Functional near infrared spectroscopy in the multimodal assessment of working memory impairments following traumatic brain injury

A Thesis
Submitted to the Faculty
of
Drexel University
by
Anna Caterina Merzagora
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
May 2010
DEDICATION

To my parents.
ACKNOWLEDGMENTS

I wish to thank most sincerely all the people that made this research possible through their support, advice and understanding.

First, I would like to thank my supervisors and mentors, Dr. Banu Onaral, Dr. Maria Schultheis and Dr. Meltem Izzetoglu. The multidisciplinary nature of this work has in fact been made possible by his unique team of advisors. In particular, I would like to thank Dr. Onaral for her trust and guidance, for the sincere encouragement to succeed in my career and for being a great source of inspiration. I would also like to thank Dr. Schultheis for her mentorship and for having educated me to interact successfully with researchers outside my field of expertise. Last but not least, I would like to recognize Dr. Meltem Izzetoglu: with her I shared many profitable discussions about science and life, she kindly guided me and always provided precious feedback on my work.

I would also like to extend my gratitude to Dr. Kambiz Pourrezaei, Dr. Patricia A. Shewokis, Dr. Kurtulus Izzetoglu and Dr. Robi Polikar. They all provided valuable guidance and constructive criticism at different stages of this research. Many thanks go also to Dr. Antonio Oliviero and Dr. Guglielmo Foffani at the Fundación Hospital Nacional de Parapléjicos para la Investigación y la Integración in Toledo (Spain) for kindly hosting me in their lab and for making a valuable collaboration possible. A grateful thanks goes also to my friends and colleagues in the Optical Brain Imaging team and to all staff members in the department.
Table of Contents

LIST OF TABLES ........................................................................................................... VIII
LIST OF FIGURES .......................................................................................................... IX
ABSTRACT ....................................................................................................................... X

CHAPTER 1: INTRODUCTION ......................................................................................... 1
  1.1 MOTIVATION ....................................................................................................... 1
  1.2 APPROACH AND CONTRIBUTION ..................................................................... 4
  1.3 CONCLUSIONS AND FUTURE WORK ............................................................... 7

CHAPTER 2: DISCRIMINATIVE VALIDATION OF FNIRS ......................................... 9
  2.1 INTRODUCTION ................................................................................................. 9
  2.2 METHODS .......................................................................................................... 11
    2.2.1 Subjects ....................................................................................................... 11
    2.2.2 Neuropsychological evaluation .................................................................... 12
    2.2.3 Experimental paradigm ............................................................................... 14
    2.2.4 fNIRS data acquisition ............................................................................... 17
    2.2.5 fNIRS data processing and analysis ............................................................. 18
  2.3 RESULTS ............................................................................................................ 21
    2.3.1 Subjects demographic and neuropsychological assessment ....................... 21
    2.3.2 N-back behavioral performance ................................................................... 22
    2.3.3 Neuroimaging data ...................................................................................... 23
  2.4 DISCUSSION ....................................................................................................... 28

CHAPTER 3: CLASSIFICATION CAPABILTIY OF FNIRS ........................................ 36
3.1 INTRODUCTION ........................................................................................................ 36

3.2 METHODS.................................................................................................................. 37

3.2.1 Subjects and experimental paradigm ................................................................. 37

3.2.2 fNIRS data acquisition and processing ............................................................... 38

3.2.3 fNIRS-based classification .................................................................................. 38

3.2.4 Characterization of classification performance ................................................... 40

3.3 RESULTS................................................................................................................... 42

3.4 DISCUSSION............................................................................................................. 44

CHAPTER 4: CLASSIFICATION CAPABILITY OF MULTIMODAL APPROACH .. 47

4.1 INTRODUCTION ...................................................................................................... 47

4.2 METHODS................................................................................................................ 49

4.2.1 Subjects and experimental paradigm .................................................................. 49

4.2.2 Data acquisition and processing ....................................................................... 50

4.2.3 Classification ..................................................................................................... 51

4.2.3 Characterization and statistical evaluation of classification performance ........... 54

4.3 RESULTS................................................................................................................... 55

4.3.1 EEG results ....................................................................................................... 55

4.3.2 Classification results ........................................................................................ 56

4.4 DISCUSSION............................................................................................................. 60

CHAPTER 5: DEMONSTRATION OF EXPLORATORY APPLICATIONS .......... 65

5.1 INTRODUCTION ...................................................................................................... 65

5.2 EFFECTS OF ANODAL tDCS ON PREFRONTAL HEMODYNAMIC ACTIVITY ....... 66

5.2.1 Introduction ....................................................................................................... 66

5.2.2 Methods ............................................................................................................ 68

5.2.3 Results .............................................................................................................. 72
## List of Tables

2.1 Demographics and neuropsychological assessment of participants ........................................... 22
2.2 Behavioral performance in n-back task ......................................................................................... 23
2.3 Between-group comparison, averaging across load and region .................................................... 24
2.4 Between-group comparison, separately for region and averaging across load ............................... 24
3.1 Descriptives of fNIRS-based classification and permutation test results .................................... 43
4.1 P300 peak amplitude by load condition and overall ..................................................................... 56
4.2 Descriptives of results for different classification approaches .................................................... 57
4.3 Effect sizes comparing single-modality and multimodal classification ...................................... 59
4.4 Effect sizes comparing fNIRS-based and EEG-based classification ........................................... 59
5.1 Behavioral performance in n-back task ......................................................................................... 82
5.2 Behavioral performance in ‘Low performing’ and ‘High performing’ groups ............................... 86
5.3 Classification accuracy for different classification approaches .................................................... 88
List of Figures

2.1 N-back protocol........................................................................................................................................16
2.2 Hemodynamic response compared between controls and TBI subjects........................................26
2.3 Relation between response time and metabolic ratio in healthy controls.................................27
2.4 Relation between % Correct and left DLPFC $H_{\max}$ in TBI subjects........................................28
3.1 Confusion matrix........................................................................................................................................39
3.2 fNIRS-based classification compared with permutation test results........................................43
4.1 Comparison of performances obtained by different classification approaches........................57
5.1 Schematic representation of tDCS experimental setup.................................................................70
5.2 Spatio-temporal hemodynamic patterns following real tDCS....................................................73
5.3 Spatio-temporal $H_{\text{bO}}$ patterns following real and sham tDCS...............................................74
5.4 Temporal $H_{\text{bO}}$ course following 15 min real tDCS.................................................................76
5.5 EEG and fNIRS features for ‘Low performing’ and ‘High performing’ groups......................87
A.1 Random subspace method applied to fNIRS-based classification..............................................93
A.2 Pseudo-code for modified leave-one-out (mLOO) algorithm.......................................................95
A.3 SVM k-fold calibration for fNIRS-based classification.................................................................96
A.4 Bagging applied to EEG-based classification...............................................................................98
A.5 Pseudo-code for mLOO algorithm for EEG-based classification..............................................99
A.6 SVM k-fold calibration for EEG-based classification.................................................................100
A.7 Random subspace method applied to feature-level fusion classification..........................101
A.8 Pseudo-code for mLOO algorithm for feature-level fusion classification........................102
A.9 Random subspace method and bagging for applied to decision-level fusion..................103
A.8 Pseudo-code for mLOO algorithm for decision-level fusion classification......................104
Abstract

Functional near infrared spectroscopy in the multimodal assessment of working memory impairments following traumatic brain injury
Anna Caterina Merzagora
Meltem Izzetoglu, Ph.D.; Banu Onaral, Ph.D.; Maria Schultheis, Ph.D.

A frequent consequence of traumatic brain injury (TBI) is cognitive impairment, which results in significant disruption of an individual’s everyday living. To date, most clinical rehabilitation interventions still rely on behavioral observation, with little or no quantitative information about physiological changes produced at the brain level.

Functional brain imaging modalities have been extensively used in the study of cognitive impairments following TBI. However, the applications of these technologies to rehabilitation have been limited. This is due in part to the expensive or invasive nature of these modalities or because they rely on experimental tasks that are not ecologically valid in reference to real-world behaviors. Additionally, studies of cognitive impairments have most commonly depended on a single imaging modality. However, different modalities glean differ aspects of the brain activity and each may offer distinct and often complementary strengths. Therefore, combining multiple technologies could offer improved understanding of brain-behavior relationship under pathological conditions.

The objective of this study is to apply, for the first time, functional near infrared spectroscopy (fNIRS), and its integration with electroencephalography (EEG), to the assessment of working memory impairments following TBI. fNIRS provides a localized measure of prefrontal hemodynamic activation, which is susceptible to TBI, and it does
so in a noninvasive, affordable and wearable way, thus partially overcoming the limitations of other modalities. EEG offers a cost-effective and simple measure of brain electrical activity and it has been employed in a few studies on traumatic brain injury, showing abnormal patterns of neural activity. The combination of two modalities therefore offers information about the spatial location of the recorded activity and it takes advantage the good temporal resolution of EEG.

Participants included six TBI subjects and eleven healthy controls. Standard neuropsychological tests probing attention and working memory were administered. Brain activation measurements were collected during a visual n-back task, designed to incrementally vary the working memory load and often used in neuroimaging studies.

