# Curiosity Cloning

## Neural Modelling for Image Analysis

**Authors:** Ashkan Yazdani[1], Frederic Dufaux[1], Thien M. Ha[1], Touradj Ebrahimi[1], Dario Izzo[2], Christos Ampatzis[2]
**Affiliation:** [1]Polytechnique Fédérale de Lausanne, [2]Advanced Concepts Team, European Space Agency.

**Date: 15/01/2010**

**Contacts:**

Touradj Ebrahimi, Polytechnique Fédérale de Lausanne
e-mail: Touradj Ebrahimi <touradj.ebrahimi@epfl.ch>

Advanced Concepts Team
Fax: +31(0)715658018
e-mail: act@esa.int

Available on the ACT website
http://www.esa.int/act

**Ariadna ID:** 08/8201a
**Study Type:** Standard
**Contract Number:** 21949/08/NL/CB

# Curiosity Cloning - Neural Modelling for Image Analysis (Final Report)

Ashkan Yazdani[1], Frédéric Dufaux[1], Thien M. Ha[1], Touradj Ebrahimi[1]
Dario Izzo [2], Christos Ampatzis[2]

[1]Ecole Polytechnique Fédérale de Lausanne
[2]Advanced Concepts Team, European Space and Technology Research Center (ESTEC)

[1] {ashkan.yazdani,touradj.ebrahimi}@epfl.ch
[2] {Dario.Izzo, Christos.Ampatzis}@esa.int

03/02/2010

**Abstract**

Curiosity is an emotion in human and animals, which motivates them to explore their environment for the purpose of learning, investigation and gaining information. This emotion can be replicated in artificial intelligent systems so that artificial agents such as robots or specific programs can show the similar behaviour when they observe a phenomenon that is new and unexpected to them. If one or several physiological markers can be proven to have strong correlation with curiosity arousal, they can be used for the purpose of elaboration of an artificial curiosity module in machines. For instance, space exploration robots in extra orbital mission can benefit from such module and can select the scientifically relevant images among the immense image datasets that they autonomously collect using several high-definition sensors. Consequently they will be able to discard the rest of images and transmitting only the selected images back to earth. One of the physiological signals that is known to alter during some emotions such as surprise and curiosity is the brain electrical. In this research, electroencephalogram signal is analyzed to study whether it can be beneficial for the purpose of curiosity detection. To this end, several experiments have been performed. It has been shown that for most of the subjects, a strong correlation exists between the curiosity arousal and amplitude of P300 component of the EEG signal.

# Contents

# Chapter 1

# Introduction

*NOTE: The content of this report also appeared (partially) in published scientific papers [1] and part of it has been submitted for publication to peer-reviewed journals. The work here reported was part of a project involving another research group who performed and analyzed the same experiments independently [2]. As a consequence the text in the early chapters (Introduction and Experiments) is similar to that appearing in the report detailing the results from the other group.*

Autonomy is not only a highly desirable, but it is also a critical technology in many scenarios of space exploration. For example, the communication delay between a rover exploring Mars and the Earth can vary, depending on the relative positions of the planets around the sun, in the order of several minutes. Clearly, there may be critical situations where the rover needs to display autonomous decision-making to overcome obstacles, avoid hazards or even to pick up a sample from the ground, or take an image from the environment and so on.

Operational examples of on-board autonomy include the selective acquisition of interesting images. The pioneering Autonomous Sciencecraft Experiment (ASE) [3] has been active on-board the Earth Observing-1 mission since 2003. The EO-1 mission has developed and validated unique technology, and more specifically instruments, for Earth observation. ASE allows the spacecraft to autonomously identify and monitor scientifically interesting events observed on the Earth from the satellite's optical payload, by making use of algorithms rooted in data analysis and planning and scheduling. In particular, the image analysis software of ASE aims at extracting static features characteristic of regions such as land, ice, snow, water, and thermally hot areas, while allowing the detection of dynamic features such as regions of change or activity (e.g., floods, ice melt, lava flows etc.). After static features are extracted using threshold-based classification and after dynamic events are detected by comparing spectra changes across consecutive images, a score is attributed to each detected event. Subsequently the planning and scheduling software prioritizes data for downlink, discards data or reschedules the spacecraft for subsequent observations connected to the observed phenomena. More recent work has been focusing on applying more sophisticated machine learning algorithms, such as Support Vector Machines (SVMs) for the on-board detection of certain phenomena, such as active sulfur springs [4]. Despite the very limited available spacecraft processing power and memory, the difficulty of the task related to the slight sulfur signature and the very few training examples to perform supervised learning, the results are very encouraging. In 2007, the two NASA MER rovers Spirit and Opportunity received an update which aimed at endowing them with some basic autonomy based on pattern matching. More specifically, the update made them able to detect dust-devils and clouds in the Martian landscape [5]. This constituted the first on-board science analysis process on Mars, and so far the only example

of selective data acquisition by exploratory rovers. The algorithm (still in use) is essentially based on the detection of changes between subsequent pictures and works well whenever the acquisition campaigns are run in still conditions. The picture interest is, for the two rovers, thus related to the amount of moving objects in the picture itself. While these examples are a big step forward in the technological development of intelligent space agents displaying autonomous properties, much more research effort is needed to understand fully the implications and also the challenges involved in providing space agents with autonomous decision-making capabilities.

A key point, at the center of current technological developments, is the design of algorithms able to classify sensor readings (e.g. images) according to their degree of scientific interest. The main difficulty lies in the definition of what is scientifically interesting. Such a definition not only encompasses expected and known aspects, but also a rough definition of the unexpected. Typically, researchers involved in creating such algorithms would first define the characteristics that a scientifically interesting image possesses (e.g. a dust devil, a given rock type or a cloud) and then use this definition to let the agent decide upon an image interest. More specifically they would train machine learning algorithms (e.g. based on pattern matching) to detect predefined features considered to be scientifically interesting by the experts they talked to. This a-priori definition of scientifically interesting elements in a given image works well enough to recognize what we expect and already know to be interesting, but it fails to detect anything that falls out of those defined boundaries. Contrary to such an approach, machine learning algorithm could be trained directly to classify what is scientifically interesting and what is not, without knowing further information on these two very broad classes. This could potentially allow for broader and more fuzzy classification borders, which could result in algorithms able to return not only the strictly defined and expected, but also a set of images with potentially unexpected, but relevant properties. The challenge when following this approach becomes how to then create a training set for a classifier. One option is to resort to what is typically referred to as the interviewing or interrogation technique. Expert scientists would be interviewed on a particular set of pictures (as for example in [6]), being asked to simply classify or rank them; subsequently a computer would be trained to have a similar response to the one of the interviewed scientist. In this way, the computer has to automatically extract the relevant features that guided the expert's decision-making and learn to use them in such a way so to mirror the expert's classification.

Despite the simplicity of such a methodology, there are various drawbacks involved. For example, it requires the scientists to undergo long and time-consuming sessions of image classification that may prove to be particularly tiring and cumbersome, which in turn can result in the acquisition of a noisy training set. Moreover, this approach is subject to the fuzziness of the scientist's reasoning when placing a highly cognitive judgment upon each picture. In other words, the scientist will repeatedly consciously filter the image, eventually merging even contradictory verdicts to one binary classification or a ranking. In the following, we present the rationale that lies behind and the implementation of an alternative approach to creating such a training set for a classifier; in particular, the information about the expert's classification is extracted directly from the classification of his/her brainwaves.

# Chapter 2

# Motivation

It is well known from neurophysiological studies that when subjects look at images which arouse mental response, their parietal cortex is excited in a very characteristic way: a synchronized peak in the global electrical activity of large groups of neurons in the parietal area arises after approximately 300 ms after the stimulus (image) presentation. This electrical activity can be recorded with an Electro-EncephaloGraphy (EEG) instrument as an electric positive potential wave and is commonly referred to as P300 (see [7] for a good introduction to the P300 wave). The P300 as an event-related potential (ERP) shows interesting features: its magnitude is associated with the level of attention the stimulus arouses (i.e. with the difficulty of the classification task), it cannot be fine controlled, and it is reported to be, at least partially, independent from consciousness.

We propose to extract the picture rating information using the EEG signal recorded while the expert is presented with the pictures in a Rapid Serial Visual Presentation (RSVP) experiment. Our set-up is inspired by related work performed by Gerson et al. [8] who present an original experiment where a simple image ranking task is performed by ranking images according to the amplitude of the P300 brainwave recorded during a RSVP experiment based on the oddball paradigm. The oddball paradigm refers to experimental setups where a target stimulus is presented among more frequent background stimuli. The results of the experiment, and the vast pre-existing literature available on the detection and use of the P300 wave for different applications, suggests the potentiality of processing EEG signals recorded during a RSVP experiment using machine learning techniques to extract a classification of the images mirroring the classification the subject made when presented with the images. In other words, the EEG signal could be used to define image classes rather then having the scientist analyze and explicitly perform such a classification.

We aim to demonstrate that correlating the level of attention with the corresponding sensorial stimulus, it is possible to assign a scientific interest level to the stimulus presented. Moreover, since the P300 shows attention arousal at its very beginning, it is possible to classify the interest-level of an image quicker than by directly interviewing the subject, and to remove any bias operated by the subject's conscious filtering. In order to train an appropriate machine learning algorithm, it is necessary to gather and process a large set of images. Therefore, reducing the time dedicated to the analysis of an image can have drastic effects on the total time required for the processing of the total dataset. Thus, if the suggested approach ends up successful, it will have distinct advantages. The scientist should be able in principle to process a larger number of images per minute, their final classification would potentially be less prone to contain fuzziness, that is, it might be less susceptible to conscious filtering, and the reliability of the classification could thus be potentially be much higher.

Looking for interesting features is looking for the unexpected, highly unusual, or odd. In

other terms, scientific interest is associated with the picture's features which arouse speculation, interest, or particular attention. In this research, the data set obtained by evaluating P300 signals associated with each picture, are later used to train and test a classifier which reacts to stimuli showing the same level of scientific attention that had been monitored from the scientists. In short, we present an experiment where scientists' scientific attention is somehow replicated—or "cloned"–into an artificial system.

# Chapter 3

# Experiments

## 3.1 Description of the experiments performed

### 3.1.1 Phase 1 - Calibration, Reliability vs Speed, Subconscious Perception and Learning

The aim of the first phase of the experiments was multi-fold. The most basic objective was to confirm that the P300 signal can be reliably detected with the used experimental set-up and with the available tools. The next goal was to analyze how P300 detection reliability is affected by the rate of the image presentation (i.e. the number of images presented per second). Then it was checked if P300 activity is evoked also in situations when the image presentation rate rules out conscious perception of visual stimuli. Finally, the impact of the learning effect on the detection of P300 was assessed.

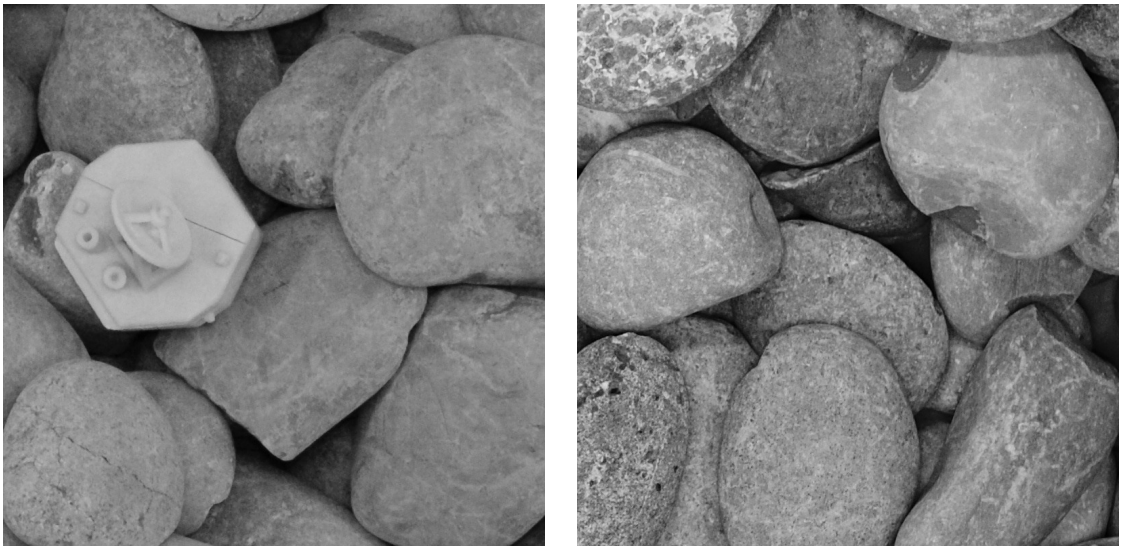
In order to fulfill these objectives, the classical oddball paradigm has been used throughout the first phase of experiments. Visual stimuli consisted of a subset of 3204 images of grey stones luminated with a uniform ambient light. 25 of those images contained in addition to the stones, a sand model of a spacecraft, thus constituting oddball images. The spacecraft position was different in each of these images but the object itself was clearly visible in all cases. Examples of background and oddball images for this first phase experiments are given in figure 3.1.

The first phase was divided into 4 experiments each related to one of the aforementioned scientific goals. Every experiment involved the presentation of one or more image sequences to experiment subjects. The subjects were instructed to count the images containing the spacecraft model and were made familiar with examples of an oddball and non-oddball image. After that, the actual sequence of the images was presented with the EEG signals being recorded, always preceded by a countdown screen of duration 5 seconds that allowed the subjects to prepare for the experiment, reducing the surprise effect of the sequence start.

The parameters of the first experiment, further referred to as the *Calibration* experiment, are summarized in Table 3.1. The goal here was to verify that the experimental setup allows for a reliable P300 detection. The experiment involved 4 subjects, and 5 different sequences

| *No. of subjects* | *No. of sequences* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|
| 4 | 5 | 40 | 4 | 2 | 500/500 | 40 |

Table 3.1: Parameters of the Calibration experiment



(a) (b)

Figure 3.1: Examples of an oddball (a) and non-oddball (b) images used for Phase 1.

| *No. of subjects* | *No. of sequences* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|
| 4 | 5 | 40 | 4 | 2 | 500/500 | 40 |
| 4 | 5 | 67 | 7 | 2 | 300/300 | 40 |
| 4 | 5 | 133 | 13 | 2 | 150/150 | 40 |
| 4 | 5 | 200 | 20 | 2 | 100/100 | 40 |
| 4 | 5 | 400 | 40 | 2 | 50/50 | 40 |

Table 3.2: Parameters of the Reliability vs Speed experiment

of images. Each of these sequences consisted of 40 images, 4 of which were oddball images. Oddballs were placed randomly in the image sequence. The experiment was repeated twice (and with the same 5 sequences) for each subject after an arbitrary rest period. Every image was presented to the subject for 500 milliseconds (Image Display Period, IDP), after which a neutral background appeared for another 500 milliseconds (Inter Image Period, IIP), resulting in a one image per second presentation rate. Thus, the presentation of one complete image sequence in this experiment took 40 seconds. The relatively low image presentation rate in this experiment should allow a very reliable detection of the P300 signal.

The second experiment was aimed at understanding how fast the images can be presented to the subjects while still registering a P300 response. The parameters of this experiment, further referred to as *Reliability vs Speed* are presented in Table 3.2. Image sequences of different lengths were presented to the subjects with increasing image presentation rate. The number of images was adjusted to the change in presentation rate, so that the total duration of one sequence stayed equal to 40 seconds. The number of oddball images present in the sequence was adjusted accordingly, so that the ratio of the number of oddball images to the number of

| *No. of subjects* | *No. of sequences* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|
| 4 | 10 | 300 | 1 | 2 | 33.3/0 | 10 |
| 4 | 10 | 600 | 1 | 2 | 16.7/0 | 10 |

Table 3.3: Parameters of the Subconscious Perception experiment

| *No. of subjects* | *No. of sequences* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|
| 4 | 5 | 100 | 10 | 5 | 100/100 | 20 |

Table 3.4: Parameters of the Learning experiment

non-oddball images was kept on the same level (10%). The oddballs were placed randomly in the sequences. As for the first part of the experiment all parameters are identical to the ones used in the Calibration experiment and the results of the latter were re-used.

The third issue addressed in this phase of experiments was to check that brain activity can be detected and related to oddballs even when the image presentation rate is too high to allow conscious perception. Thus, a much higher image presentation rate than in the first two experiments has been used, and no inter-image blank was used (IIP=0). Two timing options have been used, resulting in displaying 30 and 60 images per second respectively, which is higher than the commonly agreed threshold of conscious perception, being 20 images per second [9]. For these two options, 10 different image sequences have been used, each of them containing exactly one oddball image (this fact however was not known to the subject). The oddball image placement was random, however it was enforced that it is placed within the first third of the sequence for 3 out of 10 sequences, within the middle third for 4 out of 10 sequences and within the last third for remaining 3 sequences. All parameters of this experiment further referred to as *Subconscious Perception* are summarized in table 3.3.

Finally, the issue of learning the image sequence by the subject in the case of a subsequent presentation of the same image sequence, and its impact on ERP detection was addressed. In this experiment, further referred to as *Learning*, a slightly different protocol than in previous ones was used. Each of the subjects was shown 5 different image sequences, but each one of them was repeated 5 times one time after another. Moreover the subject was made aware of this fact in advance, being also instructed that "the same image sequence is going to be repeated 5 times". Relatively high image presentation rates have been used in order to allow the subjects to make mistakes and thus observe the learning effect, if present. All parameters of this experiment are given in Table 3.4.

### 3.1.2 Phase 2 - Scientific Expertise

The second phase of the experiments aimed to answer questions concerning the relation between ERPs and expert knowledge or scientific curiosity. In order to meet these objectives, a special set of visual stimuli has been used, as well as two types of experimental subjects – a person who has profound scientific knowledge about the stimuli and non-experts.

