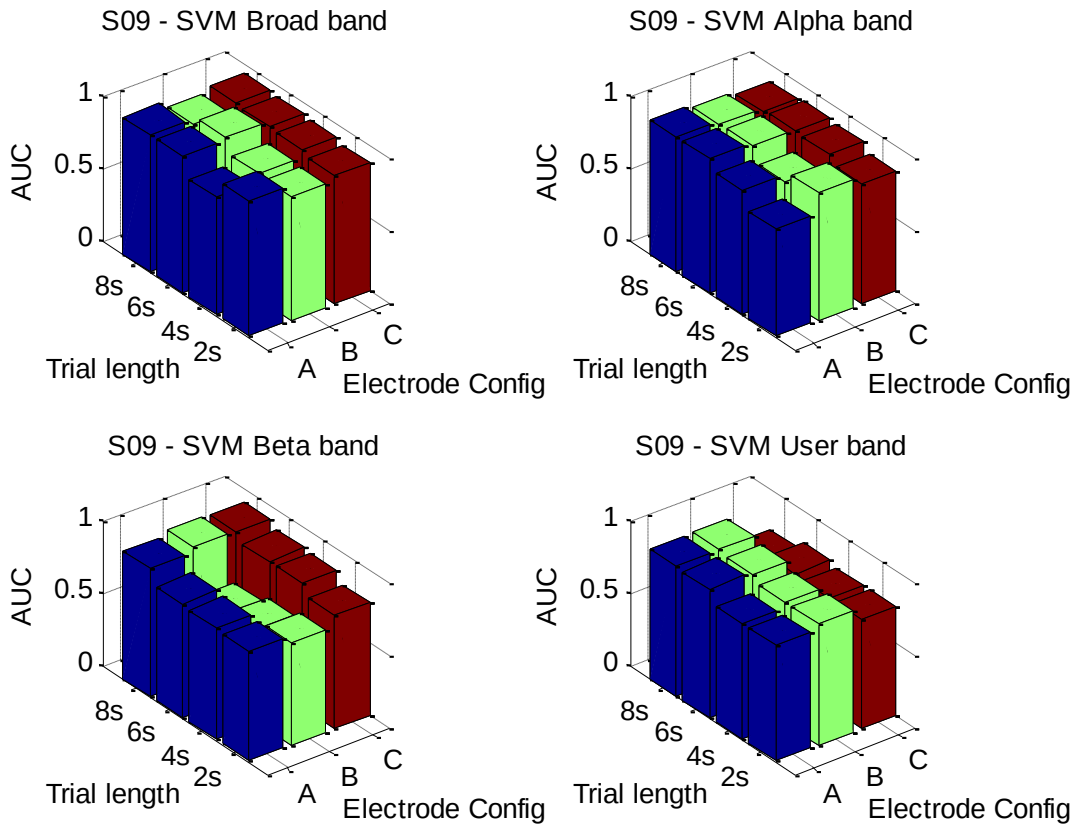


Figure F8: CSP + SVM AUCs for all trial lengths and all electrode configurations for run T2 in stroke subject S09