Overall, the results from the research presented in this thesis provide first evidence of the ability of fNIRS to reveal differences between TBI and healthy subjects in working memory tasks, suggesting a dysfunction in the matching between cognitive demands and cortical resources in TBI subjects. Moreover, this study has demonstrated that fNIRS measures can distinguish between the two groups and these data have been investigated relative to the results obtained by EEG alone or by the combination of fNIRS and EEG. Each of the two modalities revealed unique strengths that can contribute to the classification between the two groups. Therefore, successful combination of fNIRS and EEG for this application takes advantage of the complementary strengths of the individual modalities in order to improve the classification between normal and TBI cases.
CHAPTER 1: INTRODUCTION

1.1 Motivation

A commonly observed consequence of traumatic brain injury (TBI) is cognitive impairment (Binder, 1986; Levin and Grossman, 1978; McMillan, 1997), which results in significant disruption of an individual’s everyday living. Furthermore, while several neurorehabilitation strategies are available, the choice of a successful treatment still relies on behavioral observation and little or no quantitative information is available about the physiological changes produced at the brain level by the specific intervention.

Functional brain imaging represents an important tool for the investigation of brain function. Functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have been extensively used in the study of cognitive impairments following TBI (Belanger et al., 2007; Dubroff and Newberg, 2008; Strangman et al., 2005). Functional neuroimaging techniques, especially PET and fMRI, have been applied to the examination of cognitive impairment after TBI. Early work using PET examined differences in activation patterns between TBI and matched healthy controls using paradigms of attention and vigilance (Humayun et al., 1989; Ruff et al., 1994). Performance on measures of executive functioning (i.e., Wisconsin Card Sorting Task) and prefrontal cortical activation have also been examined using PET (Kirkby et al., 1996; Lombardi et al., 1999). More recent work has examined spatial working memory and aspects of verbal memory using PET (Chen et al., 2003; Levine et al., 2002; Ricker et
al., 2001). The technological development of fMRI has made this technique the leading modality for the neuroimaging of cognition. In TBI, we are aware of only two studies that have employed functional neuroimaging to directly examine the effect of a rehabilitation intervention. In the first study, Kim and colleagues (2004) examined cerebral organizations using fMRI after a rehabilitation intervention (i.e., constrained induced therapy) for patients with TBI and stroke. Specifically, the researchers scanned participants both before and after a 2-week, 7hrs/day constrained induced movement therapy intervention and found brain activation ipsilateral to the injury before the intervention and a switch to more typical contralateral activation after the intervention. In the second study, Strangman and colleagues (2008) evaluated the ability to predict the outcomes of memory rehabilitation using fMRI measures in combination with age, education, injury severity and pre-intervention memory scores. The findings of this study supported the notion that left prefrontal activity is related to strategic verbal learning and the magnitude of this activation predicted success in response to cognitive memory rehabilitation strategies. These two studies serve to exemplify the potential benefits of neuroimaging technologies in rehabilitation research, by demonstrating that functional neuroimaging can help characterize the neurophysiological changes underlying recovery.

Yet, to date, the application of these technologies for rehabilitation purposes has been limited.

In part, this is due to the expensive or invasive nature of these neuroimaging modalities or due to their reliance on experimental tasks that are not ecologically valid in reference to real-world functional behavior. In fact, both fMRI and PET have significant technological and methodological constraints that limit their full application in the evaluation of rehabilitation, including a dependence on expensive high-end technologies (e.g., requiring high-field magnet or a cyclotron for the production of radioactive tracers).
Additionally, these technologies are not portable, therefore limiting their working conditions and preventing from the deployment in the clinician’s office or in the field (i.e. during rehabilitation therapies at home or in the community). fMRI and PET also suffer from low temporal resolution and are highly sensitive to head movements. These techniques also rely on experimental tasks that are typically artificial, simplistic and not ecologically valid in reference to real-world functional behaviors which are the target of rehabilitation interventions. Finally, PET has the added drawback of employing ionizing radiation, which does not allow for repeated scans. In sum, these limitations circumscribe the utility of PET and fMRI in supporting rehabilitation’s objectives of improving functional recovery after TBI. Two additional aspects need also to be taken into consideration. First, existing neuroimaging studies are based on protocols designed to investigate only one discrete cognitive domain at a time. By contrast, functional behavior is more complex and results from the interaction of multiple components (e.g. attention, orienting response, working memory or planning). Second, most neuroimaging studies of TBI-related cognitive impairments have so far depended on one single imaging modality at a time. However, different imaging modalities glean differ aspects of the brain activity, such as the electrical, hemodynamic and metabolic activity, and offer distinct and often complementary sets of strengths (for example different spatial or temporal resolutions). Therefore the combination of multiple technologies could offer some advantages in the understanding of the multifaceted brain-behavior relationship under pathological conditions following TBI.

Hence, there is an unmet need for multimodal neuroimaging tools amenable for use in natural settings to assist in cognitive rehabilitation after TBI.
1.2 Approach and contribution

The premise of this thesis is to address the need for neuroimaging tools in neurorehabilitation by employing functional near infrared spectroscopy (fNIRS) and its integration with electroencephalography (EEG) for the investigation of TBI-related cognitive impairments.

fNIRS provides a measure of the prefrontal hemodynamic activation, which has been demonstrated to be susceptible to TBI (Langfitt et al., 1986; Levin et al., 1982), and it does so in a noninvasive, affordable and wearable way, thus partially overcoming the limitations of other modalities. Similar to its fMRI counterpart, fNIRS is a non-invasive method for studying functional activation through monitoring changes in the hemodynamic properties of the brain. However, unlike the commonly used BOLD (blood oxygen level dependent) based fMRI techniques, which derive contrast from the paramagnetic properties of deoxyhemoglobin, fNIRS is based on the intrinsic optical absorption of blood. As a result, fNIRS has the ability to simultaneously record not only concentration changes in deoxyhemoglobin but in oxy-hemoglobin and total hemoglobin as well. In addition, unlike the majority of commonly used fMRI techniques, which typically have an intrinsically and instrumentation limited acquisition rate, the temporal resolution of hemoglobin detection with fNIRS is not acquisition limited and can be up to hundreds of hertz, much faster than the hemodynamic response itself. In this respect, fNIRS potentially provides a more complete temporal picture of brain hemodynamics compared with fMRI (although offering lower spatial resolution and being limited by depth of light penetration in adult humans). EEG offers a cost-effective and simple measure of brain electrical activity and it has been employed in a few studies on traumatic brain injury (Hensel et al., 2004; Keren et al., 1998; Lew et al., 2005). The
integration of these two modalities, EEG and fNIRS, offers a means to partially overcome the limitations of each of the individual modalities by combining their complementary aspects. fNIRS offers reasonably good spatial resolution of the prefrontal hemodynamic response to the local neuronal activity, while EEG is a direct measure of electrophysiological activity and provides information about the processes ongoing in different areas of the brain, with temporal resolution of milliseconds. The proposed ‘composite measure’ is likely to lead to a more robust metric of cognitive activity.

The overall objective of the current study is to investigate the applicability of fNIRS, and its integration with EEG, to assessment of working memory after TBI. Working memory is defined as the storage, monitoring and manipulation of information and regarded as a fundamental mechanism to executive functioning (Baddeley, 1986; Swanson, 1999). Working memory processes involve the dorsolateral prefrontal cortex (DLPFC) (Cabeza and Nyberg, 2000; Owen et al., 2005), which can be monitored by fNIRS and which is vulnerable to TBI (McDonald et al., 2002). Importantly, working memory deficits have also been demonstrated to be highly related to functional outcome following TBI (Brooks et al., 1987; Ciaramelli et al., 2006).

This objective was reached in three sequential steps, categorized under three different aims. For all three steps, we used a set of fNIRS and EEG data simultaneously collected from 11 healthy subjects and 6 traumatic brain injury subjects during a working memory (n-back) task.

1) In the first step (presented in Chapter 2), we first examine the discriminative ability of fNIRS in a comparison of working memory between healthy subjects and subjects with TBI. Understanding the differences in the hemodynamic activation patterns between
the two groups, as recorded by fNIRS, needs to be examined before fNIRS can be employed in more refined studies on the recovery of working memory deficits during neurorehabilitation. We hypothesize that an altered task-related pattern of activation will be present in TBI subjects and that, in particular, TBI subjects will present higher hemodynamic activation values compared to healthy controls.

*Contribution and significance:* This is the first study that investigates the use of fNIRS in the assessment of working memory in TBI subjects in comparison to healthy controls. Possible differences in the hemodynamic activation patterns between the two groups, as recorded by fNIRS, need to be examined before fNIRS can be further investigated in studies that monitor recovery of working memory deficits during neurorehabilitation.

2) In the second step (presented in Chapter 3), these differences are used to classify between the two groups of individuals (TBI and healthy). We expect that fNIRS will be significantly better than chance in separating between TBI subjects and healthy controls.

*Contribution and significance:* While this is a gross classification, it is the first step needed towards a classification of levels of impairments and the development of fNIRS as a clinically useful tool. Additionally, the use of fNIRS to classify between groups of subjects and, in particular, between individuals with and without cognitive impairments, has not been previously reported in the literature. (The only previous attempt at using fNIRS for the classification between groups of individuals is presented in Chapter 5).