The visual stimuli used in the second phase of the experiments were taken from the Eu-

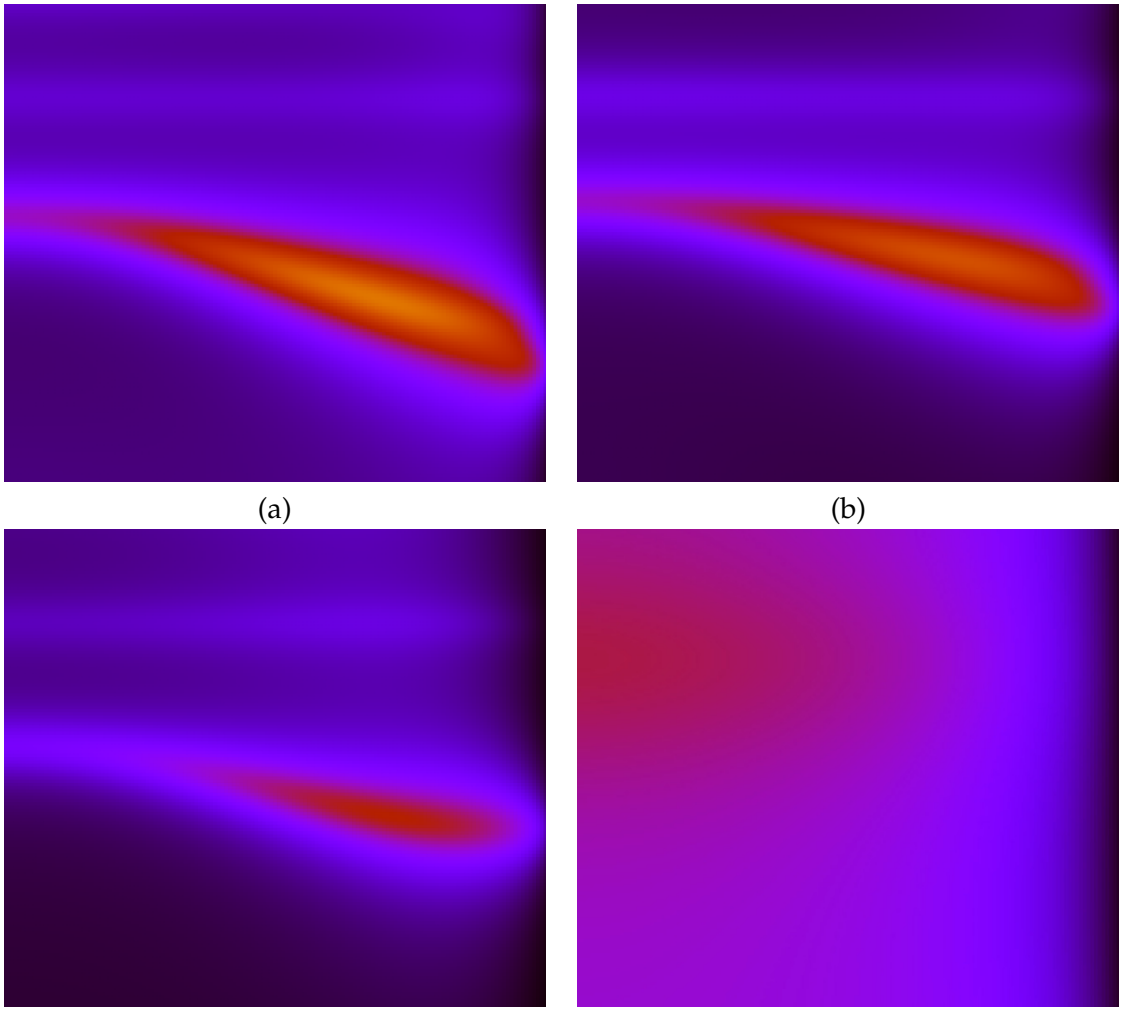



Figure 3.2: Examples of a target (a), obvious oddball (b), non-obvious oddball (c) and background (d) images used in the first experiment of the second phase.

ropean Space Agency's database of "multilayer coatings for thermal applications" [1]. The database contains images obtained during the process of designing a multilayered material exhibiting predefined thermal emissivity profiles (which are called *targets*). Spectral directional properties of a material can be presented as 2-dimensional contour plots with axes representing angle and wavelength parameters and with the colour of the point representing the magnitude of the target parameter (for example emittance). Different materials, including the ideal target solution, correspond to different plots which appear as different 2-dimensional contours. However, as a material matching exactly the desired properties is not obtainable, the best found solution will only be similar to a certain degree to the ideal target solution. This "degree of similarity" is related to a simple pattern matching process (e.g. the image looks similar to the target image) in non-expert subjects, and to more complex cognitive processes in the expert (e.g. consideration on the physics of the emissivity profiles, experience of what can be considered a good match for the emissivity pattern). The image sets used were taken from different optimization experiments for different desired ideal properties of the material and for

[1]The database can be visited at the link `www.esa.int/gsp/ACT/nan/op/bigrunresults.htm`

| *No. of subjects* | *No. of targets* | *No. of sequences per target* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|---|
| 4+1 | 2 | 5 | 50 | 3+3 | 2 | 500/0 | 25 |

Table 3.5: Parameters of the Expertise experiment

solutions of different quality. The contours were plotted in a normalized range of parameter values and stripped from the axes and the legend.

In this phase, two experiments were conducted. The first one, called *Expertise* was designed to find out if there is a difference in P300 responses between subjects who possess scientific knowledge about presented stimuli and non-expert subjects. The experiment used a modification of the oddball paradigm, with two types of oddballs: obvious and non-obvious. In each session, the non expert subject was presented an image corresponding to the target solution and instructed to "look for similar images". The subject was also shown an example image considered an obvious oddball in order to be informed about the amount of acceptable differences between target solution and "good" solutions. Then a sequence of images was presented, which contained plots of materials with properties different from the ideal target (background images), very similar to the target (obvious oddballs) and slightly similar to the target (non-obvious oddballs). Examples of such images are shown in Figure 3.2, whilst the parameters of the experiment are presented in Table 3.5.

This second phase of experiments aimed to assess the potentiality to go beyond the oddball paradigm, and to push further the initial study of [8] who demonstrated that two distinct classes may be extracted from the brainwaves by classifying the P300 response. We believe that it is actually possible to go one step beyond and to distinguish three distinct classes by classifying these same brainwaves.

In total 5 subjects were used, 1 expert (the European Space Agency's scientist conducting the aforementioned study on multilayered materials) and 4 non-experts. Two different target images were used, with 5 image sequences prepared for each of them. Every sequence contained 3 obvious and 3 non-obvious oddballs. As in previous experiments, every measurement was conducted twice. A moderately fast image presentation rate without the Inter-Image Period was used, which resulted in sequences of 25 seconds in length.

The second experiment of phase 2 to which we will refer to as the *Curiosity* experiment, was conducted on the expert subject only. No target image has been used. Non-interesting background images were mixed with potentially interesting oddball images selected by researchers preparing the image sequences, and which represented material properties that may evoke a subject's curiosity. The subject was instructed to "look for interesting properties in the displayed images". Parameters of the experiment are shown in Table 3.6. Differently from the Expertise experiment, the (expert) subject is no longer asked to perform pattern matching. Instead, with this experiment we wish to assess the potentiality of a subject's scientific curiosity being imprinted on his brain wave activity. Should we be able to subsequently train an artificial system that displays similar curiosity and attention properties to the ones of the scientist, that machine would be able to look for scientifically interesting features in images in the same way the scientist would. A visionary scenario could thus include a robot on Mars evaluating images by using the scientific curiosity of certain scientists back in earth which it has learned to imitate.

| *No. of subjects* | *No. of sequences* | *Images in seq.* | *Oddballs in seq.* | *Repetitions* | *IDP/IIP (ms)* | *T (s)* |
|---|---|---|---|---|---|---|
| 1 | 5 | 50 | 10 | 2 | 750/0 | 37.5 |

Table 3.6: Parameters of the Curiosity experiment

## 3.2 Experimental Setup

Each of the experiments described above was carried out at Multimedia Signal Processing Group (MMSPG) laboratory at EPFL. Care was taken to replicate the experimental environment as accurately as possible and the ITU-R BT. 500-11 recommendation [10] was used as a baseline. Users were facing a LCD monitor on which image sequences were presented. On the second monitor, the expert was monitoring the signals while looking at the subject to detect any moving , blinking artifacts. To ensure precise timings during image projection, an image visualization software named Curiosity Cloning Viewer (*CCViewer)* was developed [11] and used throughout the project.[2]

The EEG was recorded at 2048 Hz sampling rate from 32 active electrodes placed at the standard positions of the 10-20 international system. A Biosemi Active Two amplifier was used for amplification and and 24-bit analog to digital conversion of the EEG signals. In parallel, a second group operated at the Swiss Federal Institute of Technology (EPFL) in Lausanne, Switzerland and recorded with a 36 channel device. The EEG signals were acquired at 2048 Hz and 24-bit sampling rate from 32 electrodes that were placed on the scalp of the subjects according to the 10-20 international electrode positioning system. A Biosemi Active Two amplifier was used for amplification and analog to digital conversion of the recorded EEG signal.

Signal processing and machine learning algorithms were implemented with MATLAB. Furthermore, the online access to the EEG signals were implemented as dynamic link libraries (DLLs) in C. The DLLs were accessed from MATLAB via a MEX interface.

For the second experiment, an eye-tracker was used during the observation of image sequences. This eye-tracker was synchronized with EEG acquisition and provided information about region of interest inside each image.

## 3.3 Subjects

Subjects of this research were mostly PhD students of MMSPG laboratory at EPFL (except subject 5, who was the expert from ESA)(cf. Table3.7). Subjects one to four (all male, age $28 \pm 1.4$ , all right-handed) participated in the reliability vs. speed experiment. Subjects five to nine (1 female, 4 male, age $28.4 \pm 4.39$, all right-handed) participated in the expertise experiment. As the high speed image presentation can cause epileptic seizures on subjects with some known neurological deficits, we asked each subject before the experiment about their medical history and none of the subjects had known neurological deficits.

[2]The software has been released under BSD license and can be downloaded from `sourceforge.net/projects/ccviewer/`

| Subject | Gender | Age | Right/Left handed |
|---|---|---|---|
| 1 | Male | 28 | Right |
| 2 | Male | 27 | Right |
| 3 | Male | 27 | Right |
| 4 | Male | 30 | Right |
| 5 | Male | 33 | Right |
| 6 | Male | 31 | Right |
| 7 | Male | 22 | Right |
| 8 | Female | 26 | Right |
| 9 | Male | 30 | Right |

Table 3.7: Some Information about subjects of this research

# Chapter 4

# Signal processing and machine learning

## 4.1 Overview

In this section, a block diagram of the Implemented system is briefly introduced. Figure 4.1 illustrates the processing blocks used in the aforementioned experiments.

As it can be seen, the image sequences are first presented to the subjects of the system using the ccviewer software (see section 3.2). An EEG acquisition device, which is synchronized with image presentation protocol, acquires the EEG signals using an electro-cap which is placed on the scalp of the subject. After pre-amplification and Digitalization of the EEG signal a block of preprocessing filters are applied to the raw signal to remove the various artifacts from the signal. Wavelet decomposition is then used to break down the signal into different frequency components. After choosing the appropriate frequency components of the signal, it is windowed into one second long windows (single trials). In the next step, Features are

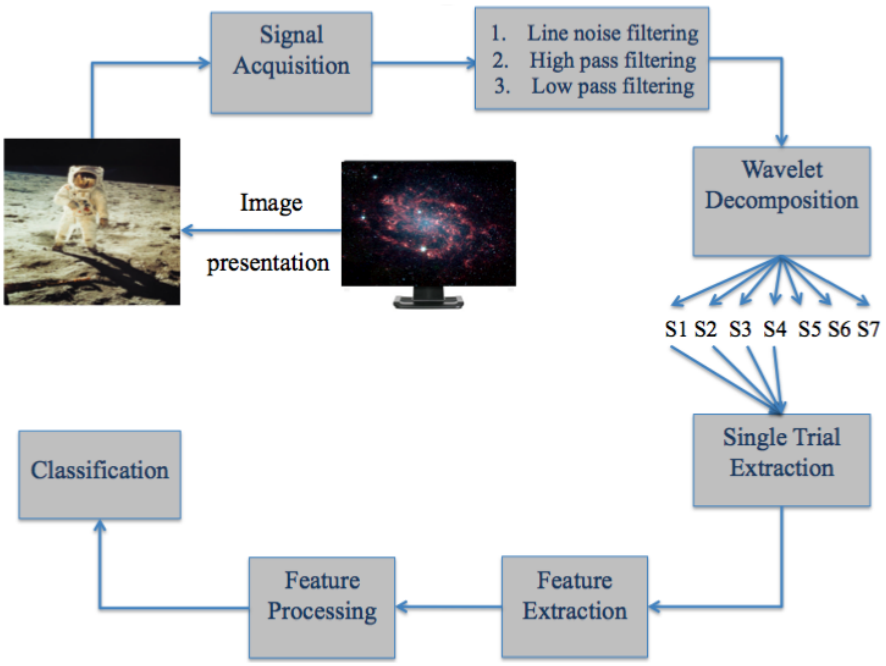



Figure 4.1: An overview of the components of implemented BCI system to recognize curiosity

extracted from the signal and finally the extracted feature vectors are classified into either target or non-target classes. In the next sections, more detailed information will be given about different blocks of the system.

## 4.2 Experimental Schedule

Each subject completed two recording sessions. The second session was performed on another day but for all subjects the time between the first and the last session was less than one weeks. The following protocol was used in each of the runs.

- Subjects were asked to perform the covert task, namely counting silently how often the target images was presented in a given sequence (For example: "Now please count how often you see a target image").
- A gray image was displayed on the screen and a countdown from five to zero was appeared to let the subjects know that the exact time that image sequence will start to be presented.
- As soon as the countdown was finished, a random sequence of images was presented and the EEG was recorded.
- After each run subjects were asked what their counting result was. This was done in order to monitor performance of the subjects.

It is worth of mention that the duration of one session including setup of electrodes and short breaks between runs was approximately 100 minutes for reliability vs. speed experiment and 30 minutes for expertise experiment.

## 4.3 Feature Extraction and Processing

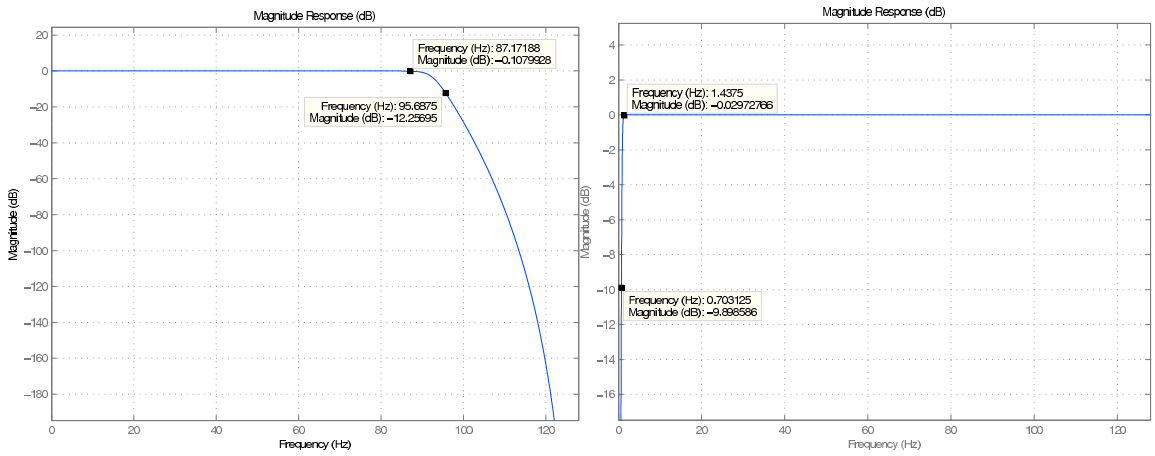
After the signal has been acquired, various offline analysis techniques are needed to preprocess the data and remove different artifacts and also to extract appropriate features from it. In this section we will describe the preprocessing and feature extraction techniques used in this study.

### 4.3.1 Referencing

The average signal from the T7 and T8 electrodes was used for referencing (see Figure 4.7. With BioSemi systems, every electrode or combination of electrodes can be the "reference". When no reference is selected, the signals are displayed with respect to the CMS (ground mastoid) electrode. This mode does not provide the full CMRR, and should only be use as a quick check of the electrodes. Only after a reference is selected, the full 80 dB CMRR is achieved.

### 4.3.2 Lowpass Filtering

A 12th order forward-backward Butterworth bandpass filter was used to filter the data. This filter had zero phase shift and it's cut-off frequency was set to 95 Hz. Figure 4.2a illustrates the magnitude response of this filter. the rational behind lowpass filtering is to remove all non-EEG artifacts from the signal as the frequency boundary of this signal is limited to 90 Hz.




(a) (b)

Figure 4.2: Magnitude response of the used filters

### 4.3.3 Downsampling

In order to reduce the number of samples, the EEG was downsampled from 2048 Hz to 256 Hz. The Decimation process employs an eight-order Lowpass Chebyshev Type I filter to filter the data and then resamples the resulting smoothed signal at the lower rate.

### 4.3.4 Highpass and Notch Filtering

In order to remove the low frequency drifts a 5th order high pass filter with the cut-off frequency of 0.5 Hz and zero phase-shift behavior was used. Figure reffig:lpfiltb illustrates the magnitude response of this filter. To remove the city line noise from the signal a 50 Hz notch filter was then applied to the signal.

### 4.3.5 Discrete Wavelet Decomposition (P300 Extraction)

Wavelet transform is particularly effective for representing various aspects of signals such as trends, discontinuities, and repeated patterns where other signal processing approaches fail or are not as effective. It is especially powerful for non-stationary signal analysis. The wavelet-based time–frequency decomposition of the signal can be used in object detection, feature extraction, and time-scale or space-scale analysis.

EEG signals contain non-stationary transient events. The traditional method used to analyze these time series signals has been Fourier transform. In Fourier transform, the signal is transformed to a complex exponential function (or a sinusoidal function) and the result is a signal in the frequency domain. The traditional signal analysis of the EEG records has been based on the Fast Fourier Transform algorithm. The infinite basis functions used in Fourier analysis are suitable for extracting frequency information from periodic, non-transient signals. Fourier transform, however, cannot capture the transient features in a given signal and the time–frequency information is not readily seen in the transformed Fourier coefficients. The frequency spectrum of a signal as a result of the Fourier transform is not localized in time. This implies that Fourier coefficients of a signal are determined by the entire signal support. Consequently, if additional data are added over time, Fourier transform coefficients will change. Any local behavior of a signal cannot be easily traced from its Fourier transformation.

As it can be seen in Figure 4.1, the preprocessed data is analyzed in time-frequency domain using Discrete Wavelet Transform (DWT). The transform of a signal is just another form of representing the signal. It does not change the information content present in the signal. The Wavelet Transform provides a time-frequency representation of the signal. It was developed to overcome the shortcoming of the Short Time Fourier Transform (STFT), which can also be used to analyze non-stationary signals. While STFT gives a constant resolution at all frequencies, the Wavelet Transform uses multi-resolution technique by which different frequencies are analyzed with different resolutions.

The wavelet analysis is done similar to the STFT analysis. The signal to be analyzed is multiplied with a wavelet function just as it is multiplied with a window function in STFT, and then the transform is computed for each segment generated. However, unlike STFT, in Wavelet Transform, the width of the wavelet function changes with each spectral component. The Wavelet Transform, at high frequencies, gives good time resolution and poor frequency resolution, while at low frequencies, the Wavelet Transform gives good frequency resolution and poor time resolution.

The Continuous Wavelet Transform (CWT) is provided by following equation , where x(t) is the signal to be analyzed. $\psi(t)$ is the mother wavelet or the basis function. All the wavelet functions used in the transformation are derived from the mother wavelet through translation (shifting) and scaling (dilation or compression).

$$X_{WT} = \frac{1}{\sqrt{\mid s \mid}} \int x(t)\psi^*(\frac{t-\tau}{s})dt$$

The mother wavelet used to generate all the basis functions is designed based on some desired characteristics associated with that function. The translation parameter $\tau$ relates to the location of the wavelet function as it is shifted through the signal. Thus, it corresponds to the time information in the Wavelet Transform. The scale parameter $s$ is defined as $\mid \frac{1}{frequency} \mid$ and corresponds to frequency information. Scaling either dilates (expands) or compresses a signal. Large scales (low frequencies) dilate the signal and provide detailed information hidden in the signal, while small scales (high frequencies) compress the signal and provide global information about the signal. Notice that the wavelet transform merely performs the convolution operation of the signal and the basis function. The above analysis becomes very useful as in most practical applications, high frequencies (low scales) do not last for a long duration, but instead, appear as short bursts, while low frequencies (high scales) usually last for entire duration of the signal.

The Wavelet Series is obtained by discretizing CWT. This aids in computation of CWT using computers and is obtained by sampling the time-scale plane. The sampling rate can be changed accordingly with scale change without violating the Nyquist criterion. Nyquist criterion states that, the minimum sampling rate that allows reconstruction of the original signal is $2\omega$ radians, where $\omega$ is the highest frequency in the signal. Therefore, as the scale goes higher (lower frequencies), the sampling rate can be decreased thus reducing the number of computations.

The Wavelet Series is just a sampled version of CWT and its computation may consume significant amount of time and resources, depending on the resolution required. The DWT, which is based on sub-band coding is found to yield a fast computation of wavelet transform. It is easy to implement and reduces the computation time and resources required. In CWT, the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales.