3) In the third step (presented in Chapter 4), the classification between TBI and healthy group obtained using fNIRS-related features is compared with the results obtained using a combined fNIRS/EEG approach.

*Contribution and significance:* A few studies have previously explored EEG in neurorehabilitation applications for subjects with brain injuries and in a few cases fNIRS and EEG have been simultaneously recorded in previous studies. However, this is the
first study that has investigated the actual fNIRS-EEG integration for a multimodal approach to the assessment of cognitive impairments. The results presented in this dissertation indicate that, if fNIRS and EEG are appropriately combined, a better metric of the impairment level after TBI and of the functional rehabilitation following interventions can be developed.

1.3 Conclusions and future work

The results of this thesis suggest that fNIRS is able to uncover differences between healthy subjects and subjects with TBI in the hemodynamic patterns in response to working memory load. This study demonstrates also that fNIRS measures can distinguish between the two groups and these data are investigated relative to the results obtained by EEG alone or by the ‘composite measure’ based on the combination of these two technologies: each of the two modalities reveals unique strengths that can contribute to the classification between the two groups.

Building on the results presented in this thesis, Chapter 5 offers future research directions and presents two exploratory studies, which deserve further examination. The first one focuses on the advantage that fNIRS measures offer in the development of therapeutic interventions (Merzagora et al., 2010) based on transcranial direct current stimulation (tDCS). The second one focuses on the cognitive assessment aspect, presenting a preliminary proof-of-concept of the applicability of fNIRS and EEG in the
assessment of cognitive performance in healthy subjects (Merzagora et al., 2009b). If further investigated, these exploratory studies can lead to a broader deployment of fNIRS and its combination with EEG in the cognitive neurorehabilitation field.
CHAPTER 2: DISCRIMINATIVE VALIDATION OF FNIRS

2.1 Introduction

Several cognitive domains have been frequently reported to be impaired in traumatic brain injury individuals (TBI). These domains include short-term and working memory, speed of information processing, attention and executive functions in general (Binder, 1986; Leclercq et al., 2000; Levin and Grossman, 1978; McDowell et al., 1997; McMillan, 1997; Proctor et al., 2000; Ziino and Ponsford, 2006).

Working memory impairments are regarded as the most significant and most disabling deficits that are endured following TBI (Mateer, 1999; Millis et al., 2001). The term ‘working memory’ defines the online storage, monitoring and manipulation of information (Baddeley and Hitch, 1974). Working memory is considered to play a critical role in higher-level cognitive processes, such as problem solving, reasoning, planning, language understanding and guidance of goal-oriented behaviors. Deficits of working memory (and memory in general) have been found to be slower to recover than other cognitive functions (Lezak, 1979), with residual impairments even several years after an injury (Millis et al., 2001; Zec et al., 2001). Furthermore, reduced working memory abilities are highly related to an individual’s functional outcome following TBI (Brooks et al., 1987; Ciaramelli et al., 2006). As a consequence, a vast number of studies have focused on working memory impairments in individuals who have sustained TBI.

Studies of working memory processes in healthy individuals, performed using functional magnetic resonance (fMRI) and positron emission tomography (PET), have
uncovered the involvement of regions in the frontal and prefrontal cortex, in the temporal cortex and in the parietal cortex (Cabeza and Nyberg, 2000; Cohen et al., 1997; Owen et al., 2005). Specifically in n-back tasks, a common measure of working memory performance, activation has been found in Broadmann areas 9 and 46, roughly corresponding to the dorsolateral prefrontal cortex (DLPFC). This suggests the involvement of these areas in tasks that require online storage, monitoring and the manipulation of information.

A common mechanism of TBI is the “coup-contrecoup”; because of this, the DLPFC is regarded as highly vulnerable in TBI (Langfitt et al., 1986; Levin et al., 1982) and several functional neuroimaging studies of working memory in the TBI population have identified modified task-related patterns of activation in the DLPFC following brain injury. As an example, Christodoulou and colleagues (Christodoulou et al., 2001) used fMRI to investigate the performance of severe TBI patients in Paced Auditory Serial Addition Test and found a more dispersed activation in TBI subjects compared to the control group. Another group (McAllister et al., 1999; McAllister et al., 2001) tested the working memory processes in TBI patients using an auditory n-back test. Although all the structural MRI scans were normal and the performance of TBI patients in the n-back test was similar to that of healthy controls, the activation changes in response to increasing working memory load differed between the two groups. When the working memory load was moderate, the TBI group showed a greater extent of activation increase in the bilateral frontal and parietal regions; in agreement with the findings of Scheibel and colleagues (Scheibel et al., 2007). Furthermore, in contrast to the control group, the TBI subjects seemed to fail in the activation of additional areas for increased working memory loads. Studies examining working memory impairments following sports-related concussions can also be added to the body of evidence supporting the notion of a
dysfunctional DLPFC activation. In a study by Lovell and colleagues (Lovell et al., 2007), a cohort of 28 individuals with concussion underwent two sets of fMRI scanning while performing a working memory task: the first scan was performed within approximately one week of the injury and the second after clinical recovery. The results of this study suggest that individuals with task-related DLPFC hyperactivation within the first week of injury demonstrated a more prolonged clinical recovery.

Taken together, we know that the DLPFC plays an important role in working memory and that damages to the DLPFC are closely related to impairments of working memory abilities in traumatic brain injury. This relationship has been established based on traditional methods but not with fNIRS. To address this, the first aim of this study was to examine the ability of fNIRS to identify hemodynamic differences in the working memory-related activity of the DLPFC in the TBI population. This would in fact be the initial step needed to pave the way for further studies of fNIRS in a range of neurorehabilitation applications.

### 2.2 Methods

#### 2.2.1 Subjects

A total of 17 adults participated in the study: 11 healthy subjects (age mean±standard deviation: 31±13; 95% confidence interval [22, 40]) and 6 subjects with traumatic brain injury (age mean±standard deviation: 42±10; 95% confidence interval [31
In order to ensure that no age difference existed between the control group and the TBI group, an independent samples t-test was performed with a 5% level of significance. All participants were right-handed, with vision correctible to 20/20, denied any history of neurological disorders, psychiatric illness or substance abuse.

In order to be included in the study, participants with traumatic brain injury needed to present an injury diagnosis as defined by the NIDRR Traumatic Brain Injury Model Systems National Database (Harrison-Felix et al., 1996): “damage to brain tissue caused by an external mechanical force, as evidenced by loss of consciousness due to brain trauma, post-traumatic amnesia, skull fracture, or objective neurological findings that can reasonably be attributed to TBI on physical examination or mental status examination.” Verification and severity of TBI were documented using the Glasgow Coma Scale (GCS) scores obtained from the individual's medical record, but the GCS was not an exclusionary standard: if the GCS score was not available, loss of consciousness and/or post-traumatic amnesia was used to define the severity of TBI. Subjects with multiple acquired brain injuries were excluded from the study.

The study was approved by Drexel University Institutional Review Board and all participants gave their written informed consent after a detailed explanation of the procedure.

2.2.2 Neuropsychological evaluation

A cognitive assessment was administered to establish the cognitive characteristics of the study participants, specifically targeting attention and working memory. In particular, a battery of standardized clinical behavioral measures was used to evaluate performance in a range of cognitive domains.
Working memory:

- **Paced Auditory Serial Addition Test (PASAT):** Complex information processing was assessed by means of the Levin adaptation of the PASAT (Brittain et al., 1991). This task requires the patient to add auditorily-presented randomized single digits so that each digit is added to the one immediately preceding it. Fifty digits were presented via a cassette tape with a presentation speed of 2.4 s between digits. Performance was evaluated by calculating the number of correct responses.

- **WAIS-III Letter-Number Sequencing:** This test assesses working memory and attention. It requires the subject to order a series of numbers and letters, presented orally, according to a pre-specified order (i.e. numbers in ascending order, followed by letters in alphabetical order). Each item is comprised of three trials and with each additional item the length of the presented sequence is increased (Wechsler, 1997).

Attention:

- **WAIS-III Digit Span:** This test uses two segments, forward and backward. Each segment consists of seven pairs of random single digit sequences that the examiner reads aloud at a rate of approximately one number per second. Both segments depend upon auditory attention and working memory to perform effectively. In the Digit Span forward segment, the subject is instructed to repeat the string of digits in the same order in which they are presented by the examiner. Conversely, in the Digit Span backward segment, the subject is instructed to repeat the string of digits in the reverse order (Wechsler, 1997).