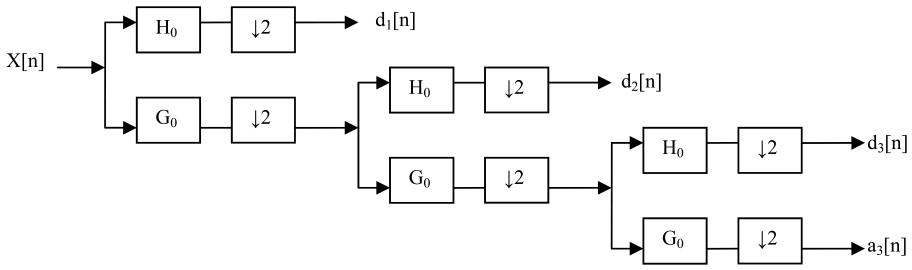



Figure 4.3: Three-level wavelet decomposition tree

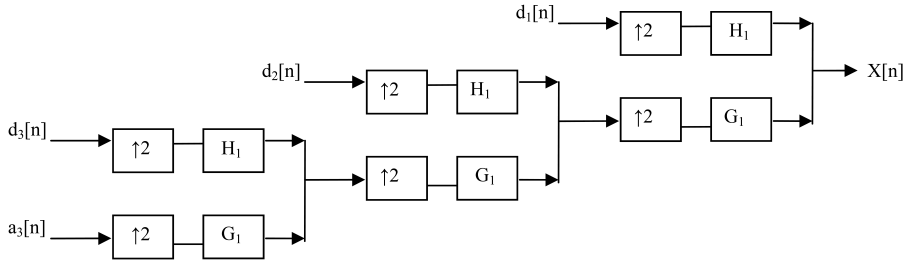



Figure 4.4: Three-level wavelet reconstruction tree

Filters are one of the most widely used signal processing functions. Wavelets can be realized by iteration of filters with rescaling. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by upsampling and downsampling (subsampling) operations [12]. The DWT is computed by successive lowpass and highpass filtering of the discrete time-domain signal as shown in figure 4.3. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time mutiresolution to discrete-time filters. In the figure, the signal is denoted by the sequence x[n], where $n$ is an integer. The low pass filter is denoted by $G_0$ while the high pass filter is denoted by $H_0$. At each level, the high pass filter produces detail information, d[n], while the low pass filter associated with scaling function produces coarse approximations, a[n].

At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. In accordance with Nyquist's rule if the original signal has a highest frequency of $\omega$, which requires a sampling frequency of $2\omega$ radians, then it now has a highest frequency of $\omega/2$ radians. It can now be sampled at a frequency of $\omega$ radians thus discarding half the samples with no loss of information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. Thus, while the half band low pass filtering removes half of the frequencies and thus halves the resolution, the decimation by 2 doubles the scale.

With this approach, the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients, a[n] and d[n], starting from the last level of decomposition.

Figure 4.4 shows the reconstruction of the original signal from the wavelet coefficients. Basically, the reconstruction is the reverse process of decomposition. The approximation and

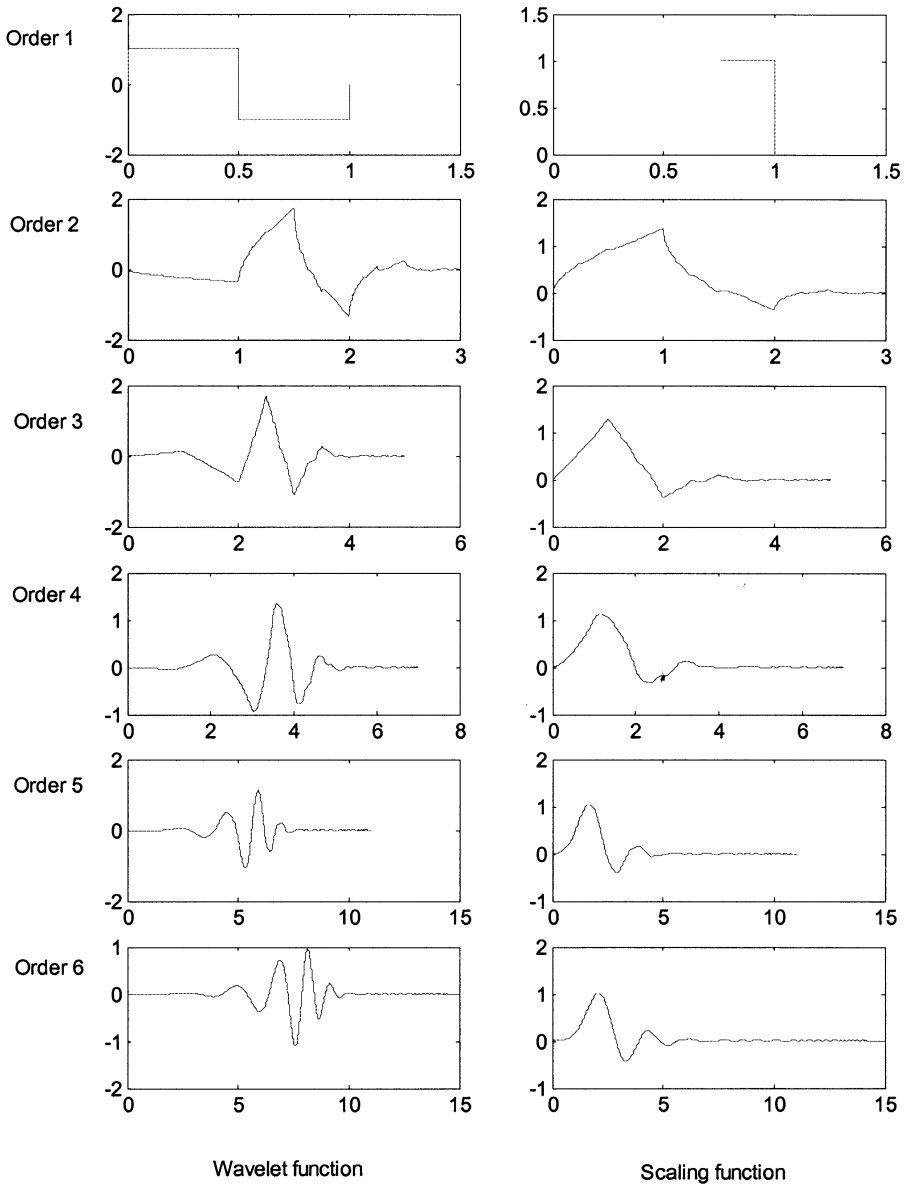



Figure 4.5: Daubechies wavelet and scaling functions of different orders [13]

detail coefficients at every level are upsampled by two, passed through the low pass and high pass synthesis filters and then added. This process is continued through the same number of levels as in the decomposition process to obtain the original signal. The Mallat algorithm works equally well if the analysis filters, $G_0$ and $H_0$, are exchanged with the synthesis filters, $G_1$ and $G_1$.

As we know, EEG contains a wide range of frequency components. However, the range of clinical and physiological interests is between 0.3 and 30 Hz. This range is classified approximately in a number of frequency bands as follows [14]:

- Delta($< 4Hz$):Delta rhythms are slow brain activities preponderant only in deep sleep stages of normal adults. Otherwise, they suggest pathologies.
- Theta(4–8 Hz): This EEG frequency band exists in normal infants and children as well as during drowsiness and sleep in adults. Only a small amount of theta rhythms appears in

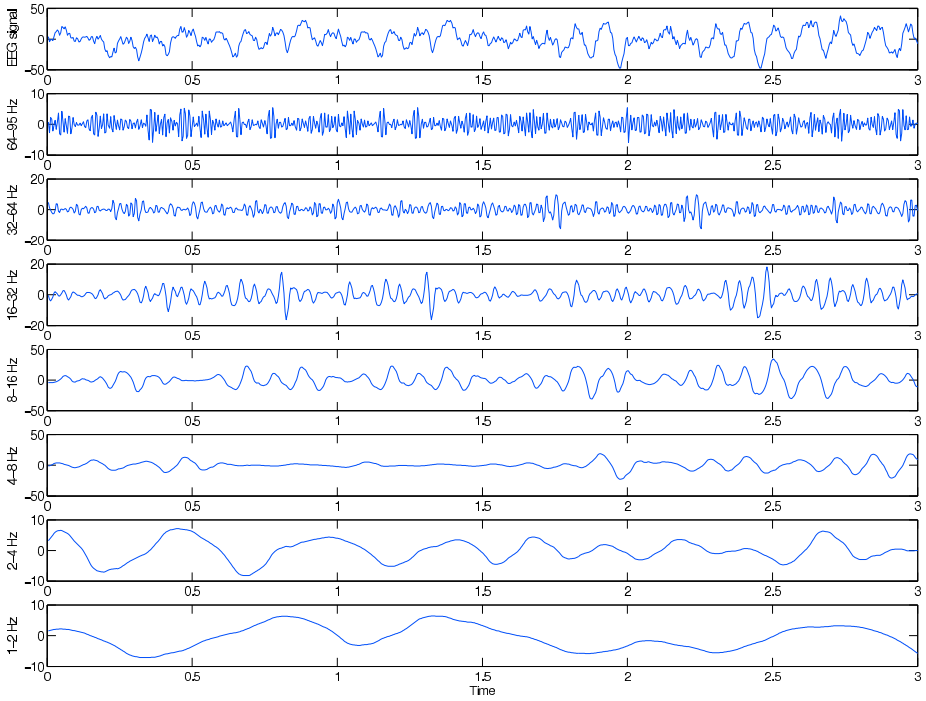



Figure 4.6: three second long EEG signal and its decomposition using db4 wavelet

the normal waking adult. Presence of high theta activity in awake adults suggests abnormal and pathological conditions.

- Alpha(8–14 Hz): Alpha rhythms exist in normal adults during relaxed and mentally inactive awakeness. The amplitude is mostly less than $50\mu V$ and appears most prominent in the occipital area. Alpha rhythms are blocked by opening the eyes (visual attention) and other mental efforts such as thinking.

- Beta(14–30 Hz): Beta activity is mostly marked in frontocentral region with less amplitude than alpha rhythms. It is enhanced by expectancy states and tension.

- Gamma($> 30Hz$): Gamma rhythms have a high frequency band and usually are not of clinical and physiological interests and therefore often filtered out in EEG recordings.

Wavelet decomposition thus can be considered as a tool to break down EEG signal into its different bands. The Daubechies' family of wavelets [15] is one of the most commonly used orthogonal wavelets satisfying the admissibility conditions, thus allowing reconstruction of the original signal from the wavelet coefficients. Examples of wavelet and scaling functions for Daubechies' family of orthogonal wavelets are shown in Figure 4.5 Daubechies' wavelet family is designed with the maximum regularity (or smoothness).

Daubechies wavelets of different orders (2, 3, 4, 5, and 6) were investigated for the analysis of epileptic EEG records. This family of wavelets is known for its orthogonality property and efficient filter implementation. Daubechies order 4 and higher wavelet was found to be the most appropriate for analysis of epileptic EEG data. The lower order wavelets of the family were found to be too coarse to represent EEG spikes properly. In this study, Daubechies order 6 was used for decomposition of EEG signals into different bands. Figure 4.3 illustrates a typical EEG signal used in this study and its decomposition using DWT.

As it can be inferred from Figure 4.6, the detail bands d1, d2, d3, d4, d5, d6 equal 64-128 Hz, 32-64 Hz, 16-32 Hz, 8-16 Hz , 4-8Hz, and 2-4 Hz, respectively and the approximation band

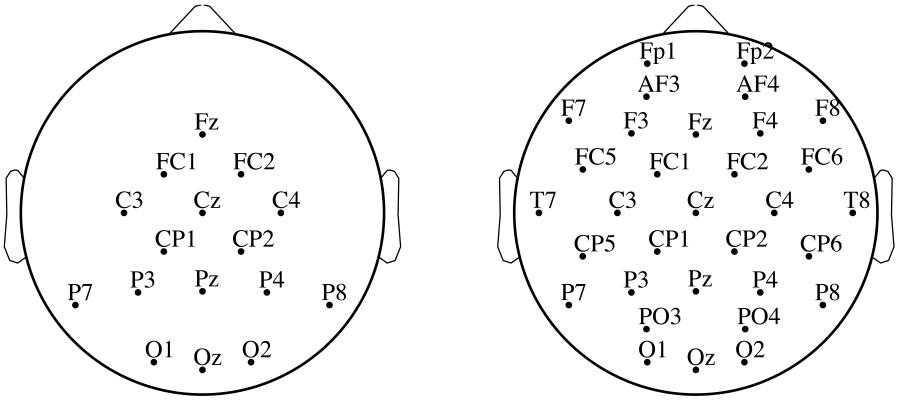



Figure 4.7: Electrode configuration used in this study (left) and 32 channel signal acquisition (right)

a6 contain 0-2Hz components. However, due to the previous lowpass and highpass filtering d1 only contains 64-95 Hz and a6 contains only 1-2 Hz frequencies. In figure 4.1 the [a6, d6, d5, d4, d3, d2, d1] signals are represented as [s1, s2, s3, s4, s5, s6, s7]. It has been shown [16] that the delta and theta frequencies that construct the major part of P300 response, however it was discussed that alpha frequencies have also minor roles in P300 construction. Therefore, we used these three bands to reconstruct the P300 signal. In other words s1, s2, s3, and s4 signals were summed together to reconstruct the P300 signal.Consequently, the reconstructed signal is comprised of 1-16 Hz frequency components.

### 4.3.6 Single trial Extraction

Single trials of duration 1000 ms were extracted from the data. Single trials started at stimulus onset, i.e. at the beginning of the intensification of first image in the image sequence, and ended 1000 ms after stimulus onset. In faster experiments thus, there is some overlap between the single trials of consequent images. For instance, for IDP=IIP=150 ms , the last 700 ms of each trial were overlapping with the first 700 ms of the following trial.

### 4.3.7 Feature Extraction

After extraction of single trials, the reconstructed signal was again downsampled from 256 Hz to 32 Hz. To this end, an eight-order Lowpass Chebyshev Type I filter with the cut-off frequency of 12.8 Hz was used to filter the data and then resamples the resulting smoothed signal at the lower rate. Therefore, the decimated signal contains frequency components of 1-12.8 Hz and each single trial contains 32 samples. 16 electrodes then were chosen and the signals of these electrodes were concatenated to form a feature vector corresponding to that single trial. figure 4.7 shows the selected electrodes and also the 32 electrode system used for signal acquisition.

The dimensionality of the feature vectors was $N_e \cdot N_s$ where $N_e$ denotes the number of electrodes and $N_s$ denotes the number of temporal samples in one trial. Due to the trial duration of 1000 ms and the downsampling to 32 Hz, $N_s$ always equaled 32. Depending on the electrode configuration, $N_e$ equal eight, sixteen, or thirty-two.

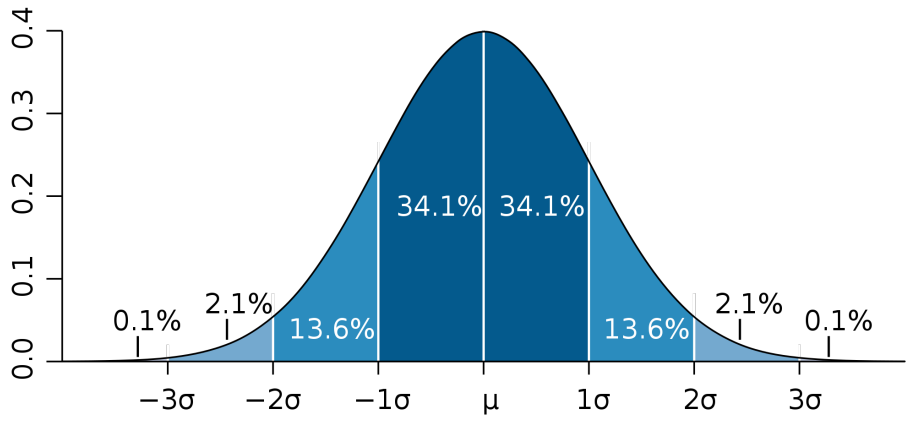



Figure 4.8: Probability distribution of normalized data

### 4.3.8 Feature Matrix Normalization

Once a feature vector corresponding to each single trial has been created, they were put together to form the feature matrix. Different rows of the feature matrix contains different samples of the single trials and thus it can be seen that the dynamic ranges of the rows in the feature matrix are not the same. In order to unify the dynamic range of the rows and give the same weight to all feature points a normalization technique was used as follows.

$$\bar{X} = \frac{X - M}{S}$$

where X denotes the feature vector, $M$ represents a vector of mean value, and $S$ denotes a vector of standard deviation values calculated over the samples. Consequently, each dimension of normalized feature matrix has the mean value of zero and standard deviation of one.

In the next step, a simple technique was used to reject the artifacts. As it can be seen in Figure4.8, Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about %68 of the set (dark blue), while two standard deviations from the mean (medium and dark blue) account for about %95, and three standard deviations (light, medium, and dark blue) account for about %99.730. For outlier removal purpose, the samples which were located outside four standard deviation from the mean, were removed from the feature matrix of each class, separately. These samples account for %99.993 of the set.

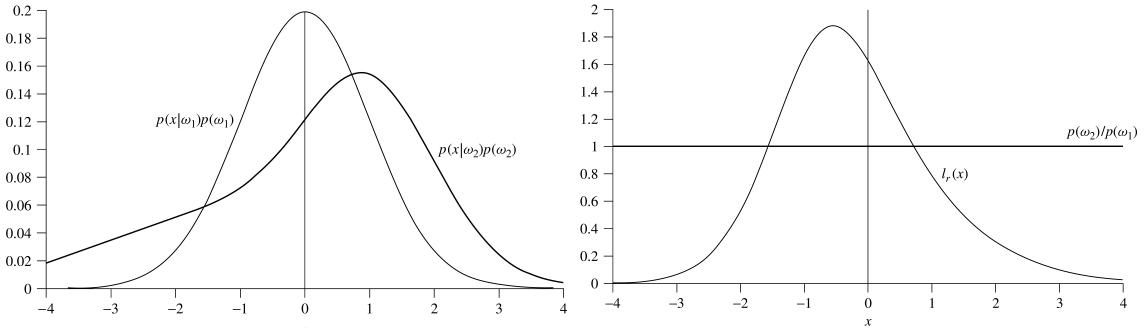
## 4.4 Classification

### 4.4.1 Bayesian inference

#### Bayes decision rule for minimum error

Consider C classes, $\omega_1, ..., \omega_C$, with a priori probabilities (the probabilities of each class occurring) $p(\omega_1), ..., p(\omega_C)$, assumed known. If we wish to minimize the probability of making an error and we have no information regarding an object other than the class probability distribution then we would assign an object to class $\omega_j$ if

$$p(\omega_j) > p(\omega_k) \quad k = 1, ..., C; \quad k \neq j$$




(a) (b)

Figure 4.9: a. $p(x \mid \omega_j)p(\omega_j)$, for classes $\omega_1$ and $\omega_2$
b. Likelihood function

This classifies all objects as belonging to one class. For classes with equal probabilities, patterns are assigned arbitrarily between those classes. However, we do have an observation vector or measurement vector $x$ and we wish to assign $x$ to one of the $C$ classes. A decision rule based on probabilities is to assign emphxto class $\omega_j$ if the probability of class $\omega_j$ given the observation $x$, $p(\omega_j \mid x)$, is greatest over all classes $\omega_1, ..., \omega_c$. That is, assign $x$ to class $\omega_j$ if

$$p(\omega_j \mid x) > p(\omega_k \mid x) \quad k = 1, ..., C; \quad k \neq j$$

This decision rule partitions the measurement space into $C$ regions $\Omega_1, ..., \Omega_C$ such that if $x \in \Omega_j$ then $x$ belongs to class $\omega_j$. The a posteriori probabilities $p(\omega_j \mid x)$ may be expressed in terms of the a priori probabilities and the class-conditional density functions $p(x \mid \omega_i)$ using Bayes' theorem as:

$$p(\omega_i \mid x) = \frac{p(x \mid \omega_i)p(\omega_i)}{p(x)}$$

and so the decision rule may be written as assign $x$ to $\omega_j$ if

$$p(x \mid \omega_j)p(\omega_j) > p(x \mid \omega_k)p(\omega_k) \quad k = 1, ..., C; \quad k \neq j$$

This is known as Bayes' rule for minimum error. For two classes, this decision rule may be written as

$$l_r(x) = \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{p(\omega_2)}{p(\omega_1)}$$

the function $l_r(x)$ is called the likelihood ratio. Figure 4.9 illustrates a two class problem, the distribution of two classes and the likelihood function. It can be shown that this decision rule corresponds with the minimization of error [17] and thus it is called Bayes decision rule for minimum error.