- **Conner’s Continuous Performance Test (CPT II):** This test measures primarily attention. The task presents a series of letters to the subject, who is instructed to press a button whenever any letter except the target letter “X” appears on the computer screen (Conners and Jett, 1999).
Executive functions / Attention:

- **Stroop Color-Word Test**: This task is based upon the finding that it takes longer to name the color of color patches than to read words and even longer to read printed color names when the printed ink is in a different color than the name of the color word (Stroop, 1935). This latter observation has been attributed to a number of possible factors, including the presence of a response conflict, failure of response inhibition and poor selective attention. The task involves three trials. In the first trial, the subject is asked to read out loud and as quickly as possible a list of color names printed in black ink. The second trial involves naming the color of blocks of Xs (i.e. XXXXX) printed in one of three ink colors (blue, red, green). The final trial involves the naming of colored ink; on this trial, however, the colored ink spells out the name of another color. Therefore, the last trial requires the subject to inhibit the automatic response of reading the color name and offer the desired response: the color of the ink. The Stroop test has satisfactory reliability and age effects have been limited to the final color-word interference trial (Golden and Freshwater, 2002).

The results of the neuropsychological evaluation of the two groups were compared using Mann-Whitney tests with a 5% level of significance (based on a Kolmogorov-Smirnov test, the neuropsychological data were found to be not normally distributed, so a nonparametric statistical test was used).

2.2.3 **Experimental paradigm**

Participants were seated in a dimly lit, sound attenuated room and were asked to perform a verbal *n-back* task. N-back has been widely used to investigate working
memory processes and a variety of versions have been employed, as for example auditory, verbal and spatial n-back tasks (Owen et al., 2005). For this study a verbal n-back was chosen to allow a direct comparison with previous fMRI studies that have investigated working memory in TBI individuals (McAllister et al., 1999; McAllister et al., 2001). Stimuli were single consonant presented in a pseudo-random sequence on a computer screen. Four load conditions were used to incrementally vary the working memory load from zero to three items. In the 0-back condition, subjects responded by pressing a button (with their dominant hand) any time the letter on the screen matched a single pre-specified letter (e.g. “X”). In the 1-back condition, the match was defined as any letter identical to the one immediately preceding it (i.e., one trial back). In the 2-back and 3-back conditions, the matches were defined as any letter that was identical to the one presented two or three trials back, respectively. This strategy incrementally increased working memory load from the 0-back to the 3-back condition: while the requirements of the stimulus encoding and response are constant across conditions, the demands to maintain and update increasingly greater amounts of information at higher loads differ.

The n-back task was built on the STIM platform (Neuroscan Inc., El Paso, TX). In all of the four n-back conditions, the pseudo-random sequences of consonants were centrally presented on a visual display; the probability of match and non-match occurrence was 33% and 76% respectively. Stimulus duration was 0.5 s and stimulus onset asynchrony was 3 s. Seven repetitions, each containing the four load conditions (0-, 1-, 2- and 3-back), were presented to the subjects and the order of task conditions was randomized within and across subjects: this approach was taken, based on previous studies (Izzetoglu et al., 2007; McAllister et al., 1999; McAllister et al., 2001), in order to avoid inducing in the task-related hemodynamic activity the existence of systematic influences of one load condition to another. Subjects were given visual instructions regarding the task condition.
to be performed at the start of each trial repetition (5 s duration). Each presentation of the load conditions was comprised of 20 stimuli and was followed by a 10 s rest period (Figure 2.1). In each presentation, a number of stimuli were non-matches repeats that were included as foils (e.g. 1-back repeats in a 2-back task). Prior to performing the task, subjects were pre-practiced to ensure that they understood the task instructions and were capable of performing the task.

Figure 2.1 N-back protocol.
In the visual n-back task four conditions were used to incrementally vary the working memory load for zero items (0-back condition) to three items (3-back condition). Each of the four conditions was repeated seven times, in a random order.

The response times and the percentage of correctly categorized stimuli (% Correct) were tested for significant differences between control subjects and TBI subjects using Mann-Whitney tests, both at the single-load level and globally with a 5% level of significance (based on a Kolmogorov-Smirnov test, the neuropsychological data were found to be not normally distributed, so a nonparametric statistical test was used).
Additionally, we examined the correlation (Kendall’s $\tau_b$) between the % Correct in 2-back and 3-back (the task conditions expected to be demanding of working memory) and the results of the PASAT and WAIS-III Letter-Number Sequencing tests (clinically established measures for working memory).

2.2.4 fNIRS data acquisition

The hemodynamic activity of the prefrontal cortex was recorded using a continuous-wave fNIRS device first described by Chance et al. (Chance et al., 1998) and further developed at Drexel University (Philadelphia, PA). The system consisted of three modules: a flexible headpiece, a control box for hardware management and a computer that runs the data acquisition. The headpiece holds four light sources and 10 photodetectors, with a source-detector separation of 2.5 cm, providing a penetration depth of approximately 1.25 cm. The four light sources were activated in turns: each source shone light with input intensity $I_0$ and the four photodetectors surrounding the currently active source measured the intensity $I$ of the emerging light. The arrangement of sources and detectors on the headpiece and the configuration for data acquisition yields a total of 16 active optodes, which were designed to image cortical areas that correspond to the dorsal and inferior frontal cortices (Izzetoglu et al., 2005). Each source emitted light at two different wavelengths in the near-infrared spectrum, namely at 730 and 850 nm, and measures of emerging light intensity were obtained for each optode with a sampling frequency of 2 samples/second.
2.2.5 fNIRS data processing and analysis

Changes in light absorption, as measured by fNIRS at each of the two wavelengths, were converted to changes in concentration of oxyhemoglobin (HbO$_2$) and deoxyhemoglobin (HHb).

It is generally assumed (Delpy et al., 1988) that the attenuation that the shined light undergoes when travelling through the tissue reflects a linear superimposition of two processes, absorption and scattering, and can be represented as:

$$OD_\lambda = \log_{10} \frac{I_{0,\lambda}}{I_\lambda} = A_\lambda + S_\lambda \quad \text{(Eq. 2.1)}$$

$OD_\lambda$ represents the light attenuation at the wavelength $\lambda$ expressed in optical density (OD) units; $I_{0,\lambda}$ is the intensity of the input light at the wavelength $\lambda$; $I_\lambda$ is the intensity of the light at the wavelength $\lambda$ measured by the detector; $A_\lambda$ and $S_\lambda$ represent the light attenuation caused respectively by absorption and scattering at the wavelength $\lambda$.

In the near-infrared region (between 700 and 900 nm), HbO$_2$ and HHb are the two main chromophores; when taking into account only their contribution to light absorption, the term $A_\lambda$ in (Eq.1) can be written as

$$A_\lambda = (\epsilon_{\text{HbO$_2$},\lambda} \cdot C_{\text{HbO$_2$},\lambda} + \epsilon_{\text{HHb},\lambda} \cdot C_{\text{HHb},\lambda}) \cdot r_{sd} \cdot DPF_\lambda \quad \text{(Eq. 2.2)}$$

based on the modified Beer-Lambert law (mBLL) (Delpy et al., 1988). In (Eq.2), $\epsilon_{\text{HbO$_2$}}$ and $\epsilon_{\text{HHb}}$ are the specific absorption coefficients of, respectively, HbO$_2$ and HHb at the wavelength $\lambda$; $C_{\text{HbO$_2$}}$ and $C_{\text{HHb}}$ are the concentrations of HbO$_2$ and HHb in the sampled volume of tissue; $r_{sd}$ is the physical source-detector separation; and $DPF_\lambda$ is the differential pathlength factor at the wavelength $\lambda$. The $DPF_\lambda$ corrects the $r_{sd}$ to give a better estimate of the real length of the path traveled by photons as a consequence of scattering and absorption. The values for $\epsilon_{\text{HbO$_2$}}$, $\epsilon_{\text{HHb}}$ and $r_{sd}$ are considered time-independent and spatially constant for the adult forehead (Delpy et al., 1988).
$S_i$ is generally considered a constant factor, dependent on the geometry (Obrig and Villringer, 2003). This allows the differential representation of the mBLL:

$$\Delta OD_i(t) = OD_i(t) - OD_{i,\text{control}} = \log_{10} \frac{I_{i,\text{control}}}{I_i(t)} =$$

$$= (A_i(t) + S_i) - (A_{i,\text{control}} + S_i) = A_i(t) - A_{i,\text{control}} =$$

$$= (\epsilon_{HbO2,i} \cdot \Delta C_{HbO2}(t) + \epsilon_{HHb,i} \cdot \Delta C_{HHb}(t)) \cdot r_{id} \cdot DPF_i$$  \hspace{1cm} (Eq. 2.3)

with $\Delta C_{HbO2}(t)$ and $\Delta C_{HHb}(t)$ representing the differences in the HbO2 and HHb concentrations between the instant in time $t$ and the control condition.

In continuous-wave fNIRS, the differential mBLL is used and light attenuation is measured at two (or more) wavelengths in order to calculate $\Delta C_{HbO2}(t)$ and $\Delta C_{HHb}(t)$. The continuous-wave fNIRS system used in this study measured light attenuation at two wavelengths (namely, 730 and 850 nm), therefore $\Delta C_{HbO2}(t)$ and $\Delta C_{HHb}(t)$ were computed through the solution of a system of two equations:

$$\begin{cases}
\Delta OD_{730\text{nm}}(t) = (\epsilon_{HbO2,730\text{nm}} \cdot \Delta C_{HbO2}(t) + \epsilon_{HHb,730\text{nm}} \cdot \Delta C_{HHb}(t)) \cdot r_{id} \cdot DPF_{730\text{nm}} \\
\Delta OD_{850\text{nm}}(t) = (\epsilon_{HbO2,850\text{nm}} \cdot \Delta C_{HbO2}(t) + \epsilon_{HHb,850\text{nm}} \cdot \Delta C_{HHb}(t)) \cdot r_{id} \cdot DPF_{850\text{nm}}
\end{cases}$$  \hspace{1cm} (Eq. 2.4)

thus leading to:

$$\Delta C_{HbO2}(t) = \frac{\epsilon_{HbO2,730\text{nm}} \cdot \Delta OD_{730\text{nm}}(t)}{r_{id} \cdot DPF_{730\text{nm}}} - \frac{\epsilon_{HbO2,850\text{nm}} \cdot \Delta OD_{850\text{nm}}(t)}{r_{id} \cdot DPF_{850\text{nm}}}$$

$$\Delta C_{HHb}(t) = \frac{\epsilon_{HHb,730\text{nm}} \cdot \Delta OD_{730\text{nm}}(t)}{r_{id} \cdot DPF_{730\text{nm}}} - \frac{\epsilon_{HHb,850\text{nm}} \cdot \Delta OD_{850\text{nm}}(t)}{r_{id} \cdot DPF_{850\text{nm}}}$$  \hspace{1cm} (Eq. 2.5)

Overall, the two hemodynamic variables, $\Delta C_{HbO2}(t)$ and $\Delta C_{HHb}(t)$, were measured for each of the 16 optodes. A third hemodynamic variable, $\Delta C_{HbTOT}(t)$, was derived as the sum of $\Delta C_{HbO2}(t)$ and $\Delta C_{HHb}(t)$ and is considered as a good indicator of variations in the regional cerebral blood volume (Tuchin, 2002).

From the overall data, epochs locked to the beginning of a single load condition presentation were extracted. Each epoch lasted 70 s and a 5 s rest baseline was included.
The baseline was used as a control in the mBLL for the calculation of the hemodynamic variables. A total of 28 epochs were extract, 7 for each of the load conditions.

Features describing the hemodynamic response were extracted from single epochs. For each of the three traditional hemodynamic variables (HbO\textsubscript{2}, HHb and HbTOT) the maximum value was extracted: \( H_{\text{max}} \text{HbO}_2 \) for oxyhemoglobin, \( H_{\text{max}} \text{HHb} \) for deoxyhemoglobin and \( H_{\text{max}} \text{HbTOT} \) for total hemoglobin. Additionally, we investigated the maximum ratio (\( MR \)) between HHb and HbO\textsubscript{2}. This metabolic ratio can in fact provide information about the usage of ‘metabolic resources’, since HHb is a byproduct of the consumption of the oxygen delivered by HbO\textsubscript{2}.

In order to reduce the data dimensionality, signals recorded from the optodes approximately overlying the left (optodes 3,4,5,6) and right (optodes 11,12,13,14) dorsolateral prefrontal cortex (DPLFC) were averaged together before feature extraction.

The regions of interest for the n-back task in the prefrontal area were in fact identified based on previous neuroimaging studies (Cabeza and Nyberg, 2000; Owen et al., 2005).

The following statistical analyses were conducted to compare control and TBI group at a 5% level of significance:

1. **Comparison of global hemodynamic activity**: a Mann-Whitney test was performed on each of the hemodynamic features (\( H_{\text{max}} \text{HbO}_2, H_{\text{max}} \text{HHb}, H_{\text{max}} \text{HbTOT} \)) and the metabolic ratio \( MR \), averaged across regions of interest and working memory load;

2. **Hemispheric distribution of hemodynamic activity**: two Mann-Whitney tests were performed on each of the hemodynamic features and the metabolic ratio, averaged across working memory load but separately for each region of interest (left and right DLPFC);

3. **Load effect on hemodynamic activity**: for the region of interest(s) that were found significantly different between control and TBI group, additional Mann-Whitney tests
were performed on each of the hemodynamic features and on the metabolic ratio, separately for the single working memory loads; 

4. Correlation between hemodynamic activity and n-back performance: we computed the correlation (Kendall’s \( \tau_b \)) between each of the hemodynamic features \( (H_{max_{HbO2}}, H_{max_{HHb}}, H_{max_{HbTOT}}) \) and the metabolic ratio \( MR \) and the behavioral performance in the n-back task (% Correct and response time) at the single trial level. 

For the Mann-Whitney tests the effect size was measured using the Cohen’s \( d \).

2.3 Results

2.3.1 Subjects demographic and neuropsychological assessment

Demographic characteristics of the study participants are provided in Table 2.1. Gender distribution did not differ between the TBI and healthy control (HC) group, as tested with a two-samples Kolmogorov-Smirnov test (Kolmogorov-Smirnov \( Z(15)=0.4187, p=0.995 \)). The two groups did not differ significantly with respect to age, as tested by an independent samples t-test (\( t(15)=-1.742, p=0.102 \)). Based on independent-samples Mann-Whitney U tests, the control and TBI groups differed significantly in the performance of the WAIS-III Digit Span test (\( z(15)=-3.358, p=0.001 \)), in the WAIS-III Letter-Number Sequencing test (\( z(15)=-2.145, p=0.032 \)) and in PASAT 2.4” (\( z(15)=-2.566, p=0.010 \)). This result was expected since it indicates deficits in working memory based on traditional clinical measures.
Table 2.1 Demographics and neuropsychological assessment of participants.
The study participants comprised 11 control subjects and 6 TBI subjects, with no
difference in terms of age at the group level. The participants were tested with a battery
of standard neuropsychological tests to evaluate their performance in a range of cognitive
domains. The table reports the mean, the standard deviation (SD) and the 95% confidence
interval of the mean (LL: lower level; UL: upper level).

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>TBI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean ± SD</td>
<td>95% C.I.</td>
</tr>
<tr>
<td></td>
<td>[LL, UL]</td>
<td></td>
</tr>
<tr>
<td>Number of subjects</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Age</td>
<td>31 ± 13</td>
<td>[22, 40]</td>
</tr>
<tr>
<td>Years post-injury</td>
<td>18 ± 9</td>
<td>[8, 23]</td>
</tr>
<tr>
<td>WAIS-III Digit Span (scaled score)</td>
<td>11 ± 2</td>
<td>[10, 12]</td>
</tr>
<tr>
<td>WAIS-III Letter-Number Sequencing (scaled score)</td>
<td>9 ± 2</td>
<td>[9, 11]</td>
</tr>
<tr>
<td>PASAT 2.4” – t-score</td>
<td>46 ± 6</td>
<td>[42, 49]</td>
</tr>
<tr>
<td>Stroop – Interference t-score</td>
<td>56 ± 12</td>
<td>[49, 64]</td>
</tr>
<tr>
<td>CPT-II Omission – t-score</td>
<td>57 ± 33</td>
<td>[35, 79]</td>
</tr>
<tr>
<td>CPT-II Commission – t-score</td>
<td>50 ± 10</td>
<td>[44, 57]</td>
</tr>
<tr>
<td>CPT-II Hit RT (ms)</td>
<td>367 ± 76</td>
<td>[317, 418]</td>
</tr>
</tbody>
</table>

2.3.2 N-back behavioral performance

As expected, increasing the working memory load led to a decrease in
performance and this decrease was more considerable in TBIs than in HCs (Table 2.2).
The TBI and control groups did not differ significantly in terms of % Correct (defined as percentage of stimuli that received a correct response from the subjects) or response times; this result was true both globally and at the single-load level.
The % Correct from 2-back and 3-back were significantly correlated with the results of
PASAT 2.4” (2-back: $\tau_b = 0.784$, p<0.001; 3-back: $\tau_b = 0.728$, p<0.001) and with the results of WAIS-III Letter-Number Sequencing (2-back: $\tau_b = 0.627$, p=0.007; 3-back: $\tau_b = 0.528$, p=0.029), thus confirming that also for this specific cohort of subjects n-back is a
good test for working memory.
Table 2.2 Behavioral performance in n-back.

The behavioral performance of the subjects in the n-back task was evaluated in terms of % Correct, defined as the percentage of stimuli that received a correct response, and in terms of response time. Behavioral performance was assessed both globally and at the single-load level. The control group and the TBI group did not differ in terms of behavioral performance. The table reports the mean, the standard deviation (SD) and the median on the behavioral performances.