**Bayes decision rule for minimum risk**

In the previous section, the decision rule selected the class for which the a posteriori probability, $p(\omega_j \mid x)$, was the greatest. This minimized the probability of making an error. We now consider a somewhat different rule that minimizes an expected loss or risk. This is a very important concept since in many applications the costs associated with misclassification depend upon the true class of the pattern and the class to which it is assigned. For example, in a medical diagnosis problem in which a patient has back pain, it is far worse to classify a patient with

severe spinal abnormality as healthy (or having mild back ache) than the other way round. We make this concept more formal by introducing a loss that is a measure of the cost of making the decision that a pattern belongs to class $\omega_i$ when the true class is $\omega_j$ . We define a loss matrix $\Lambda$ with components

$$\lambda_{ji} = cost of assigning x to \omega_i when x \in \omega_j$$

In practice, it may be very difficult to assign costs. In some situations, $\lambda$ may be measured in monetary units that are quantifiable. However, in many situations, costs are a combination of several different factors measured in different units – money, time, quality of life. As a consequence, they may be the subjective opinion of an expert. The conditional risk of assigning a pattern $x$ to class $\omega_i$ is defined as

$$l^i = \sum_{j=1}^{C} \lambda_{ji} p(\omega_j \mid x)$$

The average risk over $\Omega_i$ will be

$$r^i = \int_{\Omega_i} l^i(x) p(x)\, dx = \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(\omega_j \mid x) p(x)\, dx$$

and the overall expected cost or risk is

$$r = \sum_{i=1}^{C} r^i = \sum_{i=1}^{C} \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(\omega_j \mid x) p(x)\, dx$$

The above expression for the risk will be minimized if the regions $\Omega_i$ are chosen such that if

$$\sum_{j=1}^{C} \lambda_{ji} p(\omega_j \mid x) p(x) \leq \sum_{j=1}^{C} \lambda_{jk} p(\omega_j \mid x) p(x) \quad k = 1, ..., C$$

then $x \in \Omega_i$. This is known as the Bayes decision rule for minimum risk. Intuitively, assigning risk to decision will move the decision line horizontally in figure4.9.

### 4.4.2 Support Vector Machines

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 [18]. The SVM classifier is widely used in bioinformatics (and other disciplines) due to its high accuracy, ability to deal with high-dimensional data, and flexibility in modeling diverse sources of data [19]. SVMs belong to the general category of kernel methods [20]. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classifier to data that have no obvious fixed-dimensional vector space representation. In [21] a complete explanation about SVMs and different kernels is given. Here we briefly describe how SVMs can be useful for pattern classification. For a given discriminating hyperlane we denote by $x_+$ and $x_-$ the closest point to the hyperpalne among the positive and negative examples respectively. The norm of a vector $W$ denoted by $\|W\|$ is its length, and is given by $\sqrt{W^T W}$. A unit vector $\hat{W}$ in the direction of

$W$ is given by $W/\|W\|$ and has$\|\hat{W}\|$. From simple geometric considerations the margin of a hyperplane $f$ with respect to a dataset $D$ can be seen to be:

$$m_D(f) = \frac{1}{2}\hat{W}^T(x_+ - x_-) = \frac{1}{\|W\|}$$

where $\hat{W}$ is a unit vector in the direction of $W$, and we assume that $x_+$ and $x_-$ are equidistant from the decision boundary.

Now that we have the concept of a margin we can formulate the maximum margin classifier. We will first define the hard margin SVM, applicable to a linearly separable dataset, and then modify it to handle non-separable data. The maximum margin classifier is the discriminant function that maximizes the geometric margin $1/\|W\|$ which is equivalent to minimizing $\|W\|^2$. This leads to the following constrained optimization problem:

$$\underset{W,b}{\text{minimize}}\ \frac{1}{2}\|W\|^2 \qquad s.t \qquad y_i(W^T x_i + b) \geq 1 \qquad i = 1, ..., n$$

The constraints in this formulation ensure that the maximum margin classifier, classifies each example correctly, which is possible since we assumed that the data is linearly separable. In practice, data is often not linearly separable; and even if it is, a greater margin can be achieved by allowing the classifier to misclassify some points. To allow errors we replace the inequality constraints in the previous equation with

$$y_i(W^T x_i + b) \geq 1 - \varepsilon_i \quad i = 1, ..., n$$

where $\varepsilon_i \geq 0$ are slack variables that allow an example to be in the margin $0 \leq \varepsilon_i \leq 0$ also called a margin error) or to be misclassified ($\varepsilon_i > 1$). Since an example is misclassified if the value of its slack variable is greater than 1, $\sum_i \varepsilon_i$ is a bound on the number of misclassified examples. Our objective of maximizing the margin, i.e. minimizing $\frac{1}{2}\|W\|^2$ will be augmented with a term $C\sum_i \varepsilon_i$to penalize misclassification and margin errors. The optimization problem now becomes:

$$\underset{W,b}{\text{minimize}}\ \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\varepsilon_i \qquad s.t \qquad y_i(W^T x_i + b) \geq 1 - \varepsilon_i \qquad \varepsilon_i \geq 0$$

The constant $C > 0$ sets the relative importance of maximizing the margin and minimizing the amount of slack. This formulation is called the soft-margin SVM, and was introduced by Cortes and Vapnik [22].

Many datasets encountered in bioinformatics and other areas of application are unbalanced, i.e. one class contains a lot more examples than the other. Unbalanced datasets can present a challenge when training a classifier and SVMs are no exception— see [23] for a general overview of the issue. A good strategy for producing a high-accuracy classifier on imbalanced data is to classify any example as belonging to the majority class; this is called the majority-class classifier. While highly accurate under the standard measure of accuracy such a classifier is not very useful. When presented with an unbalanced dataset that is not linearly separable, an SVM that follows the previous equation will often produce a classifier that behaves similarly to the majority-class classifier. An illustration of this phenomenon is provided in Figure 4.10.

The crux of the problem is that the standard notion of accuracy (the success rate, or fraction of correctly classified examples) is not a good way to measure the success of a classifier applied to unbalanced data (see section 4.5), as is evident by the fact that the majority-class classifier

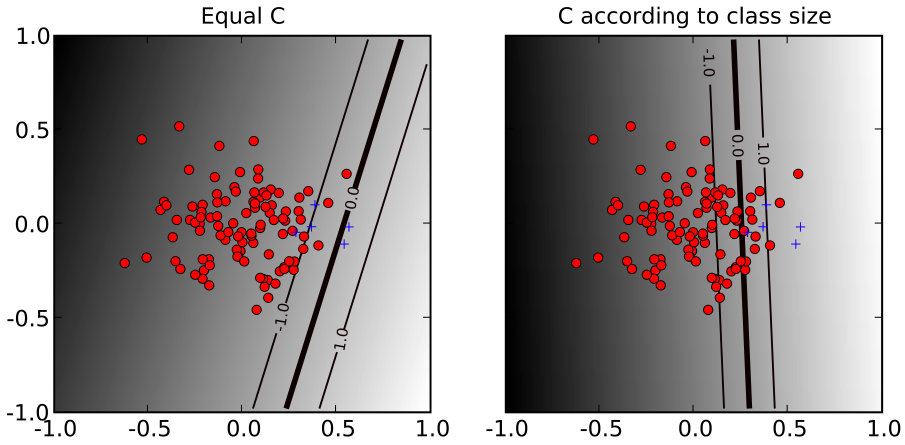



Figure 4.10: When data is unbalanced and a single soft-margin is used, the resulting classifier (left) will tend to classify any example to the majority-class. The solution (right panel) is to assign a different soft-margin constant to each class (see text for details).

performs well under it. The problem with the success rate is that it assigns equal importance to errors made on examples belonging the majority class and errors made on examples belonging to the minority class. To correct for the imbalance in the data we need to assign different costs for misclassification to each class. Before introducing the balanced success rate we note that the success rate can be expressed as:

$$P(success \mid +)P(+) + P(success \mid -)P(-)$$

where $P(success \mid +)$ and $P(success \mid -)$ are the estimate of probabilities of success in classifying positive and negative samples, respectively. $P(+)$ and $P(-)$ are the fraction of positive and negative samples. Thus, the balanced success rate modifies this expression to:

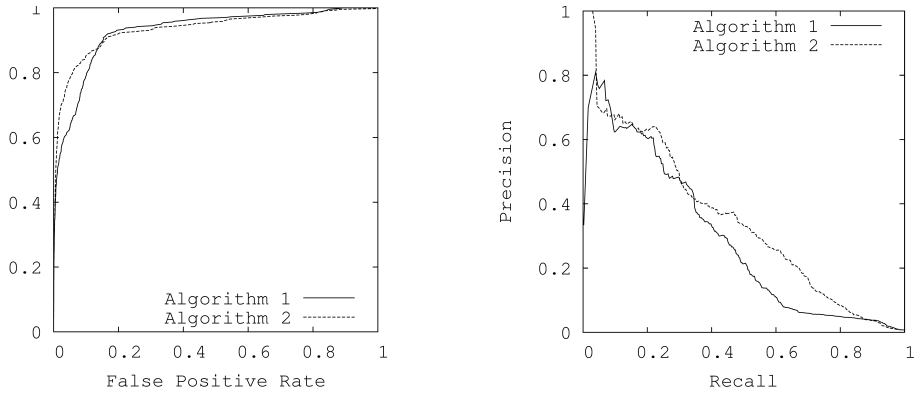
$$BSR = \frac{P(success \mid +) + P(success \mid -)}{2}$$

which averages the success rates in each class. The majority-class classifier will have a balanced successrate of 0.5. A balanced error-rate is defined as 1 ? BSR. The BSR, as opposed to the standard success rate, gives equal overall weight to each class in measuring performance. A similar effect is obtained in training SVMs by assigning different misclassification costs (SVM soft-margin constants) to each class. The total misclassification cost, $C\sum_{i=1}^{n}\varepsilon_i$ is replaced with two terms, one for each class:

$$C\sum_{i=1}^{n}\varepsilon_i \longrightarrow C_+\sum_{i\in I_+}\varepsilon_i + C_-\sum_{i\in I_-}\varepsilon_i$$

where $C_+$ and $C_-$ are the soft-margin constants for the positive and negative examples and $I_+$ and $I_-$ are the sets positive and negative samples, respectively. To give equal overall weight to each class we want the total penalty for each class to be equal. Assuming that the number of misclassified examples from each class is proportional to the number of examples in each class, we choose $C_+$ and $C_-$ such that

$$C_+n_+ = C_-n_-$$




(a) Comparison in ROC space

(b) Comparison in PR space

Figure 4.11: The difference between comparing algorithms in ROC vs PR space [31]

where $n_+$ and $n_-$ are the number of positive and negative examples, respectively. Or in other words:

$$\frac{C_+}{C_-} = \frac{n_-}{n_+}$$

This provides a method for setting the ratio between the soft-margin constants of the two classes, leaving one parameter that needs to be adjusted. This method for handling unbalanced data is implemented in several SVM software packages, e.g. LIBSVM [24].

## 4.5 Evaluation

In machine learning, current research has shifted away from simply presenting accuracy results when performing an empirical validation of new algorithms. This is especially true when evaluating algorithms that output probabilities of class values. [25] have argued that simply using accuracy results can be misleading. They recommended when evaluating binary decision problems to use Receiver Operator Characteristic (ROC) curves, which show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. However,ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution. [26] have recommended using cost curves to address this issue. Cost curves are an excellent alternative to ROC curves, but discussing them is beyond the scope of this report.

Precision-Recall (PR) curves, often used in Information Retrieval [27, 28] have been cited as an alternative to ROC curves for tasks with a large skew in the class distribution [29, 30]. An important difference between ROC space and PR space is the visual representation of the curves. Looking at PR curves can expose differences between algorithms that are not apparent in ROC space. Sample ROC curves and PR curves are shown in 4.11(a) and 4.11(b) respectively. These curves, taken from the same learned models on a highly-skewed cancer detection dataset, highlight the visual difference between these spaces [31]. The goal in ROC space is to be in the upper-left-hand corner, and when one looks at the ROC curves in 4.11(a) they appear to be fairly close to optimal. In PR space the goal is to be in the upper-right-hand corner, and the PR curves in 4.11(b) show that there is still vast room for improvement.

The performances of the algorithms appear to be comparable in ROC space, however, in PR space we can see that Algorithm 2 has a clear advantage over Algorithm 1. This difference exists because in this domain the number of negative examples greatly exceeds the number of

positives examples. Consequently, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis. Precision, on the other hand, by comparing false positives to true positives rather than true negatives, captures the effect of the large number of negative examples on the algorithm's performance.

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table. The confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives. False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative. Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative. A confusion matrix is shown in 4.13(a). The confusion matrix can be used to construct a point in either ROC space or PR space. Given the confusion matrix, we are able to define the metrics used in each space as in 4.13(b). In ROC space, one plots the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labeled. In PR space, one plots Recall on the x-axis and Precision on the y-axis. Recall is the same as TPR, whereas Precision measures that fraction of examples classified as positive that are truly positive. 4.13(b) gives the definitions for each metric. We will treat the metrics as functions that act on the underlying confusion matrix which defines a point in either ROC space or PR space.

In an information retrieval scenario, Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search, and Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (which should have been retrieved). In a statistical classification task, the Precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been). A Precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labeled correctly) whereas a Recall of 1.0 means that every item from class C was labeled as belonging to class C (but says nothing about how many other items were incorrectly also labeled as belonging to class C).

Often, there is an inverse relationship between Precision and Recall, where it is possible to increase one at the cost of reducing the other. For example, an information retrieval system (such as a search engine) can often increase its Recall by retrieving more documents, at the cost of increasing number of irrelevant documents retrieved (decreasing Precision). Similarly, a classification system for deciding whether or not, say, a fruit is an orange, can achieve high Precision by only classifying fruits with the exact right shape and color as oranges, but at the cost of low Recall due to the number of false negatives from oranges that did not quite match the specification.

Usually, Precision and Recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single measure, such as the F-measure, which is the weighted harmonic mean of precision and recall. A measure that combines Precision and Recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-

score:

$$F = \frac{precision \cdot recall}{precision + recall}$$

This is also known as the $F_1$ measure, because recall and precision are evenly weighted. It is a special case of the general $F_\beta$ measure (for non-negative real values of $\beta$):

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

In this research, we considered the curiosity recognition problem as a typical detection problem, since the end goal is to detect only images which excited the curiosity in subjects and to discard other images in the dataset. Therefore a high recall is needed to select enough target images from the dataset. However, if the high recall rate results in low precision, it wont be very beneficial to the final goal either, since there will be many non-target images among the detected images.

Moreover, the number of samples belonging to different classes are not equal in the problem that we study in this research. More precisely, the number of target images in a sequence are about %10 of the whole images in the sequence. In other words, the number of samples in non-target class is 9 times the number of samples in target class. Consequently the value of FPR is calculated over the larger samples comparing with the value of TPR. Therefore, the number of FP will be high and this is the point that is neglected with ROC-based evaluation approaches. The following example will show this problem more clearly.

As it can be seen in 4.12 The area under ROC curve equals 0.91 and the best point in the curve corresponds with the TPR value of 0.88 and FPR value of 0.16. These information look relatively good. However, analyzing this result in the precision-recall space yields the contradictory fact and that the performance in not good at all. Considering $N$, the total number of images in the sequence,16 percent of the total non-target images ($0.90 \cdot N$) which are considered by mistake as target, would be $0.16 \cdot 0.90 \cdot N$ which is equal to $0.144 \cdot N$. Using the same approach the number of TP can be written as $0.88 \cdot 0.10 \cdot N$ which is equal to $0.088 \cdot N$. Consequently the recall is equal to TPR but the precision would be:

$$precision = \frac{0.088 \cdot N}{0.144 \cdot N + 0.088 \cdot N} = 0.38$$

Consequently the f-measure for recall value of 0.88 and precision value of 0.38 equals 0.53 which is not as good as it looks in the ROC space. It is worthy of mention that the f-measure of random guess for the data under study equals 0.18.

Considering these points, In the current study, we have decided to use confusion matrix and $F_1$ measure as evaluation metrics of our classification methods.

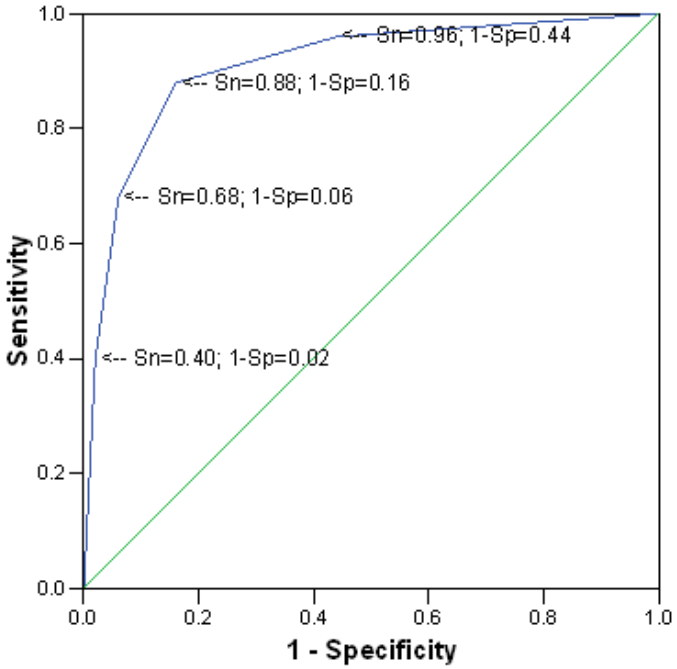



Figure 4.12: a typical roc cure for the data under study

| | actual positive | actual negative |
|---|---|---|
| predicted positive | $TP$ | $FP$ |
| predicted negative | $FN$ | $TN$ |

(a) Confusion Matrix

| | | |
|---|---|---|
| Recall | = | $\frac{TP}{TP+FN}$ |
| Precision | = | $\frac{TP}{TP+FP}$ |
| True Positive Rate | = | $\frac{TP}{TP+FN}$ |
| False Positive Rate | = | $\frac{FP}{FP+TN}$ |

(b) Definitions of metrics

Figure 4.13: Common machine learning evaluation metrics

# Chapter 5

# Results

In this chapter, the results of different experiments explained in section 3 are given and discussed. As it can be seen in Figure 4.1, several preprocessing and feature extraction techniques have been used to create a feature matrix from EEG signals. In summary, after the EEG signal had been acquired and decimated , it was filtered with so that the high frequency, very low frequency and city line noises are filtered out from it. In the next step, DWT decomposition was performed using Daubechies wavelet of order 6 and a6, d6, d5 and d4 sub-bands were summed together to create a signal which convey frequency components of 1-16 HZ. However, for the feature extraction, this signal was again decimated from 256 Hz to 32 Hz and thus the final signals is comprised of 1-12.8 Hz, due to the additional lowpass filtering during decimation.

## 5.1 Reliability vs. Speed Experiment

### 5.1.1 Study of Averaged Signals

To begin the analysis of signals relating target and non-target images, the target and non-target signals were averaged separately. Figures 5.1, 5.2, 5.3, 5.4, 5.5 illustrate the averaged signals (over the subjects), taken from six electrodes (P3, PO3, PZ, PO4, P4, CZ), for the the reliability vs. speed experiment (from IIP=IDP 500 ms to IIP=IDP 50 ms).

As it can be inferred from the figures, all the parietal, parietal-occipital and central electrodes contain the P300 benchmark. Moreover, P300 can be also observed in occipital channels. The P300 amplitude is usually larger in parietal regions compared to other parts of the scalp. Figure 5.6 illustrates the amplitude of P300 of PZ in different experiments. It can be seen that the amplitude of P300 clearly decreases as the speed of image presentation increases. In the first experiment, the largest amplitude of P300 equals 14.03$\mu$v and as the speed rate goes up, this value degrades to 9.57$\mu$v, 8.93 $\mu$v, 8.38 $\mu$v, and finally 5.66 $\mu$v in the fastest experiment. The reason for this might be twofold. First, as the speed of image presentation increases, it will become harder for subjects to detect the target images in the sequence. we observed this fact as we asked the subjects at the end of each run, how many target images they had seen in the test sequence. for the IIP and IDP values of 100 ms and lower, most of the subjects were unable to give the exact number and their answers were usually far from the real number of target images. Subject 4, could stay concentrated during the test and his answers were relatively better compare to others, however, even for him the answers were not mostly correct. Consequently, when the subjects failed to detect a target image, it is not reasonable to expect their brains to generate a P300 peak. The problem will be more noticeable, considering that the EEG segment corresponding to these missed targets are labeled as targets and will be used later for training.