<table>
<thead>
<tr>
<th></th>
<th>HC</th>
<th>TBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Correct – 0-back</td>
<td>98.6 ± 2.7</td>
<td>88.5 ± 13.2</td>
</tr>
<tr>
<td>% Correct – 1-back</td>
<td>99.2 ± 1.0</td>
<td>88.9 ± 13.1</td>
</tr>
<tr>
<td>% Correct – 2-back</td>
<td>97.8 ± 1.8</td>
<td>86.0 ± 10.7</td>
</tr>
<tr>
<td>% Correct – 3-back</td>
<td>97.1 ± 2.1</td>
<td>83.1 ± 14.5</td>
</tr>
<tr>
<td>% Correct – OVERALL</td>
<td>98.2 ± 0.9</td>
<td>86.6 ± 12.4</td>
</tr>
</tbody>
</table>

| Response time (ms) – 0-back | 504 ± 166 | 514 ± 136 |
| Response time (ms) – 1-back | 547 ± 163 | 497 ± 219 |
| Response time (ms) – 2-back | 596 ± 184 | 607 ± 419 |
| Response time (ms) – 3-back | 578 ± 201 | 653 ± 492 |
| Response time (ms) – OVERALL | 556 ± 164 | 568 ± 269 |

2.3.3 Neuroimaging data

Comparison of global hemodynamic activity

The maximum values of the hemodynamic variables ($H_{max}$) and the metabolic ratio $MR$ (maximum ratio between HHb and HbO$_2$) were compared between the TBI and healthy group using Mann-Whitney U tests. The comparison revealed that the maximum of the hemodynamic response reached by the TBI group, averaging across working memory load or region of interest, is significantly higher for each of the three hemodynamic variables ($H_{max_{HbO2}}$, $H_{max_{HHb}}$, $H_{max_{HbTOT}}$) (Table 2.3 and Figure 2.2A) but not for the metabolic ratio $MR$. 
Table 2.3 Between-group comparison, averaging across load and region.
When analyzing the hemodynamic response after averaging across working memory load and region of interest, the two groups presented statistically significant differences, with the TBI group reaching greater values in all three hemodynamic variables. The table reports the z score, the p value, the effect size d and the 95% confidence interval of the effect size (LL: lower level; UL: upper level).

(1) $df=15$, (2) Cohen’s $d=(m_{HC}-m_{TBI})/s.d_{HC}$ where a positive value indicates a larger mean of the control group compared to the TBI group.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Z</th>
<th>P</th>
<th>D</th>
<th>95% C.I. of D</th>
</tr>
</thead>
<tbody>
<tr>
<td>HmaxHbO2</td>
<td>2.348</td>
<td>0.015 *</td>
<td>-0.37</td>
<td>[-0.52 -0.21]</td>
</tr>
<tr>
<td>HmaxHHb</td>
<td>2.293</td>
<td>0.022 *</td>
<td>-0.48</td>
<td>[-0.63 -0.32]</td>
</tr>
<tr>
<td>HmaxTOT</td>
<td>3.022</td>
<td>0.003 *</td>
<td>-0.46</td>
<td>[-0.62 -0.31]</td>
</tr>
<tr>
<td>MR</td>
<td>-0.009</td>
<td>0.993</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4 Between-group comparison, separately for region and averaging across load.
When analyzing the hemodynamic response after averaging across working memory load but separately for region of interest, the two groups presented statistically significant differences in the left DLPFC, with the TBI group reaching greater values in all three hemodynamic variables. The table reports the z score, the p value, the effect size d and the 95% confidence interval of the effect size (LL: lower level; UL: upper level).

(1) $df=15$, (2) Cohen’s $d=(m_{HC}-m_{TBI})/s.d_{HC}$ where a positive value indicates a larger mean of the control group compared to the TBI group.

<table>
<thead>
<tr>
<th>Region</th>
<th>Variable</th>
<th>Z</th>
<th>P</th>
<th>D</th>
<th>95% C.I. of D</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEFT</td>
<td>HmaxHbO2</td>
<td>2.032</td>
<td>0.042 *</td>
<td>-0.35</td>
<td>[-0.52 -0.13]</td>
</tr>
<tr>
<td></td>
<td>HmaxHHb</td>
<td>2.836</td>
<td>0.005 *</td>
<td>-1.10</td>
<td>[-1.34 -0.87]</td>
</tr>
<tr>
<td></td>
<td>HmaxTOT</td>
<td>3.177</td>
<td>0.001 *</td>
<td>-0.50</td>
<td>[-0.72 -0.28]</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>0.723</td>
<td>0.470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIGHT</td>
<td>HmaxHbO2</td>
<td>1.430</td>
<td>0.153</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HmaxHHb</td>
<td>0.470</td>
<td>0.638</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HmaxTOT</td>
<td>1.113</td>
<td>0.266</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>-0.751</td>
<td>0.453</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hemispheric distribution of hemodynamic activity

Additional analyses were performed separately for the two regions of interest (left and right DLPFC), using Mann-Whitney U tests. They revealed that no difference exists between the two groups in the right DLPFC but that the activation for the TBI group is significantly higher than the activation of the control group at the left DLPFC in
terms of $H_{max_{HbO2}}$, $H_{max_{HHb}}$, $H_{max_{HbTOT}}$ but not in terms of metabolic ratio $MR$ (Table 2.4 and Figure 2.2B).

**Load effect on hemodynamic activity**

As confirmed by Mann-Whitney U tests performed at the single working memory load level, the hemodynamic response of the left DLPFC of the TBI group is significantly greater than the control group during the 0-back condition ($H_{max_{HbO2}}$: $z(15)=2.192$, $p=0.028$, E.S. $d=-0.61$, 95% C.I. of $d [-1.05 -0.17]$; $H_{max_{HbTOT}}$: $z(15)=2.013$, $p=0.044$, $d=-0.70$, E.S. 95% C.I. of $d [-1.15 -0.26]$). At the 2-back level, instead, the maximum value for HHb was higher in TBI patients than in control subjects ($H_{max_{HHb}}$: $z(15)=2.276$, $p=0.023$, E.S. $d=-1.28$, 95% C.I. of $d [-1.75 -0.82]$) (Figure 2.2C). No difference was found in the analysis of the metabolic ratio $MR$.

**Correlation between hemodynamic activity and n-back performance**

The relation between the behavioral performance in the n-back task (% Correct and response time) and the hemodynamic response ($H_{max_{HbO2}}$, $H_{max_{HHb}}$, $H_{max_{HbTOT}}$ and $MR$) was investigated by means of non-parametric correlations (Kendall’s $\tau_b$) performed separately for the healthy and the TBI group. Two separate relationships were found. On one side, the control group showed a significant positive correlation between the response time and $MR$ in both the left and right side (left DLPFC: $\tau_b=0.159$, $p<0.001$; right DLPFC: $\tau_b =0.152$, $p=0.001$): healthy subjects that showed slower response times also exhibited higher consumption of ‘metabolic resources’. On the other side, the TBI group showed a significant negative correlation between $H_{max_{HHb}}$ and % Correct in the left side ($\tau_b =-0.201$, $p=0.003$): TBI subjects that made more errors demonstrated higher values for HHb, a byproduct of ‘metabolic resources’ consumption.
Figure 2.2 Hemodynamic response compared between controls and TBI subjects. Boxplot of $H_{max}$ comparing the two groups (control vs. TBI) for each of the three hemodynamic variables. (A) When averaging across working memory loads and across region of interest, the TBI group had significantly higher hemodynamic values for each of the three variables. (B) When averaging across working memory loads and comparing the two regions of interest (left and right DLPFC), the TBI group had significantly higher hemodynamic values than healthy controls in the left DLPFC. (C) When analyzing the hemodynamic response in the left DLPFC at the single-load level, differences were found at 0-back for HbO$_2$ and HbTOT and at 2-back for HHb. The star represents the existence of a statistically significant difference.
Correlation between hemodynamic activity and n-back performance

The relation between the behavioral performance in the n-back task (% Correct and response time) and the hemodynamic response ($H_{max_{HbO2}}$, $H_{max_{HHb}}$, $H_{max_{HbTOT}}$ and $MR$) was investigated by means of non-parametric correlations (Kendall’s $\tau_b$) performed separately for the healthy and the TBI group. Two separate relationships were found. On one side, the control group showed a significant positive correlation between the response time and $MR$ in both the left and right side (left DLPFC: $\tau_b=0.159$, $p<0.001$; right DLPFC: $\tau_b =0.152$, $p=0.001$): healthy subjects that showed slower response times also exhibited higher consumption of ‘metabolic resources’. On the other side, the TBI group showed a significant negative correlation between $H_{max_{HHb}}$ and % Correct in the left side ($\tau_b =-0.201$, $p=0.003$): TBI subjects that made more errors demonstrated higher values for HHb, a byproduct of ‘metabolic resources’ consumption.

**Figure 2.3** Relation between response time and metabolic ratio in healthy controls. Comparison between the median of the response time (RT) and the median of the metabolic ratio $MR$ for the two regions of interest (left and right DLPFC) at varying task loads, in healthy controls. In both left and right DLPFC there was a significant positive relation between RT and $MR$. (Whiskers represent the M.A.D., median absolute deviation.)
Figure 2.4 Relation between % Correct and left DLPFC $H_{\text{max,HHB}}$ in TBI subjects. Comparison between the median of the percentage of correct responses (% Correct) and the median of the deoxyhemoglobin maximum change ($H_{\text{max,HHb}}$) for the left DLPFC at varying task loads, in TBI subjects. There was a significant negative relation between % Correct and $H_{\text{max,HHb}}$. (Whiskers represent the M.A.D., median absolute deviation.)

2.4 Discussion

Taken altogether, the results of this study suggest that fNIRS may be able to provide information about the differences in the DLPFC activation patterns between individuals with TBI and healthy controls in response to a working memory task.

Behavioral results

The neuropsychological tests confirmed impairments of the working memory domain for the subjects in the TBI group; this evaluation was based on significantly lower scores in the WAIS-III Digit Span, WAIS-III Letter-Number Sequencing and PASAT tests, which are clinical standard measures. This is in line with the literature on neuropsychological outcomes in TBI (Bublak et al., 2000; McDowell et al., 1997).
The behavioral performance in the n-back task surprisingly did not differ significantly between the two groups (although the percentage of stimuli that received correct response in the single load conditions and overall was in general lower for the TBI subjects and this difference was more marked for 3-back). However, other studies have found cohorts of healthy and TBI subjects that did not differ in the n-back performance (Newsome et al., 2007; Perlstein et al., 2004); furthermore, this lack of difference makes that differences in hemodynamic measures meaningful.

The significant correlation between the n-back performance (\% Correct) in 2-back and 3-back and standard neuropsychological tests of working memory (WAIS-III Letter-Number Sequencing and PASAT) additionally confirmed that the findings from the n-back test did investigate working memory in this groups of subjects.

**Comparison of global hemodynamic activity**

Overall, the maximum hemodynamic response reached by the TBI group was significantly higher than that reached by the healthy control group, when averaging across working memory load or region of interest, as shown in Figure 2.2A and Table 2.3; the Cohen’s \(d\) measured a moderate effect for each of the three hemodynamic variables \((H_{max_{\text{HbO2}}}, H_{max_{\text{HHb}}}, H_{max_{\text{HbTOT}}})\). On one side, higher values of oxyhemoglobin and total hemoglobin can be interpreted as increased availability of local resources; on the other side, higher deoxyhemoglobin values can be interpreted as higher use of the available resources (oxyhemoglobin turns into deoxyhemoglobin after the release of the oxygen it carries). This pattern of activation is consistent with findings from previous neuroimaging studies that have investigated the hemodynamic and metabolic activity in TBI subjects during working memory tests and have found increase
recruitment in the DLPFC (Lovell et al., 2007; Perlstein et al., 2004; Scheibel et al., 2007; Scheibel et al., 2003; Turner and Levine, 2008).

**Hemispheric distribution of hemodynamic activity**

A similar result (higher values for $H_{\text{max}_{\text{HbO}_2}}$, $H_{\text{max}_{\text{Hb}}}$, $H_{\text{max}_{\text{HbT}}}$ in TBI subjects) was found also when comparing the two groups in the single region of interests: TBI subjects reached significantly higher levels of activation than healthy subjects in the left DLPFC; the effect size of this difference was small to moderate for oxyhemoglobin and total hemoglobin and moderate to large for deoxyhemoglobin (Figure 2.2B and Table 2.4). Previous working memory studies of healthy subjects (Cabeza and Nyberg, 2000; Owen et al., 2005) have found that, although activation is typically bilateral, the left DLPFC plays a key role in verbal working memory (in comparison, for example, to spatial working memory) (Smith and Jonides, 1999). Therefore the results of the comparison between the healthy group and the TBI group did not show any difference in the hemispheric distribution of activity and confirmed that left DLPFC is crucial in the recruitment and use of resources. For this reason, the analysis of the working memory load effects on the hemodynamic activation was performed focusing on the left DLPFC.

**Load effect on hemodynamic activity**

The effects of working memory load were investigated in the pattern of hemodynamic activity in the left DLPFC. On one side, the healthy group showed increased DLPFC activation following the increments in working memory load (Figure 2.2C and Table 2.4), in line with the results obtained on healthy subjects in previous fMRI (Braver et al., 1997) and fNIRS studies (Izzetoglu et al., 2007). On the other side, for the lowest level of load (0-back) the TBI group shows the highest value of
oxyhemoglobin ($H_{max_{HbO2}}$) and total hemoglobin ($H_{max_{HbTOT}}$), likely indicating a higher need for resources even with very basic cognitive demands (Figure 2.2C and Table 2.4). As the working memory load is increased to 2-back, more resources are used, as indicated by higher values of deoxyhemoglobin ($H_{max_{Hb}}$). Therefore the recruitment and usage of hemodynamic resources did differ between the two groups. Additionally the results obtained from the analysis of the hemodynamic response to increased working memory load in the TBI group were consistent with previous reports (McAllister et al., 1999; McAllister et al., 2001): whereas the healthy subjects maintained their ability to further increase activation, the TBI group showed less of an increase in hemodynamic response during higher load condition (it needs to be remembered, in fact, that fNIRS-related measures carry information about the differential activation). A suggested interpretation for this result is that TBI individuals need greater cortical resources to perform a given task but they seem unable to recruit additional resources and/or areas needed to cope with the increasing demands (McAllister et al., 1999; McAllister et al., 2001).

Correlation between hemodynamic activity and n-back performance

The hemodynamic features extracted from the fNIRS recordings were also studied in correlation with the behavioral performance, separately for the healthy and TBI group.

In the control group a positive correlation was found between the metabolic ratio $MR$ and the response time: the higher the metabolic ratio, the higher the response time. The metabolic ratio $MR$ (maximum ratio between change in deoxyhemoglobin and change in oxyhemoglobin) can be considered as an indicator of resources use: it indicates how much of the available resources is used (since, as pointed out before, deoxyhemoglobin is
a by-product of the use of the oxygen carried by oxyhemoglobin). In general, the response time is regarded as an index of processing speed, defined as the amount of time needed to process a predetermined amount of information or, alternatively, as the amount of information processed within a predetermined amount of time (Deluca et al., 2004). Higher response times (i.e. lower processing speed) could arise either from individual differences or from increased processing demands from the task (higher working memory loads yielded in fact higher response times). A lower processing speed was correlated therefore with a higher use of resources in the healthy group.

The TBI group, on the other hand, showed a negative correlation between the percentage of correct responses and the maximum change in deoxyhemoglobin ($H_{\text{max,Hb}}$) in the left DLPFC. In other words, for the TBI subjects lower performances in the task (as indicated by lower percentages of correct responses) were related to higher use of resources.

In both the healthy and the TBI group, the use of more hemodynamic resources correlated with lower performance, but the aspect of performance that correlated with this were different between the two groups. In healthy subjects, it was represented by response time, indicating that lower performing healthy subjects might present difficulties in terms of information processing speed. In TBI subjects, the aspect of performance that correlated with higher resources use was the behavioral accuracy in the task (percentage of correct responses), likely indicating impairments in the information manipulation aspect of working memory. Therefore hemodynamic measures are able to provide additional information about the different working memory components that might be impaired following TBI, even when no significant difference exists between the behavioral performance of the two groups. Knowledge about these subtle differences would instead be important for the choice and planning of neurorehabilitation interventions targeting working memory.
Final remarks

Although there is no one explanation for these findings, suggestions can be made for the interpretation of the observed results.

One possible explanation for the higher overall values of oxyhemoglobin and total hemoglobin in the TBI group is a higher regional cerebral blood flow (rCBF). However this alone could not explain also the higher values of deoxyhemoglobin. Deoxyhemoglobin is in fact a by-product of oxygen consumption (the oxygen is extracted from the oxyhemoglobin, transforming it into deoxyhemoglobin), therefore increased rCBF would have a ‘wash-out’ effect, thus reducing the amount of deoxyhemoglobin in the sampled tissue (Buxton et al., 2004).

Another possibility is that the TBI group lacks the proper ability to match sufficient processing resources to processing load. Based on a popular model of working memory (Baddeley, 1986), the allocation of processing resources is performed by a central executive component and the DLPFC has a key role in this function (Smith and Jonides, 1999). Since DLPFC is particularly vulnerable to TBI, impairments can occur to the matching of resources with demand (McDonald et al., 2002).

A further possibility that does not contradict the previous speculation might come the notion of neural efficiency. Previous fMRI studies showed that a reduced processing speed is related to increases in DLPFC activation, suggesting that increases in processing speed are related to increases in the efficiency of neural processes (Rypma and D'Esposito, 1999). These findings are consistent with a more general neural efficiency concept of intelligence that emerged from a seminal PET study by Haier and colleagues (Haier et al., 1988) and that found supporting evidence from other neuroimaging studies [PET studies: (Andreasen et al., 1995; Beauchamp et al., 2003; Mehta et al., 2000); fMRI studies: (Mehta et al., 2000; Prat et al., 2007; Reichle et al., 2000; Rypma et al.,]
Based on this theory, better performing (or brighter) individuals use fewer cerebral resources to cope with task demands. Therefore, the relation found here in the healthy group between the response time and the maximum ratio between $H_{\text{max,HHb}}$ and $H_{\text{max,HbO2}}$ could find an explanation in the framework of the neural efficiency theory: (healthy) subjects with faster processing speed may be more efficient in the oxygen extraction, thus transforming more oxyhemoglobin in deoxyhemoglobin. On the other hand, the TBI group showed working memory impairments, as confirmed by the results of neuropsychological tests, and needed more cerebral resources (as indicated by higher values of oxyhemoglobin, deoxyhemoglobin and total hemoglobin). However, TBI individuals that performed better in the n-back task (as defined by the percentage of correct responses) displayed lower values of deoxyhemoglobin compared to TBI subjects that performed worse. This might suggest that better performing TBI subjects used less resources (or use the available resources more efficiently), as indicated by lower values of deoxyhemoglobin. This pattern of hemodynamic differences between the TBI and the control group suggests therefore that there is a “mismatch” not only in the supply of resources at higher working memory loads but also in their consumption.