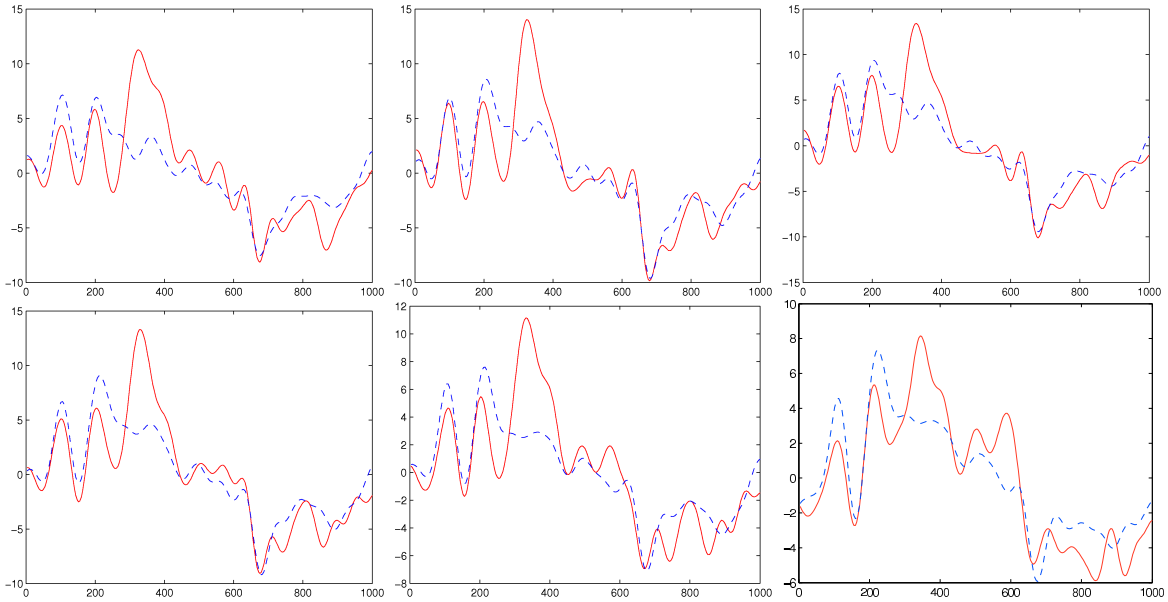


Figure 5.1: Averaged target signals (red) and non target signals (blue) for experiment one ( IIP=IDP= 500 ms ) taken from electrodes P3, PZ, PO3, PO4, P4, CZ (top left to bottom right). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

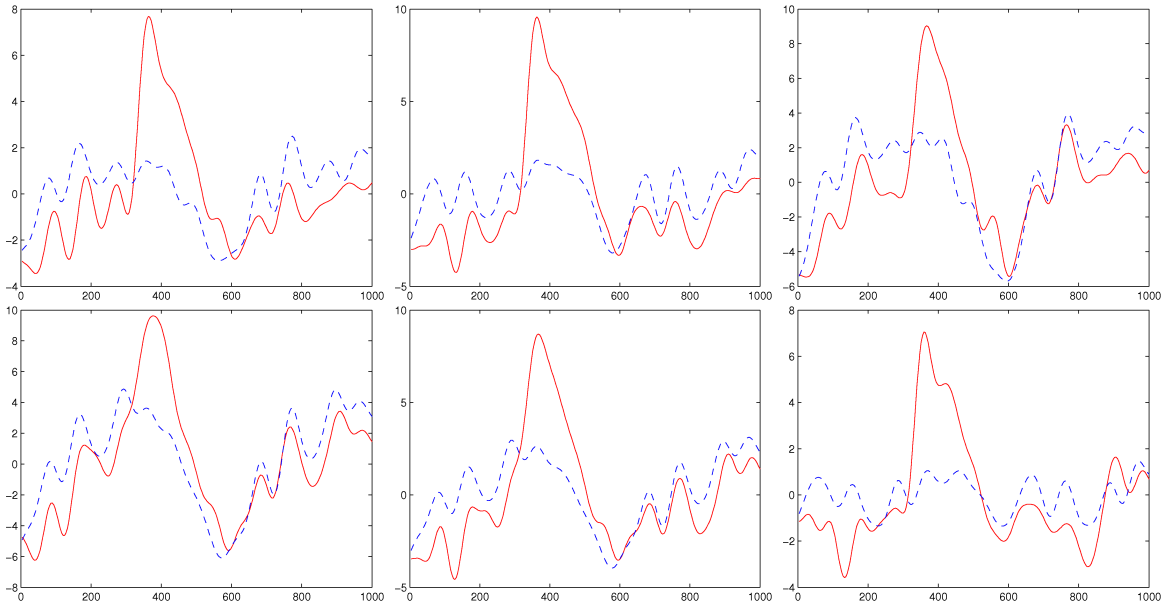


Figure 5.2: Averaged target signals (red) and non target signals (blue) for experiment two ( IIP=IDP= 300 ms ) taken from electrodes P3, PZ, PO3, PO4, P4, CZ (top left to bottom right). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

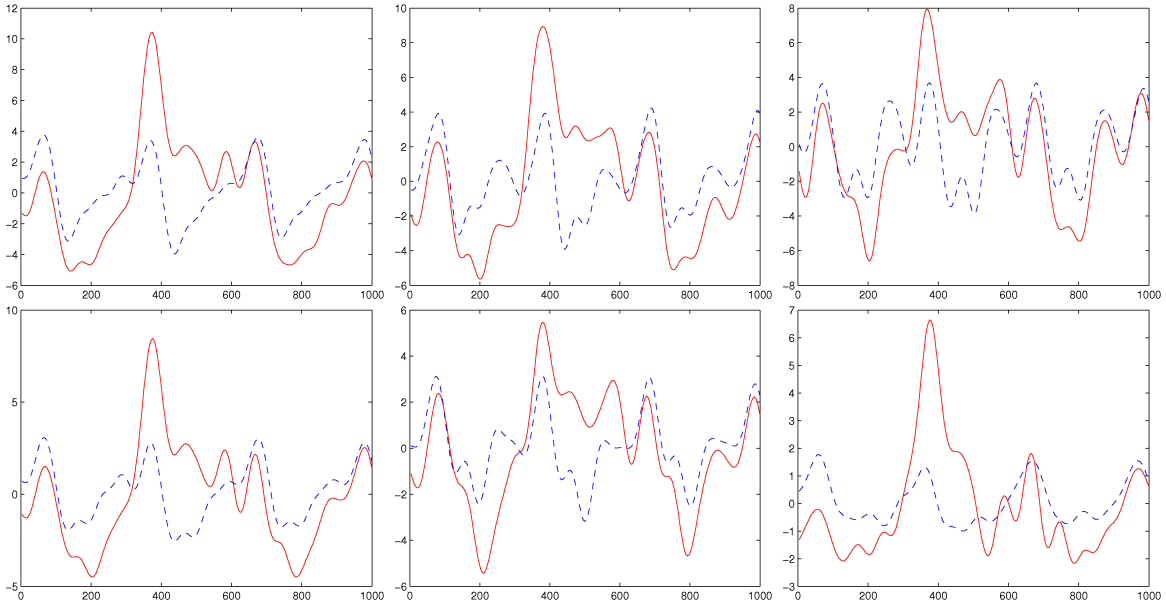


Figure 5.3: Averaged target signals (red) and non target signals (blue) for experiment three ( IIP=IDP= 150 ms ) taken from electrodes P3, PZ, PO3, PO4, P4, CZ (top left to bottom right). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

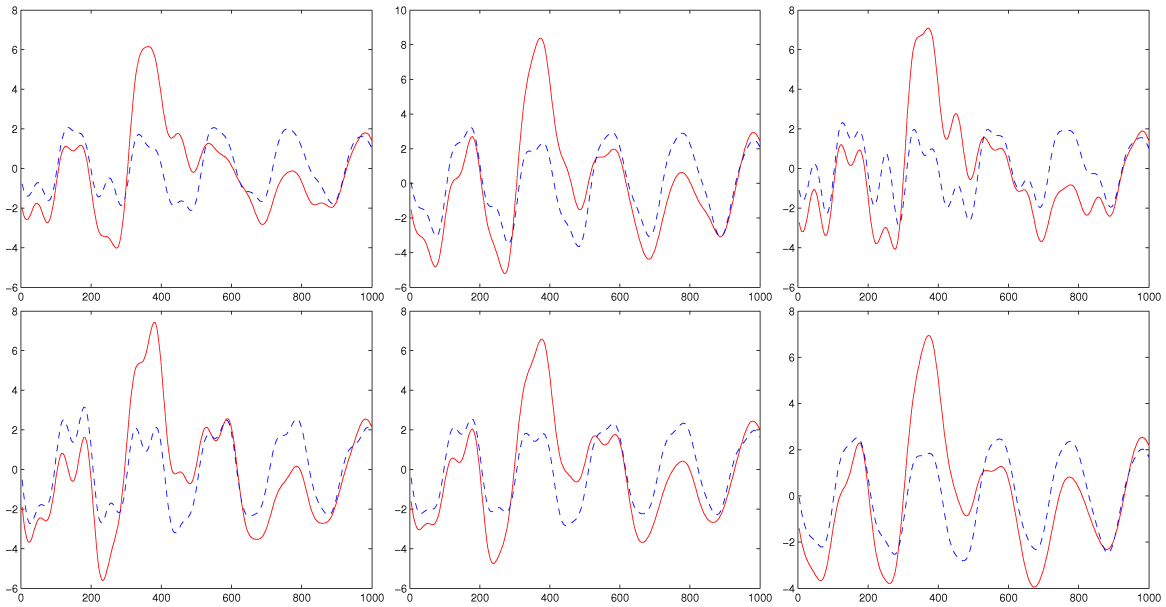


Figure 5.4: Averaged target signals (red) and non target signals (blue) for experiment four ( IIP=IDP= 100 ms ) taken from electrodes P3, PZ, PO3, PO4, P4, CZ (top left to bottom right). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

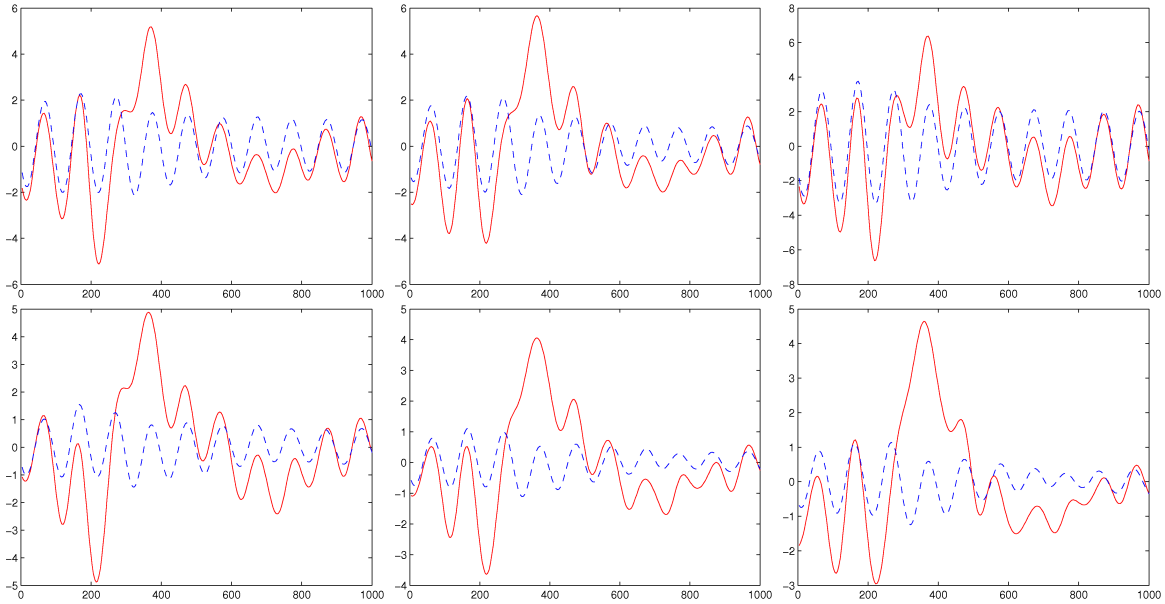


Figure 5.5: Averaged target signals (red) and non target signals (blue) for experiment five ( IIP=IDP= 50 ms ) taken from electrodes P3, PZ, PO3, PO4, P4, CZ (top left to bottom right). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

Since subjects were performing the covert task during the experiment (silently counting), there is no way to recognize which targets were detected by the subjects to only select them for the training phase.

Second, as the image presentation speed increases, the evoked potentials interfere with each other and the measured EEG signal would be superposition of different evoked potentials. For example, for the experiment where IIP and IDP values are set to 100 ms, there would be a stimuli onset every 200 ms and therefore the P300 amplitude of the target stimuli can be reduced when superimposed with N100[1] potential of the next stimuli and consequently it wont be detectable.

As it can be seen in Figures 5.3, 5.4, and 5.5, it seems that the EEG signal has some other pattern-like components apart from P300. These oscillatory patterns exist in most of the electrodes. Further studying the EEG signals revealed the fact that these oscillatory responses have always the same oscillatory frequency of the frequency of the visual stimuli (image presentation rate). Looking at the EEG signal processing literature, we noticed that these signals are another types of visual evoked potential.

Steady State Visually Evoked Potentials (SSVEP) are signals that are natural responses to visual stimulation at specific frequencies. When the retina is excited by a visual stimulus ranging from 2.5 Hz to 75 Hz, the brain generates electrical activity at the same (or multiples of) frequency of the visual stimulus. This technique is used widely with electroencephalographic research regarding vision. SSVEP's are useful in research because of the excellent signal-to-noise ratio and relative immunity to artifacts. SSVEP's also provide a means to characterize preferred frequencies of neocortical dynamic processes. SSVEP is generated by stationary lo-

[1] In neuroscience, the N100 or N1 is a large, negative-going evoked potential in EEG signal and it peaks in adults between 80 and 120 milliseconds after the onset of a stimulus, and distributed mostly over the fronto-central region of the scalp. It is elicited by any unpredictable stimulus in the absence of task demands. It is often referred to with the following P200 evoked potential as the "N100-P200" or "N1-P2" complex.

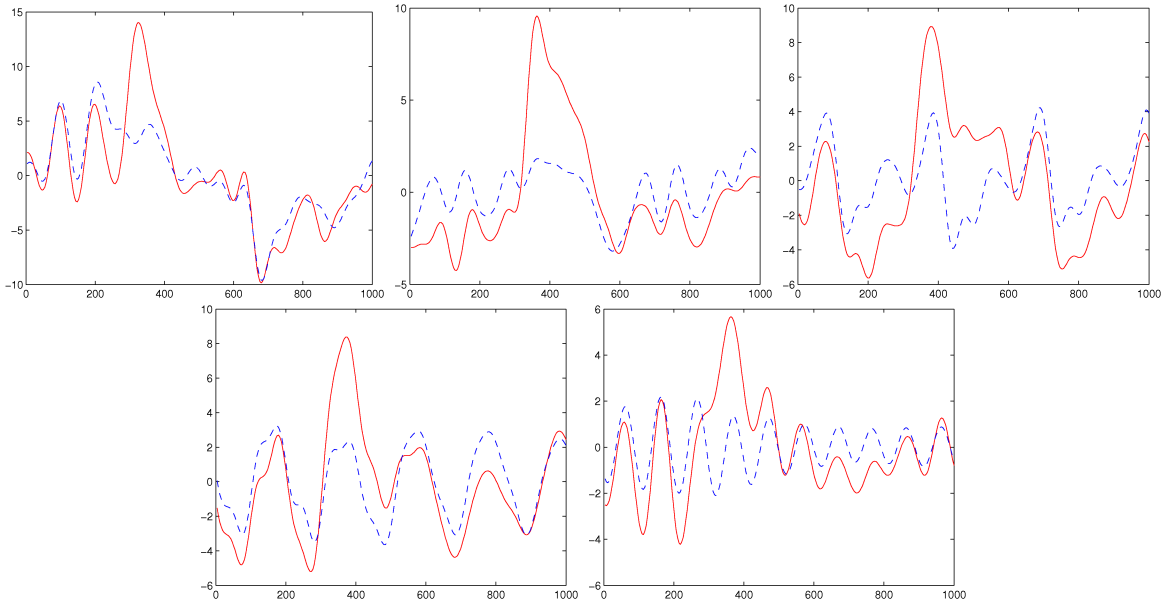


Figure 5.6: Averaged signals of PZ electrodes for different image processing rates (top left: IIP=IDP= 500 ms to bottom right: IIP=IDP= 50 ms). X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the P300 signal in $\mu$v.

calized sources and distributed sources that exhibit characteristics of wave phenomena. For detailed analysis of the frequencies of the responses the following technique was employed.

Considering that EEG signal is a non-stationary signal, using FFT is not an appropriate tool for analysis of the frequency components of this signal. However, short windows of EEG signal (less than 2 seconds) can be considered as wide-sense stationary random processes. therefore, an appropriate technique for analyzing the frequency components of the short EEG segments will be Wiener-Khinchin theorem.

The Wiener–Khinchin theorem (also known as the Wiener–Khintchine theorem and sometimes as the Wiener–Khinchin–Einstein theorem or the Khinchin–Kolmogorov theorem) states that the power spectral density of a wide-sense-stationary random process is the Fourier transform of the corresponding autocorrelation function [32].

In the continuous case:

$$S_{xx}(f) = \int_{-\infty}^{\infty} r_{xx}(\tau) e^{-j2\pi f\tau}\, d\tau$$

where

$$r_{xx}(\tau) = \mathrm{E}\left[\, x(t)x^*(t-\tau)\,\right]$$

is the autocorrelation function defined in terms of statistical expectation, and where $S_{xx}(f)$ is the power spectral density of the function x(t). Note that the autocorrelation function is defined in terms of the expected value of a product, and that the Fourier transform of x(t), does not exist in general, because stationary random functions are not square integrable.

The asterisk denotes complex conjugate, and can be omitted if the random process is real-valued.

In the discrete case the above-mentioned equation becomes:

$$S_{xx}(f) = \sum_{k=-\infty}^{\infty} r_{xx}[k] e^{-j2\pi kf}$$

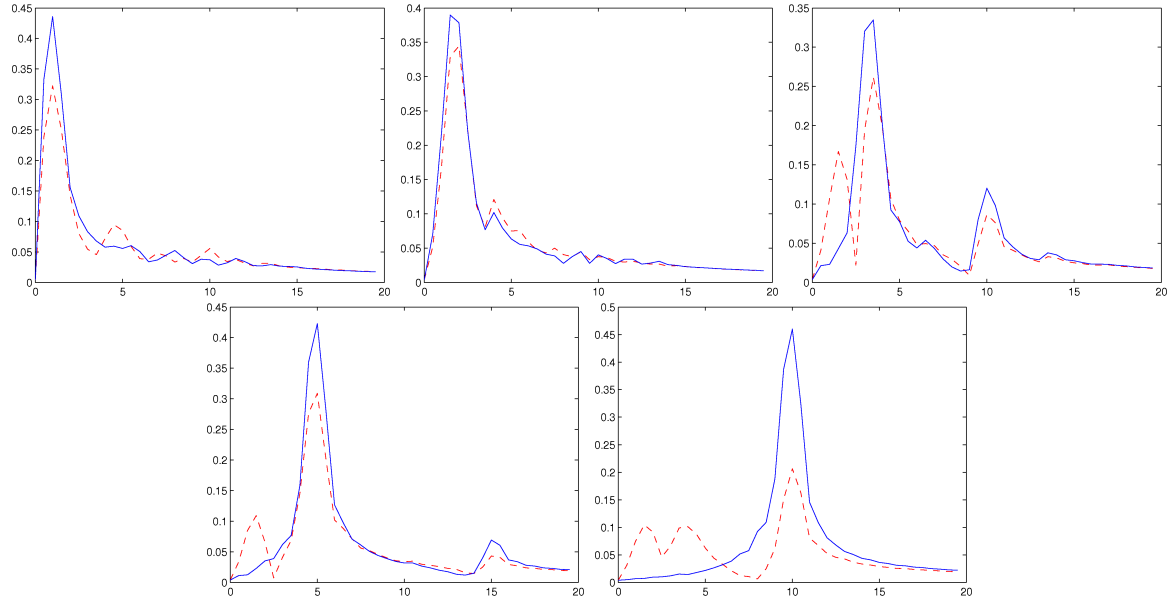


Figure 5.7: Power spectrum density of the averaged target (red) and non-target (blue) EEG signals taken from OZ electrode (top left: IIP=IDP= 500 ms, top middle: IIP=IDP= 300 ms, top right: IIP=IDP= 150 ms, bottom left: IIP=IDP= 100 ms, bottom right: IIP=IDP= 50 ms ). X axis corresponds to the frequency in Hz and Y axis displays the amplitude of the corresponding frequency component

where

$$r_{xx}[k] = \mathrm{E}\left[\, x[n]x^*[n-k] \,\right]$$

and where $S_{xx}(f)$ is the power spectral density of the function with discrete values x[n]. Being a sampled and discrete-time sequence, the spectral density is periodic in the frequency domain.

Having this technique applied to the EEG signals, the power spectrum densities of the EEG single trial segments were computed. Figure 5.7 illustrates these power spectrum densities during all experiments.

As it can be seen in this figure, after the stimulus frequency or image presentation rate exceeds 3 Hz, the SSVEP appears in the EEG signal and the same frequency will have the maximum power in power spectrum density. In this figure, it can be also seen that the target and non-target power spectrum densities differ, in that for the target EEG, the power of the oscillatory component in the stimulation frequency is lower compared to that of non-target EEG. An argument for this could be that the P300 signal and SSVEP signal superimpose each other and P300 will disturb the order in the oscillating pattern of SSVEP. Looking closer at the power spectrum densities in Figure 5.7, one can observe that the target signals still contain some frequency components that are not sub-harmonic of the oscillating frequency. These peaks are much alike the peaks that are shown for the two slow experiments and they could be considered as frequency components of P300.

One other component of the EEG signal, which could be of interest is the gamma band activity. A gamma wave is a pattern of brain waves, with a frequency between 30 to 100 Hz [33], though 40 Hz is prototypical [34]. According to a popular theory, gamma waves may be implicated in creating the unity of conscious perception (the binding problem) [35]. However, there is no agreement on the theory and some researchers suggest that whether or not gamma wave activity is related to subjective awareness is a very difficult question which cannot be answered with certainty at the present time. Figure 5.8 illustrates the averaged gamma band

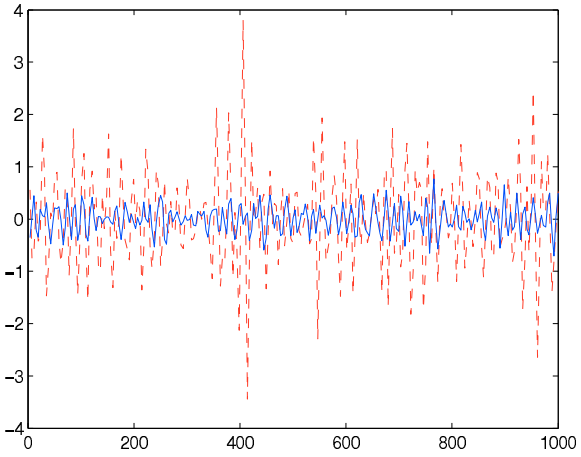



Figure 5.8: Averaged target (red) and non-target (blue) gamma band activities for the IIP=IDP=300 ms. X axis corresponds to the time after the stimulus onset in ms and Y axis displays the amplitude of the gamma band signal in $\mu$v.

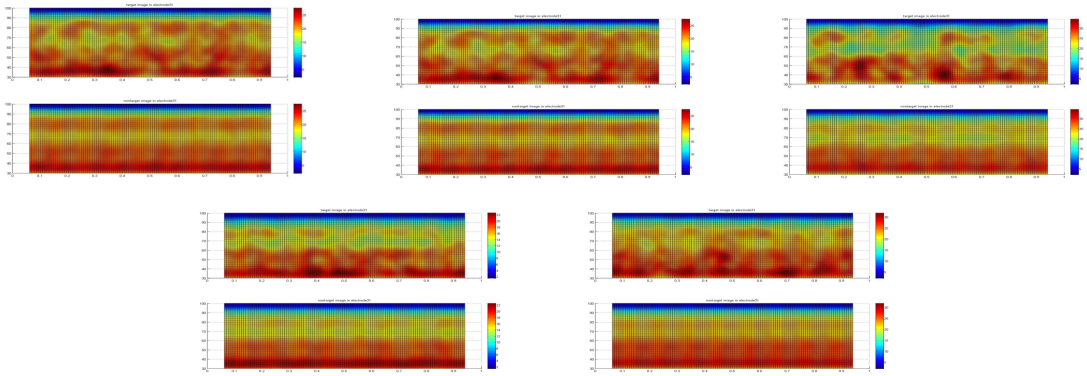


Figure 5.9: Gamma band activity of the averaged EEG signals taken from FZ electrode (top left: IIP=IDP= 500 ms, top middle: IIP=IDP= 300 ms, top right: IIP=IDP= 150 ms, bottom left: IIP=IDP= 100 ms, bottom right: IIP=IDP= 50 ms ). X axis is the time after the stimulus onset in ms and Y axis displays the amplitude of the different frequency component

signal for the second experiment:

Although, due to the chaotic and random nature of this signal the averaged signals are not directly comparable and the power of gamma band activity of target signals seems to be considerably greater that that of non-target signals, but this is mainly because the number of random activities of non-target signals were much more compared to target signals and while averaging they superimposed each other and have consequently lower power. However, it can still be seen that there is high power activity around 100 ms after the stimulus onset and also another high power activity after 300 ms after the stimulus onset. These activities are known as gamma band responses and are known to appear only if the binding mental task is performed by the subject. These activities can be witnessed in Figure 5.9.

As it can be seen in Figure 5.9, the gamma band response (40Hz) can be seen in the first 2 experiment, but for the third, fourth and fifth experiments the patterns can not be justified as and induced gamma band activity. It is worth of mention that in order to extract gamma band activities from the EEG signals, d2 and d1 frequency bands (see figure 4.1) were summed together.

### 5.1.2 Single Trial Analysis

As explained in section 4.3, 32 samples of the constructed signal which are corresponding to one-second long single trials were used as features. Therefore, the dimension of extracted feature vectors equals 512 ( $16 electrodes \times 32 samples$). Once the feature vectors have been constructed from the single trials, five-fold cross validation was performed to create different train and test sets from the dataset. This was repeated for 15 times so that different training and test sets are constructed. The following two classification approaches were then used to classify the test samples into target and non-target classes:

- Fishel Linear Discriminat Analysis (FLDA) was performed to find the optimal linear combination of the feature space. The aim of this stage is to reduce the dimension of the feature vectors and improve the accuracy of the classification at the same time by using Fisher's criterion defined as:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

  where $\mu_i$ and $\sigma_i$ represent the mean and standard deviation of each class, respectively. Let us assume that there are 2 different classes of C1, C2 for which the number of data points are N1, N2, respectively. In order to use Fisher's criterion, the so-called "within-class scatter matrix" (WCS) and "between-class scatter matrix" (BCS) are defined as:

$$S_k = \sum_{x \in C_k} (x - \mu_k)(x - \mu_k)^T$$

$$WCS = S_w = \sum_{k=1}^{2} S_k$$

$$BCS = S_b = \sum_{k=1}^{2} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

  where $x$ is the feature vector, and $\mu$ is the mean of all datapoints.

  The Fisher's criterion maximizes the trace of $\frac{S_b}{S_w}$ which is called the separability matrix. By applying the singular value decomposition to the separability matrix and discarding the insignificant eigenvalues and their corresponding columns in the eigenvector matrix, the linear transform matrix and the new dimension-reduced feature vector are obtained. It can be readily shown that the discarded eigenvalues do not have a significant role in the trace of the separability matrix. It can also be mathematically shown that the rank of the matrix $S_b$ is (L-1) and subsequently, the rank of the matrix $\frac{S_b}{S_w}$ is (L-1). In a simpler expression, they only have L-1 non-zero eigenvalues. Therefore, the final dimension of the feature vectors after this stage would be L-1 ([17]).

  Since the data used in the current study were from two different classes, hence using the aforementioned method, the dimensions of the feature vectors were reduced to one. After applying FLDA to training data, the computed mixing matrix $A$ ( this matrix was computed only using training samples) was applied to the test data to map the test feature vectors to the exact same space as the training set. Figure 5.10 illustrates the mapped feature points during 1 run of five fold cross validation for the data of subject 4 during experiment one.

  For classification, a bayesian classifier based on the training data (values along y=0 in Figure 5.10, was used. To generate a ROC curve, the minimum and maximum values of x axis over all folds was computed and then a decision line was shifted gradually from the left-most

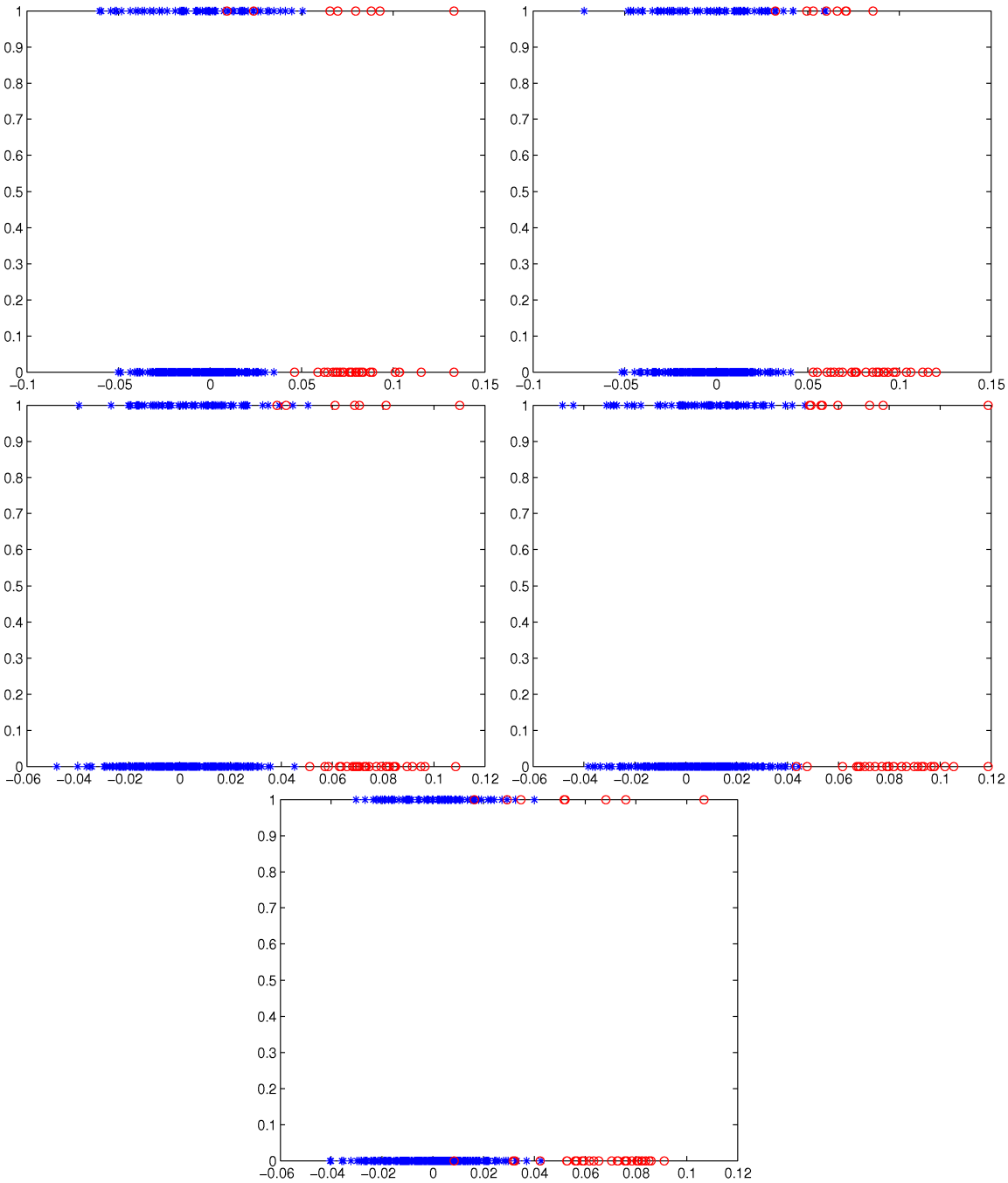



Figure 5.10: The linear combination of the feature space and maximized separability for different folds of cross validation in experiment one. X axis shows the value of the mapped combined features of target (red) and non-target (blue) feature vectors and Y axis show any value. y=0 shows the mapped features for training data and y=1 shows the mapped feature vectors from the test data

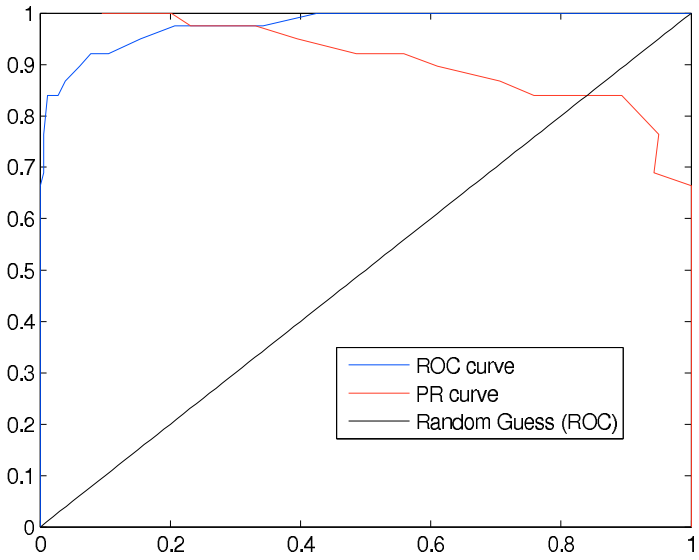



Figure 5.11: ROC and precision-recall curves created based on shifting the decision line along the combined feature axis. this curve is a result of one run of five fold cross validation for subject 4 and experiment one.

point to the right-most point. Figure 5.11 illustrates the ROC and precision recall curves created based on the data in Figure 5.10.

The test samples on the left of the decision line were classified as non-target and the test samples on the right side of the line were classified as targets. The rational behind moving the vertical decision line is that the more the decision line is shifted to the left, the greater the value of TP and naturally the greater the value of FP and vice versa. Thus, moving the decision line, acts like assigning different risks to our decision. More precisely, shifting the decision line towards left of the point of optimal minimum error Bayes classifier, would be as if we presume that detecting the targets have more importance than having false positives. In other words, we donate more importance to recall value than precision. Consequently, this classification scheme will result in decisions with high recall value but lower precision value. In other words, the output of the system will include many non-target images which were recognized as targets but on the other hand a good portion of target images will be also recognized correctly. Figure 5.12 illustrates the changing of ROC and precision recall curves for subject 4 when the image presentation speed increases.

As it can be seen in Figure 5.12, in both ROC and precision recall spaces, the performance of the classification algorithm decreases with the speed of time. In this figure three zones are clear. First, the high ROC and precision recall curves for experiment 1, second, a zone for experiment 2, 3 which look pretty similar but compared to zone 1 there is a relatively significant degradation of the curves. Third, a zone for experiment 4, and five where the performance degrades dramatically. The interesting point in this figure is that, even in the very fast experiment (IIP=IDP=50 ms) the performance of the classifier is above random guess, this can be due to the weak P300 benchmark seen in Figure 5.5 and also maybe partially due to the deference of target and non-target SSVEP signals as shown in Figure 5.7.

Tables 5.1, 5.2, 5.3, and 5.4 display the confusion matrix, obtained using the above-mentioned classification algorithm for subjects 1 to 4, respectively. Figure 5.13 illustrates the degradation of F measure, recall, and precision values for subject 4 (best subject) as the speed of image presentation increases and Figure 5.14 compares the F measure changes between all

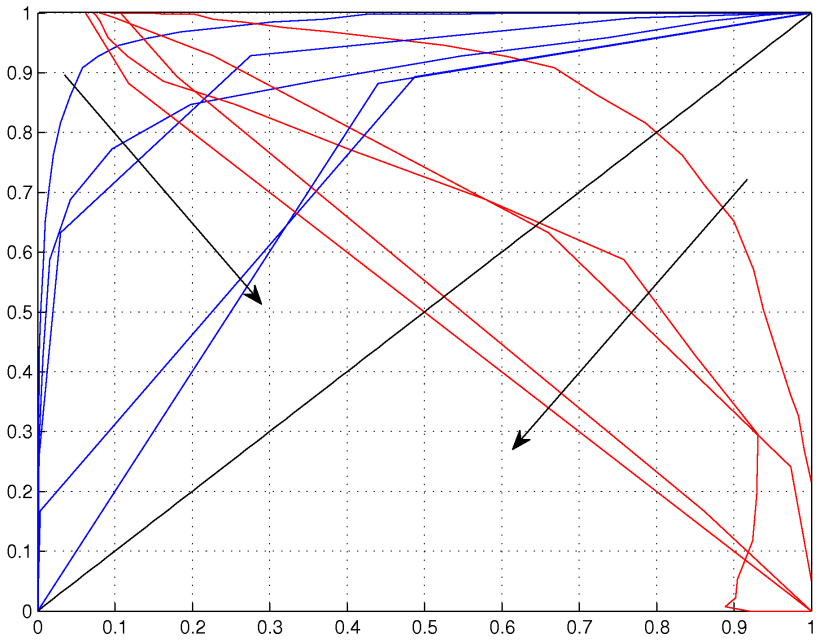



Figure 5.12: Change in the ROC and precision-recall curves with respect to different image presentation speeds. These curves are results of 15 runs of five fold cross validation for subject 4. The arrows underlines the deterioration of the performance of the classifier, as the speed is increased.

subjects. As it can be seen in Figure 5.13, this classification algorithm tries to keep the recall value constant and in return the recall value decreases as the speed increases. Therefore, as it can be seen the F measure also decreases as the speed increases.

Figure 5.14, explains that for all subjects there is a decreasing trend in F measure and in average, F measure degrades dramatically as the image presentation rate increases. It also shows that for most of the subjects, even in the fastest experiment, the F measure is above the F measure that can be obtained by random guess.

- the second classification algorithm which was used, is SVM with radial basis function kernels. To perform the classification using this method again five-fold cross validation was performed and it was repeated 15 times. In each run for training, first the extracted feature vectors was preprocessed using Principal Component Analysis (PCA) and the dimension of the feature vectors were reduced from 512 to 288. The Mixing matrix W was used to be applied later for dimension reduction of test data. In the next step, the training and test data were normalized to have the minimum and maximum values of zero and one, respectively. Finally, the training data was used to train the SVM classifier. To this end, a grid search for parameter $c$, was performed. However, the value of $g$ was fixed based on the information of the training data as follows. The radial basis kernel (also called gaussian kernel) can be expresses as:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2), \qquad for\ \gamma > 0$$

Therefore, to determine the value of $\gamma$ we computed the average distance between two feature vectors in the training set ($d$) and we set the gamma value as:

$$\gamma = \frac{1}{d}$$

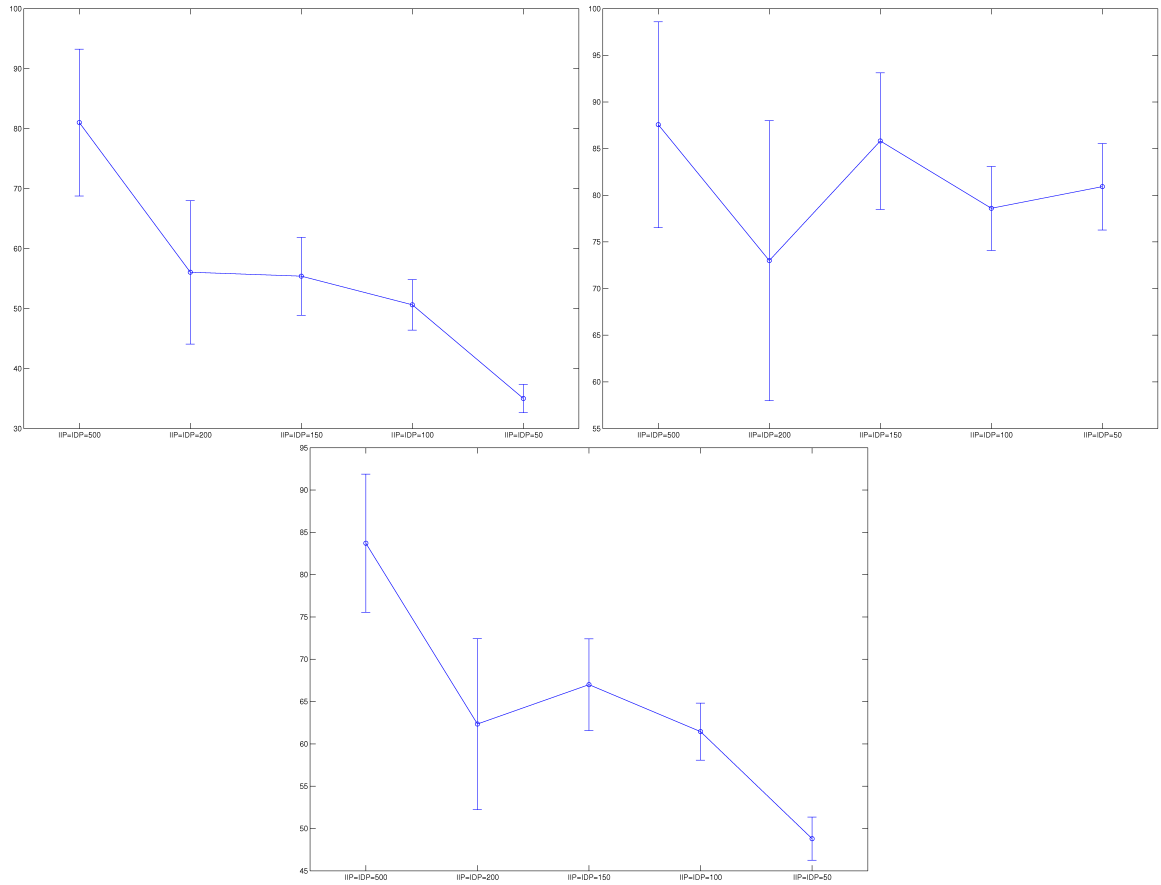



Figure 5.13: The changes in values of precision(top left), recall (top right), and F measure (bottom) for subject 4 during the experiments.

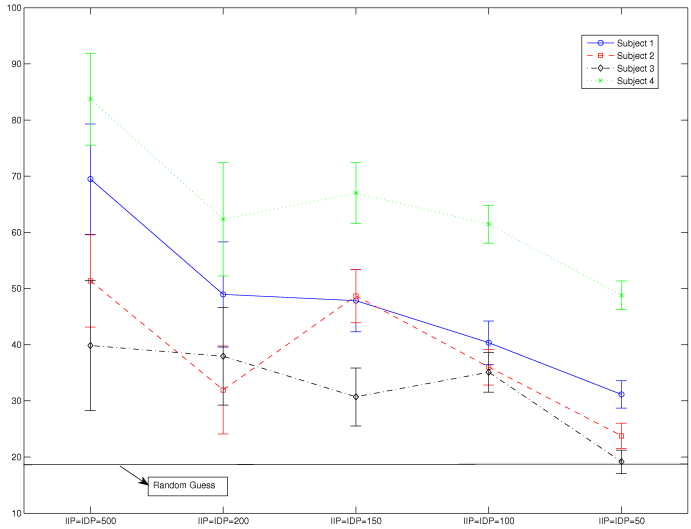



Figure 5.14: Comparison of changes in F measure values between all subjects with respect to speed.

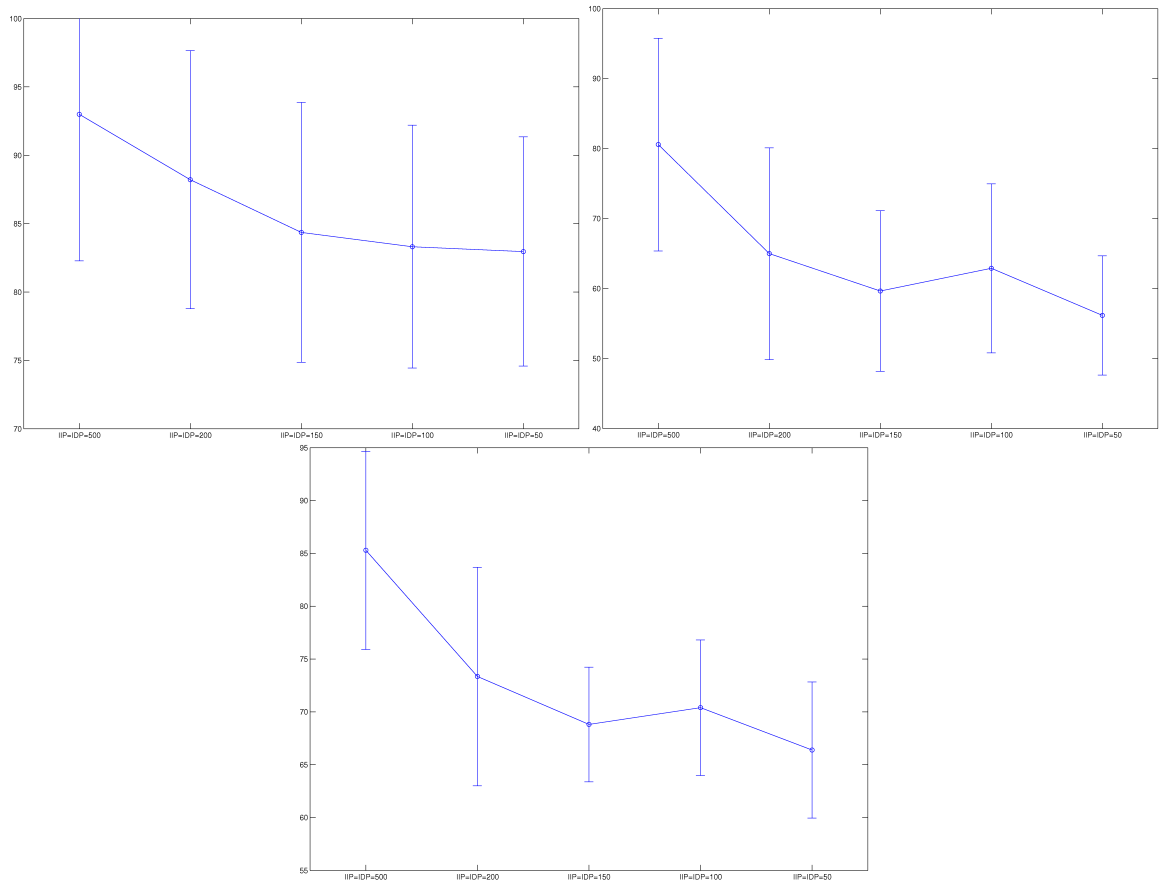



Figure 5.15: The changes in values of precision(top left), recall (top right), and F measure (bottom) for subject 4 during the experiments using SVM classifier.

In the next step the three-fold cross validation was performed on the training set (two folds as training and one fold as validation set) to determine the optimal value of $c$.

Tables 5.5, 5.6, 5.7, and 5.8 display the confusion matrix obtained when using the SVM classifier for subject 1 to subject 4, respectively. Figure 5.15 illustrates the decrease in the values of F measure, precision and recall for subject 4 ( best subject) as the image presentation rate increases and Figure 5.16 compares the F measure changes between all subjects. As it can be seen, SVM classifier will tend to keep the precision value constant. In other words, the number of false positives for this classifier is low. However, the recall rate will decrease dramatically as the image presentation rate increases and this leads to degradation of F measure.

Compared to results obtained by the first classification method, one can conclude that the former method will lead to high recall rate but low precision rate whereas, the latter results in higher precision and relatively lower recall. However, it is obvious that the results obtained by SVM are relatively higher in terms of F measure values compared to that of the former method.

## 5.2 Subconscious Perception

In this experiment, very high image presentation rate namely, 30 Hz and 60 Hz, were used. the subjects were asked again to report the number of times they could detect a target image ap-

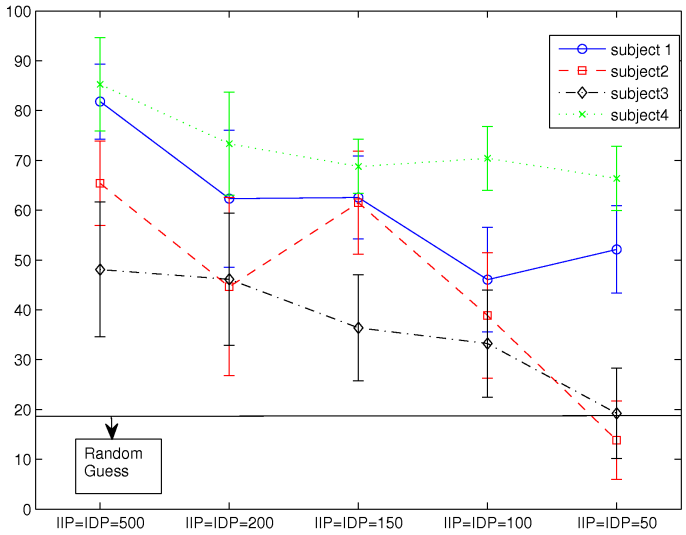



Figure 5.16: Comparison of changes in F measure values between all subjects with respect to speed using SVM classifier.

pearing on the screen. However, it was observed that due to the very high image presentation rate, all of the subjects were unable to detect any target images. In fact, using this experiment, it was observed that 32 ms and 16 ms are not enough to have a clear understanding of the image shown and thus this experiment only flashed some images to the subjects and the content of the images were mostly unknown for the subjects.

Considering the points explained above, we couldn't make any difference between the only target EEG segment and the average non-target EEG segments neither by visually monitoring them nor by processing them.

## 5.3 Learning

In this experiment, the same image sequence was repeated for five times and subjects were informed about this. However, due to the fast image presentation rate, the subjects gave different numbers of the detected target stimuli each time. Apart from this, the result obtained in this experiment didn't show any relationship between learning and degradation of the performance. The reason might be due to the fact that this experiment was performed at the last 10 minute of a very long acquisition session (100 minute long) and the subjects were mostly tired and not motivated to perform the test and stay concentrated. Even for subject 1 and subject 4 who had acceptable performance with the same image presentation rate before, the best F measure obtained were below 0.4 and didn't change much during different run.

## 5.4 Experience

### 5.4.1 Study of Averaged Signals

In this experiment another set of visual stimuli were used. The techniques which was applied for preprocessing, feature extraction and classification stages, were exactly the same techniques which was explained in previous chapter and sections. Figure 5.17 shows the averaged non-target and target signals taken from the PZ electrode. However, here in this experiment, two

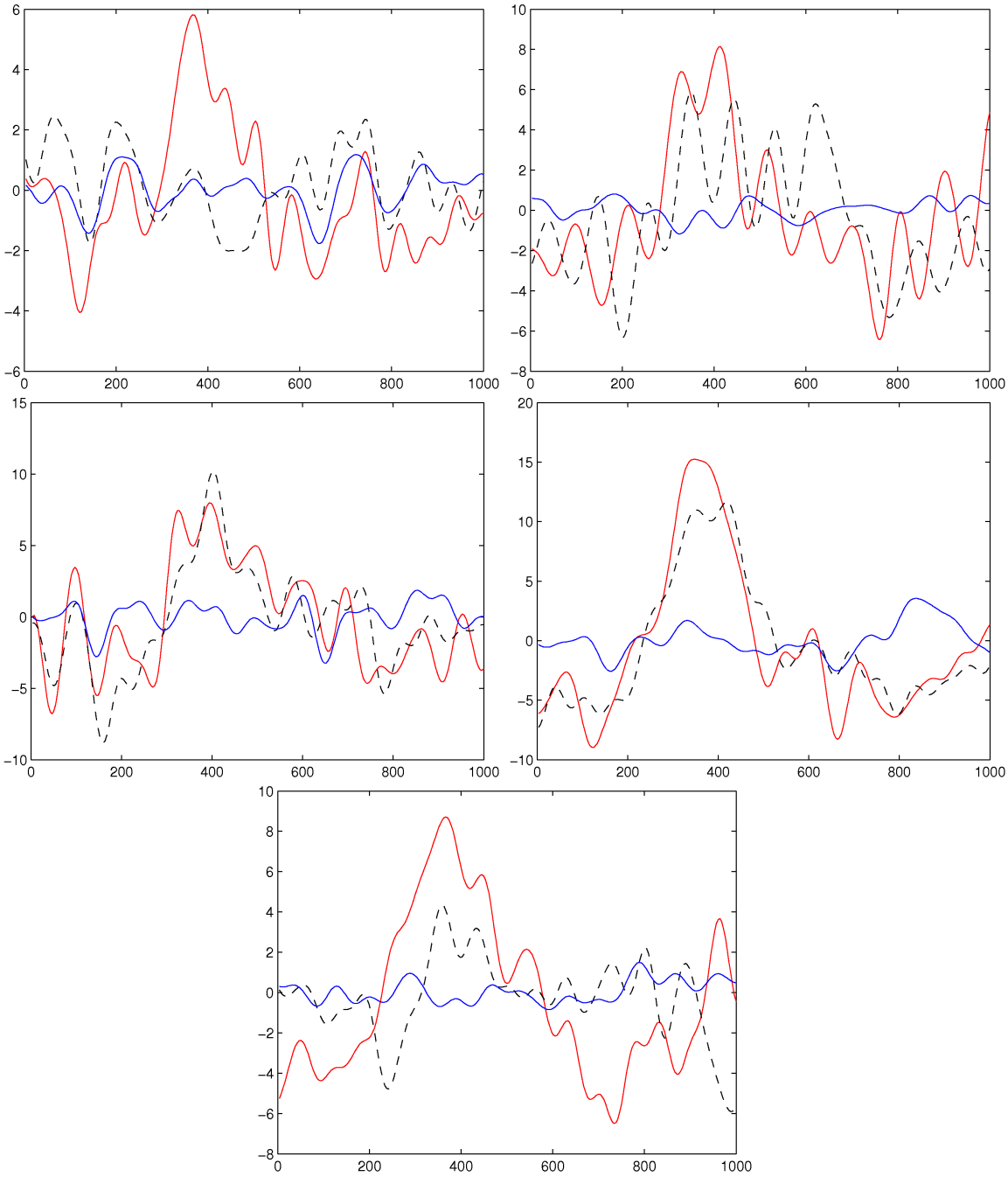



Figure 5.17: The averaged obvious target (red), non-obvious target (black) and nontarget (blue) EEG signals for different subjects (subject 5 (top left) to subject 9 bottom).

kinds of target stimuli were used. first group, namely obvious targets, were the target images which was shown before the experiment to all the subjects. the second group (non-obvious targets) were images which look somehow like obvious target images but in fact scientifically they are considered as non-targets (see Figure 3.2).

As it can be seen in Figure 5.17, most of the subjects were unable to distinguish between the obvious and non-obvious target images. This can be derived from the Figure 5.17 in that for most of the subjects there is a P300 peak occurring in EEG signal even when they see non-obvious targets. However, the amplitude of the two P300 signals (corresponding to obvious

and non-obvious targets) might be different, as the variety of non-obvious targets is big and some subject might classify some of the non-obvious images as non-target. Interestingly, the expert provided by ACT team in ESTEC was the only subject who could distinguish fully between the two subcategory of target images and his brain waves didn't show any P300 for non-obvious targets at all. The other fact that can be inferred from Figure 5.17 is that the amplitude of the P300 varies largely among the subjects. This amplitude depends on many biological factors.

### 5.4.2 Single Trial Analysis

As explained in section 4.3, 32 samples of the constructed signal which are corresponding to one-second long single trials were used as features. Once the feature vectors have been constructed from the single trials, Ten-fold cross validation was performed to create different train and test sets from the dataset. This was repeated for 15 times so that different training and test sets are constructed. In this experiment again FLDA+ Bayesian inference classifier and SVM classifier were used to classify the single trials.

Tables 5.9, 5.10, 5.11, 5.12, and 5.13 show the classification results for subject 5 to subject 9 using the two afore-mentioned classification algorithms.

## 5.5 Curiosity

In this experiment, the expert subject was asked to watch the images presented to him but this time he was not instructed to detect any target patterns. His task was simply to watch the image sequence and silently count how many times he saw a "scientifically interesting" image. In order to evaluate this data, the SVM classifier which was generated to detect the scientifically interesting images in the previous experiment was used again. In other words, the data gathered in the previous experiment from the expert subject was used to generate a classifier and this classifier was used to classify the data acquired in the current experiment. Table 5.14 shows the result of this classification. As it can be seen, an acceptable retrieval precision and F measure was obtained in this experiment. The results confirm the possibility of retrieval of scientifically interesting images in a dataset.

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $95.31 \pm 2.890$ | $23.93 \pm 13.17$ |
| Target | $4.69 \pm 2.890$ | $76.81 \pm 14.67$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $90.61 \pm 2.46$ | $27.88 \pm 16.75$ |
| Target | $9.39 \pm 2.46$ | $72.12 \pm 16.75$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $88.79 \pm 1.99$ | $28.03 \pm 9.50$ |
| Target | $11.21 \pm 1.99$ | $71.97 \pm 9.50$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $80.64 \pm 2.36$ | $34.08 \pm 6.47$ |
| Target | $19.36 \pm 2.36$ | $65.92 \pm 6.477$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $81.45 \pm 1.61$ | $29.81 \pm 6.30$ |
| Target | $18.55 \pm 1.61$ | $70.19 \pm 6.30$ |

(e)

Table 5.1: confusion matrix for subject 1 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $89.63 \pm 2.74$ | $31.35 \pm 17.37$ |
| Target | $10.37 \pm 2.74$ | $68.65 \pm 17.37$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $85.12 \pm 2.97$ | $44.19 \pm 15.20$ |
| Target | $14.88 \pm 2.97$ | $55.81 \pm 15.20$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $89.04 \pm 2.22$ | $27.36 \pm 9.30$ |
| Target | $10.96 \pm 2.22$ | $72.64 \pm 9.30$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $78.08 \pm 2.16$ | $37.88 \pm 7.04$ |
| Target | $21.92 \pm 2.16$ | $62.12 \pm 7.04$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $75.53 \pm 1.86$ | $36.61 \pm 6.92$ |
| Target | $24.47 \pm 1.86$ | $63.39 \pm 6.92$ |

(e)

Table 5.2: confusion matrix for subject 2 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $87.48 \pm 3.26$ | $45.02 \pm 17.46$ |
| Target | $12.52 \pm 3.26$ | $54.98 \pm 17.46$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $89.52 \pm 2.89$ | $44.62 \pm 15.37$ |
| Target | $10.48 \pm 2.89$ | $55.38 \pm 15.37$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $82.39 \pm 2.74$ | $45.00 \pm 10.69$ |
| Target | $17.61 \pm 2.74$ | $55.00 \pm 10.69$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $76.76 \pm 2.46$ | $37.49 \pm 07.39$ |
| Target | $23.24 \pm 2.46$ | $62.51 \pm 7.39$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $71.43 \pm 1.90$ | $43.60 \pm 7.16$ |
| Target | $28.57 \pm 1.90$ | $56.40 \pm 7.16$ |

(e)

Table 5.3: confusion matrix for subject 3 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.68 \pm 1.77$ | $12.45 \pm 11.03$ |
| Target | $2.32 \pm 1.77$ | $87.55 \pm 11.03$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $95.25 \pm 2.03$ | $27.01 \pm 15$ |
| Target | $4.75 \pm 2.03$ | $72.99 \pm 15$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $93.84 \pm 1.65$ | $14.20 \pm 7.33$ |
| Target | $6.16 \pm 1.65$ | $85.80 \pm 7.33$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $90.69 \pm 1.60$ | $21.41 \pm 4.50$ |
| Target | $9.31 \pm 1.60$ | $78.59 \pm 4.50$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $89.92 \pm 1.17$ | $19.09 \pm 4.64$ |
| Target | $10.08 \pm 1.17$ | $80.91 \pm 4.64$ |

(e)

Table 5.4: confusion matrix for subject 4 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.06 \pm 1.03$ | $23.93 \pm 13.17$ |
| Target | $4.69 \pm 2.890$ | $76.81 \pm 14.67$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.23 \pm 1.27$ | $42.84 \pm 17.02$ |
| Target | $1.77 \pm 1.27$ | $57.16 \pm 17.02$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.07 \pm 0.61$ | $49.05 \pm 9.41$ |
| Target | $0.93 \pm 9.41$ | $50.95 \pm 17.02$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.67 \pm 1.00$ | $65.72 \pm 11.06$ |
| Target | $1.33 \pm 1.00$ | $34.28 \pm 11.06$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.43 \pm 0.25$ | $61.12 \pm 9.22$ |
| Target | $0.57 \pm 0.25$ | $38.88 \pm 9.22$ |

(e)

Table 5.5: confusion matrix of SVM classifier for subject 1 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $97.56 \pm 1.55$ | $40.79 \pm 14.33$ |
| Target | $2.44 \pm 1.55$ | $59.21 \pm 14.33$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.32 \pm 0.72$ | $65.96 \pm 18.92$ |
| Target | $0.68 \pm 0.72$ | $34.04 \pm 18.92$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.99 \pm 0.77$ | $49.34 \pm 13.23$ |
| Target | $1.01 \pm 0.77$ | $50.66 \pm 13.23$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.03 \pm 0.76$ | $72.74 \pm 11.51$ |
| Target | $0.97 \pm 0.76$ | $27.26 \pm 11.51$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.85 \pm 0.22$ | $94.56 \pm 5.61$ |
| Target | $0.15 \pm 0.22$ | $5.44 \pm 5.61$ |

(e)

Table 5.6: confusion matrix of SVM classifier for subject 2 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.68 \pm 1.92$ | $69.29 \pm 21.48$ |
| Target | $1.32 \pm 1.92$ | $30.71 \pm 21.48$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.87 \pm 1.11$ | $63.69 \pm 15.86$ |
| Target | $1.13 \pm 1.11$ | $36.31 \pm 15.86$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.05 \pm 0.90$ | $75.26 \pm 11.71$ |
| Target | $0.95 \pm 0.90$ | $24.74 \pm 11.71$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.23 \pm 0.84$ | $84.60 \pm 13.34$ |
| Target | $0.77 \pm 0.84$ | $15.40 \pm 13.34$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.79 \pm 0.30$ | $89.04 \pm 6.74$ |
| Target | $0.21 \pm 0.30$ | $10.96 \pm 6.74$ |

(e)

Table 5.7: confusion matrix of SVM classifier for subject 3 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.22 \pm 1.56$ | $19.43 \pm 15.20$ |
| Target | $0.78 \pm 1.56$ | $80.57 \pm 15.20$ |

(a)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.23 \pm 0.64$ | $35.02 \pm 15.13$ |
| Target | $0.77 \pm 0.64$ | $64.98 \pm 15.13$ |

(b)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.92 \pm 0.75$ | $40.35 \pm 11.48$ |
| Target | $1.08 \pm 0.75$ | $59.65 \pm 11.48$ |

(c)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.26 \pm 1.28$ | $37.12 \pm 12.08$ |
| Target | $1.74 \pm 1.28$ | $62.88 \pm 12.08$ |

(d)

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.18 \pm 0.48$ | $43.84 \pm 8.52$ |
| Target | $0.82 \pm 0.48$ | $56.16 \pm 8.52$ |

(e)

Table 5.8: confusion matrix of SVM classifier for subject 4 in experiments a. 1, b. 2, c. 3, d. 4, e. 5

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $84.45 \pm 3.65$ | $37.78 \pm 13.31$ |
| Target | $15.55 \pm 3.65$ | $62.22 \pm 13.31$ |

(a) $F = 0.4493 \pm 0.0766$, $prec = 0.3562 \pm 0.0627$

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.55 \pm 1.49$ | $56.660 \pm 17.35$ |
| Target | $1.45 \pm 1.49$ | $43.330 \pm 17.35$ |

(b) $F = 0.5641 \pm 0.1565$, $prec = 0.8304 \pm 0.1262$

Table 5.9: confusion matrix of SVM classifier for subject 5 in experience experiment a. classification using FLDA+ Bayes inference, b. classification using SVM

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $90.56 \pm 3.13$ | $28.44 \pm 12.79$ |
| Target | $9.44 \pm 3.13$ | $71.56 \pm 12.79$ |

(a) $F = 0.6006 \pm 0.0923$, $prec = 0.5259 \pm 0.0945$

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $97.13 \pm 2.23$ | $32.67 \pm 15.01$ |
| Target | $2.87 \pm 2.23$ | $67.33 \pm 15.01$ |

(b) $F = 0.7114 \pm 0.0899$, $prec = 0.7910 \pm 0.1046$

Table 5.10: confusion matrix of SVM classifier for subject 6 in experience experiment a. classification using FLDA+ Bayes inference, b. classification using SVM

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $84.05 \pm 3.97$ | $32.22 \pm 14.32$ |
| Target | $15.95 \pm 3.97$ | $67.78 \pm 14.32$ |

(a) $F = 0.4816 \pm 0.0712$ , $prec = 0.3789 \pm 0.0712$

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $97.79 \pm 1.84$ | $53.33 \pm 21.92$ |
| Target | $2.21 \pm 1.84$ | $46.67 \pm 21.92$ |

(b) $F = 0.5685 \pm 0.1559$ , $prec = 0.7675 \pm 0.1476$

Table 5.11: confusion matrix of SVM classifier for subject 7 in experience experiment a. classification using FLDA+ Bayes inference, b. classification using SVM

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $89.11 \pm 3.65$ | $29.68 \pm 9.81$ |
| Target | $10.89 \pm 3.65$ | $70.33 \pm 9.81$ |

(a) $F = 0.5661 \pm 0.0688$ , $prec = 0.4830 \pm 0.0846$

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $97.12 \pm 2.19$ | $34.67 \pm 14.37$ |
| Target | $2.88 \pm 2.19$ | $65.33 \pm 14.37$ |

(b) $F = 0.6956 \pm 0.0777$ , $prec = 078.99 \pm 0.1304$

Table 5.12: confusion matrix of SVM classifier for subject 8 in experience experiment a. classification using FLDA+ Bayes inference, b. classification using SVM

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $77.79 \pm 5.35$ | $52.56 \pm 14.30$ |
| Target | $22.21 \pm 5.35$ | $47.44 \pm 14.30$ |

(a) $F = 0.3047 \pm 0.0842$ , $prec = 0.2270 \pm 0.0636$

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $99.14 \pm 1.05$ | $81.67 \pm 11.28$ |
| Target | $0.86 \pm 1.05$ | $18.33 \pm 11.28$ |

(b) $F = 0.3498 \pm 0.1042$ , $prec = 0.7867 \pm 0.1868$

Table 5.13: confusion matrix of SVM classifier for subject 9 in experience experiment a. classification using FLDA+ Bayes inference, b. classification using SVM

| Prediction \ Ground Truth | Non-target | Target |
|---|---|---|
| Non-target | $98.45 \pm 1.67$ | $72.67 \pm 19.02$ |
| Target | $1.55 \pm 1.67$ | $27.33 \pm 19.02$ |

Table 5.14: confusion matrix of SVM classifier for subject 5 in curiosity experiment using SVM. $F = 0.4244 \pm 0.1514$ , $prec = 0.7462 \pm 13.82$

# Chapter 6

# Conclusion

## 6.1 General Observations

In this Research, an efficient P300-based BCI system for classification of brainwaves associated to scientific stimuli was presented. It was shown that relatively high classification accuracies can be obtained for users of this system. Due to the use of the P300, only a small amount of training was required to achieve good classification accuracy for each subject. In other words, since for each individual, the amplitude and latency of P300 pattern varies only in a small dynamic range, only with a small number of training data, one can design a classifier to detect the P300 patterns for each subject individually. On the contrary, in research areas like studying the EEG changes during mental actions, psychiatric phenotypes, Audio/ visual stimulation, music expossure, etc. one needs to gather a huge amount of training data to rule out the generalization problem. It has been observed that increasing the speed of image presentation, will decrease the classification accuracy, however with image presentation frequency of 3.33 Hz, it is still possible to have relatively good classification result.

Concerning the relative performance of expert and non-expert subjects we have seen that the expert can distinguish between the obvious target and non-obvious target whereas the non-expert subjects were mostly considered both obvious and non-obvious targets as one category and were mostly unable to distinguish between them. Despite the fact that a relatively high classification accuracy can be obtained for each subject individually, but since the classification accuracy varies from subject to subject and mainly depends on the level of concentration, activeness, etc. of the subjects during experiment, it is a challenging problem to develop a general classification scheme which can work for naive subjects who have never been through a training phase before.

A comparison between the machine learning algorithms FLDA+Bayesian inference and SVM revealed that SVM clearly outperforms the former method. This was especially the case when high-dimensional feature vectors, resulting from the usage of many electrodes, were employed.

## 6.2 Differences to Other Studies

Compared to other P300-based BCI systems for target stimuli recognition, the classification accuracy is relatively high (see [36, 37]). Due to differences in experimental paradigms and subject populations the classification accuracy and bitrate obtained in other studies cannot be compared directly to those obtained in the present study. Nevertheless, several factors that might have caused the differences can be identified. These factors are described below.

- Probability of target stimuli appearance

  In the present study the probability of target stimuli was mostly %10, whereas in the experiments of Sellers and Donchin [36] and Piccione et al. [37] four-choice paradigms were used. As a consequence the target stimulus occurred with a probability of %25 in their experiments. Smaller target probabilities correspond to higher P300 amplitudes ([38]), thus the P300 in our system might have been easier to detect. In general, when designing a typical P300-based BCI, one has to take into account that subjects have to focus on a relatively small area of the display might (i.e. P300 speller). In this study, however, relatively large images were used that could help the subjects stay more concentrated.

- Inter-stimulus Interval (ISI)

  Several factors have to be kept in mind when choosing an ISI (IDP+IIP) for a P300-based BCI system. Regarding classification accuracy, longer ISIs, theoretically, yield better results. This should be the case because longer ISIs (within some limits) cause larger P300 amplitude. On the other hand, a consequence of long ISIs is a longer overall duration of runs. Many subjects might have difficulties to stay concentrated during long runs (as it was observed in learning experiment) and thus P300 amplitude and classification accuracy might actually decrease for longer ISIs. Given the complex interrelationship of several factors an optimal ISI for P300-based BCIs can only be determined experimentally. Here we have shown that an ISI of 500 ms, 400 ms, and 200 ms yield good results. Sellers and Donchin have used an ISI of 1.4 s, and Piccione et al. have used an ISI of 2.5 s. The results obtained in their studies seem to indicate that these ISIs are too long.

## 6.3 Visual Evoked Potentials

In the literature on P300-based BCI systems it is almost always assumed that the only factor allowing to discriminate target trials from non-target trials is the P300 (see [39] for an exception). However, for systems using visual stimuli this assumption might be too limited. In fact, the visual stimuli used in this study are at the center of the visual field and influence a relatively large part of visual cortex. Hence, the visual evoked potentials (VEPs) corresponding to target flashes can be also expected to differ from the VEPs corresponding to non-target flashes. This is of more importance when faster acquisition protocols are used and as it was shown in this study for fast image presentation rates, SSVEPs will appear significantly in the EEG signals. Therefore, for the system presented here, the plots of the average waveforms in the target and non-target conditions provide evidence that the P300 plays an important role for the classification of targets and non-targets. However, the possibility that the classification accuracy depends partly on the VEP responses cannot be excluded. Further research is necessary to elucidate the role of P300 and VEPs in P300-based BCI systems.

## 6.4 Electrode Configurations

The electrode configuration used in a BCI determines the suitability of the system for daily use. Clearly, systems that use only few electrodes take less time for setup and are more user friendly than systems with many electrodes. However, if too few electrodes are used not all features that are necessary for accurate classification can be captured and communication speed decreases. For P300-based BCI systems different electrode configurations have been described in the literature. Good results have been reported using only three or four midline electrodes (Fz, Cz, Pz,Oz) ([36, 37, 40]). [41], described that an eight electrode configuration consisting of the

midline electrodes and the four parietal-occipital electrodes PO7, PO8, P3, and P4. [39] employed a ten electrode configuration consisting of the midline electrodes, the parietal-occipital electrodes PO7, P08, P3, P4 and the central electrodes C3, C4. Finally [42] used a set of 25 central and parietal electrodes.

In summary, regardless of the classification algorithm that is used, the eight electrode configuration represents a good compromise between suitability for daily use and classification accuracy and seems to capture most of the important features for P300 classification.

## 6.5 Machine Learning Algorithms

Many of the characteristics of a BCI system depend critically on the employed machine learning algorithm. Important characteristics that are influenced by the machine learning algorithm are classification accuracy and communication speed, as well as the amount of time and user intervention necessary for setting up a classifier from training data. In this study, FLDA+Bayesian inference and SVM techniques were used for the classification of single trials and it has been shown that SVM offers good classification accuracy and does not constrain the practical applicability of a BCI system and is thus an interesting alternative to FLDA.

## 6.6 Human Curiosity and its Cloning

In this report we study an EEG based system intended to extract informations on human curiosity, or scientific interest. We prove that the P300 wave can be reliably detected and associated to images selected by the subject during a rapid visual image presentation experiment. We cannot conclude from the experiments and data collected during this project that such a selection can indeed be based on the subject scientific expertise, or on its curiosity. Some partial results have though been presented and indicate that such a possibility is concrete.

# Bibliography


[1] D. Izzo, L. Rossini, M. Rucinski, C. Ampatzis, G. Healy, P. Wilkins, A.F. Smeaton, A. Yazdani, and T. Ebrahimi. Curiosity cloning: neural analysis of scientific interest. 2009.

[2] G. Healy, Smeaton A.F. Wilkins, P., D. Izzo, M. Rucinski, C. Ampatzis, and E. Moraud. Curiosity cloning: Neural modelling for image analysis. Technical Report 08-8201b, European Space Agency, the Advanced Concepts Team, 2010.

[3] S. Chien, R. Sherwood, D. Tran, B. Cichy, G. Rabideau, R. Castano, A. Davies, D. Mandl, S. Frye, B. Trout, S. Shulman, and D. Boyer. Using autonomy flight software to improve science return on earth observing one. *Journal of Aerospace Computing, Information, and Communication.*, 2005.

[4] L. Mandrake, K. Wagstaff, D. Gleeson, U. Rebbapragada, D. Tran, R. Castano, S. Chien, and R. Pappalardo. Onboard detection of naturally occurring sulfur on a glacier via an onboard svm and hyperion/eo-1. In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009.

[5] A. Castano, A. Fukunaga, J. Biesiadecki, L. Neakrase, P. Whelley, R. Greeley, M. Lemmon, R. Castano, and S. Chien. Automatic detection of dust devils and clouds on Mars. *Machine Vision and Applications*, 19(5):467–482, 2008.

[6] D. Hayden, S. Chien, D. Thompson, and R. Castano. Onboard clustering of aerial data for improved science return. In *Proceedings of the IJCAI-09 Workshop on Artificial Intelligence in Space*, 2009.

[7] T. Hruby and P. Marsalek. Event-related potentials-the P3 wave. *Acta Neurobiologiae Experimentalis*, 63(1):55–63, 2002.

[8] A. Gerson, L. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Transactions on neural systems and rehabilitation engineering*, 14(2):174–179, 2006.

[9] M. Hoshiyama, R. Kakigi, S. Watanabe, K. Miki, and Y. Takeshima. Brain responses for the subconscious recognition of faces. *Neuroscience research*, 46(4):435–442, 2003.

[10] ITU Radiocommunication Assembly. Methodology for the subjective assessment of the quality of television pictures. Technical report, 1974-2002. URL http://www.dii.unisi.it/~menegaz/DoctoralSchool2004/papers/ITU-R_BT.500-11.pdf.

[11] M. Rucinski. Curiosity cloning image viewer user's manual. Technical Report CCIVUM01, European Space Agency, the Advanced Concepts Team, 2008. URL http://www.esa.int/gsp/ACT/doc/INF/pub/ACT-MAN-5100-CCIVUM01.pdf. Available on line at http://www.esa.int/act.

[12] M. Vetterli and C. Herley. Wavelets and filter banks: Theory and design. *IEEE Transactions on Signal Processing*, 40(9):2207–2232, 1992.

[13] H. Adeli, Z. Zhou, and N. Dadmehr. Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, 123(1):69–87, 2003.

[14] J. Gotman. The use of computers in analysis and display of EEG and evoked potentials. *Current practice of clinical electroencephalography*, pages 51–83, 1990.

[15] I. Daubechies et al. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math*, 41(7):909–996, 1988.

[16] V. Kolev, T. Demiralp, J. Yordanova, A. Ademoglu, and
"U. Isoglu-Alkaç. Time-frequency analysis reveals multiple functional components during oddball P300. *NeuroReport*, 8(8):2061, 1997.

[17] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Citeseer, 2001.

[18] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM New York, NY, USA, 1992.

[19] Schölkopf, B. and Tsuda, K. and Vert, J.P. Kernel methods in computational biology. *MIT press Cambridge, MA*, 2004.

[20] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge Univ Pr, 2004.

[21] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[22] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[23] F. Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*, 2000.

[24] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines, 2001.

[25] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453. Citeseer, 1998.

[26] C. Drummond and R.C. Holte. What ROC curves can't do (and cost curves can). In *Proceedings of the ROC Analysis in Artificial Intelligence, 1st International Workshop*, pages 19–26. Citeseer, 2004.

[27] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 2002.

[28] V. Raghavan, P. Bollmann, and G.S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

[29] J. Bockhorst and M. Craven. Markov networks for detecting overlapping elements in sequence data. In *Neural Information Processing Systems*, volume 17. Citeseer, 2005.

[30] R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, R.J. Mooney, A.K. Ramani, and Y.W. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, 2005.

[31] J. Davis, E. Burnside, I. Dutra, D. Page, R. Ramakrishnan, V.S. Costa, and J. Shavlik. View learning for statistical relational learning: With an application to mammography. In *Proceeding of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland*. Citeseer, 2005.

[32] KS Shanmugam. Digital and analog communication systems. *NASA STI/Recon Technical Report A*, 80:23225, 1979.

[33] J.R. Hughes. Gamma, fast, and ultrafast waves of the brain: Their relationships with epilepsy and behavior. *Epilepsy and Behavior*, 13(1):25–31, 2008.

[34] I. Gold. Does 40-Hz oscillation play a role in visual consciousness? *Consciousness and Cognition*, 8(2):186–195, 1999.

[35] W. Singer and CM Gray. Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, 18(1):555–586, 1995.

[36] E.W. Sellers and E. Donchin. A P300-based brain–computer interface: initial tests by ALS patients. *Clinical Neurophysiology*, 117(3):538–548, 2006.

[37] F. Piccione, F. Giorgi, P. Tonin, K. Priftis, S. Giove, S. Silvoni, G. Palmas, and F. Beverina. P300-based brain computer interface: reliability and performance in healthy and paralysed participants. *Clinical neurophysiology*, 117(3):531–537, 2006.

[38] C.C. Duncan-Johnson and E. Donchin. On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, 14(5):456–467, 1977.

[39] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter. BCI Competition 2003 —Data Set IIb: Support Vector Machines for the P 300 Speller Paradigm. *IEEE Transactions on Biomedical Engineering*, 51(6):1073–1076, 2004.

[40] H. Serby, E. Yom-Tov, and G.F. Inbar. An improved P300-based brain-computer interface. *IEEE Transactions on neural systems and rehabilitation engineering*, 13(1):89–98, 2005.

[41] D.J. Krusienski, E.W. Sellers, F. Cabestaing, S. Bayoudh, D.J. McFarland, T.M. Vaughan, and J.R. Wolpaw. A comparison of classification techniques for the P300 speller. *Journal of neural engineering*, 3:299–305, 2006.

[42] M. Thulasidas, C. Guan, and J. Wu. Robust classification of EEG signal for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(1):24, 2006.