The results of this study, however, cannot rule out the possibility that the different activation observed in the TBI group is due to a distinct strategy employed in the performance of the task. This explanation seems however unlikely, given the consistency with results of other independent studies.

An additional note needs to be made about the family-wise type I error and the class of statistical tests in the study. Since numerous comparisons have been performed and 4 independent variables have been considered, a possible strategy to reduce the probability
of type I error could be the use of a Bonferroni correction. However, this correction is recognized to be very conservative. Given the small sample size of the study, this should be considered a proof of concept, therefore less conservative approaches are preferable in order not to discard possible interesting findings. The significance threshold for proof of concept studies is usually set at $\alpha=0.1$ but a lower value ($\alpha=0.05$) has been used here as a way to partially control the type I errors. Additionally, given the proof of concept character of the study, we do not know in advance the variability that can be expected in the data; therefore no normalization has been applied to non-normally distributed data but non-parametric statistical tests have been preferred, because of the more powerful results that they yield.
3.1 Introduction

The hemodynamic response in the dorsolateral prefrontal cortex (DLPFC) for traumatic brain injury (TBI) subjects recorded during a working memory task using functional near-infrared spectroscopy (fNIRS) revealed alterations compared to healthy controls (see Chapter 2). After investigating fNIRS for its ability to identify differences between TBI and healthy individuals, we examined the applicability of such differences to classify between the two groups. While this is overall a gross classification, it represents the first step towards classification at a finer level. It could in fact be further developed, in future studies, to categorize TBI individuals based on their working memory impairments and to follow the changes induced by neurorehabilitation intervention.

To the best of our knowledge, this study is the first to apply fNIRS to the classification between groups of individuals and, in particular, between subjects with and without cognitive impairments. There have been, however, few other studies that have employed fNIRS signals in pattern recognition to discriminate within-subject states in brain-computer interface applications. In one of the first studies, Coyle and colleagues (Coyle et al., 2004) used average oxyhemoglobin to identify motor imagery (fist clenching) in physically able subjects. In a similar study (Sitaram et al., 2007) finger tapping was used, leading to results consistent with those reported by Coyle. Beside motor tasks, pattern recognition has been applied to fNIRS-based brain-computer
interfaces with other binary tasks, such as tasks designed to elicit emotions (Tai and Chau, 2009) or designed to distinguish between mental arithmetic and music imagery (Power et al., 2010) or between frequent and infrequent responses (Butti et al., 2007). In this study we aim at exploring the capability of fNIRS-related features to separate between TBI individuals and healthy controls.

### 3.2 Methods

#### 3.2.1 Subjects and experimental paradigm

The same subjects pool of Chapter 2 was included. It consisted of a total of 17 adults participated in the study: 11 healthy subjects and 6 traumatic brain injury subjects. All participants were right-handed, with vision correctible to 20/20, denied any history of neurological disorders, psychiatric illness or substance abuse. In order to be included in the study, participants with traumatic brain injury needed to present an injury diagnosis as defined by the NIDRR Traumatic Brain Injury Model Systems National Database (Harrison-Felix et al., 1996). Additional details about the inclusion/exclusion criteria and about the demographics of the two groups are presented in Chapter 2.

The subjects performed a visual n-back task with four load conditions (0-back, 1-back, 2-back and 3-back) that incrementally varied the working memory load, as described in Chapter 2.
3.2.2 fNIRS data acquisition and processing

The hemodynamic activity of the prefrontal cortex was recorded using a continuous-wave fNIRS device. Changes in light absorption were converted to changes in concentration of oxyhemoglobin (HbO$_2$) and deoxyhemoglobin (HHb). From the overall data, epochs locked to the beginning of a single load condition presentation were extracted and the maximum ($H_{max}$) in the single epoch was used as feature describing the hemodynamic response for each variable (HbO$_2$, HHb and total hemoglobin Hb$_{TOT}$). Additionally, the maximum ratio between HHb and HbO$_2$ was considered ($MR$). The regions of interest for this task were identified to be the left and right dorsolateral prefrontal cortex based on previous neuroimaging studies (Cabeza and Nyberg, 2000; Owen et al., 2005). fNIRS data acquisition and processing is presented in further detail in Chapter 2.

3.2.3 fNIRS-based classification

A total of 32 fNIRS-related features was used to classify between healthy (HC) and TBI subjects. The features consisted in the maximum value $H_{max}$ for the three hemodynamic variables (HbO$_2$, HHb and Hb$_{TOT}$) and the maximum ratio $MR$ in the four load conditions at the two regions of interest. Features selection was performed using a random subspace method (RSM) (Ho, 1998): based on the RSM, the decision about the label to be assigned to a given instance is taken by an ensemble of classifiers that span the decision space. Training/testing was achieved using a modified leave-one-out method (mLOO) (Merzagora et al., 2009a), which allows statistical characterization of the classification performance.

The following steps can summarize the overall procedure for classification:
1. Extract fNIRS-related features (32 features = $M_{\text{fNIRS}}$), with 7 instances for each subject

2. Assign to each instance the true label, describing the corresponding group (“TBI group” or “HC group”)

3. Implement the mLOO procedure, by leaving out in turn each instance for testing and using the other instances to create 10 random training subsets.

4. Apply the RSM by creating an ensemble of 1000 classifiers based on fNIRS-related features.

5. Using the 10 ensembles created from each of the 10 random training subsets, determine the label of the held-out instance.

In this application support vector machines (SVM) and decision trees (DT) were employed. Additional details and the full description of the classification procedures are presented in Appendix A.

![Figure 3.1 Confusion matrix.](image)

The confusion matrix is a square matrix that gives information about the number of instances from one class that are labeled as belonging to the same or another class. The confusion matrix defines also true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).
3.2.4 Characterization of classification performance

The $N$ available instances in the set $S$ were each classified 10 times using the mLOO, allowing a statistical characterization of performance indices. Based on the definition of true positives ($TP$), true negatives ($TN$), false positives ($FP$) and false negatives ($FN$) provided in the confusion matrix of Figure 3.1, the following indices were computed:

- **accuracy**, defined as the probability of correctly classifying an instance and computed as the percentage of correctly classified instances:

\[
Accuracy = \frac{TP + TN}{TP + FN + TN + FP}
\]  
(Eq. 3.1)

- **sensitivity** (or recall), defined as the probability of the test to correctly identify the “TBI group” class and computed as the ratio of the number of correctly classified “TBI group” instances to the total number of “TBI group” instances:

\[
Sensitivity = \frac{TP}{TP + FN}
\]  
(Eq. 3.2)

- **specificity**, defined as the probability of the test to correctly identify the “HC group” class and computed as the ratio of the number of correctly classified “HC group” instances to the total number of “HC group” instances:

\[
Specificity = \frac{TN}{TN + FP}
\]  
(Eq. 3.3)

- **positive predictive value** ($PPV$, or precision), defined as the probability of a sample classified as “TBI group” to truly belong to the “TBI group”:

\[
PPV = \frac{TP}{TP + FP} = \frac{sensitivity \cdot prevalence}{sensitivity \cdot prevalence + (1 - specificity)(1 - prevalence)}
\]  
(Eq. 3.4)

where prevalence is defined as the total number of TBI instances divided by the total number of instances.
- **negative predictive value** (NPV), defined as the probability of a sample classified as “HC group” to truly belong to the “HC group”:

\[
NPV = \frac{TN}{TN + FN} = \frac{\text{specificity} \cdot (1 - \text{prevalence})}{\text{specificity} \cdot (1 - \text{prevalence}) + (1 - \text{sensitivity}) \cdot \text{prevalence}}
\]  
(Eq. 3.5)

- **F<sub>1</sub> measure**, defined as the harmonic mean of PPV and sensitivity and representing the effectiveness of the classification when PPV and sensitivity are equally important (Van Rijsbergen, 1979):

\[
F_{1\text{- measure}} = 2 \cdot \frac{\text{PPV} \cdot \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}}
\]  
(Eq. 3.6)

### 3.2.5 Statistical evaluation of classification performance

In order to test if the fNIRS-based classification between TBI and HC group performed significantly better than chance, a permutation test was used. In this test, the same procedure was used as in the real classification, but the true labels were assigned randomly to each instance.

The (“real”) classification performance was compared with the results of the permutation test (“chance”) using a one-way MANOVA whose factor was the **Classification type**, with two levels: “real” and “chance”. Since all variables were highly correlated (\(|r|>0.8\)) one with the other, the one-way MANOVA was performed only on accuracy, sensitivity and specificity (in order to avoid violations of the assumptions because of multicollinearity).

All post-hoc comparisons were performed with Tukey’s honest significance test; all results were considered significant at \( p<0.05 \).
The effect size was measured using the partial $\eta^2$ for the omnibus test and using the Cohen’s $d$ for the univariate ANOVAs.

### 3.3 Results

The classification between the healthy and TBI subjects was performed using 32 features extracted from fNIRS data collected during a working memory task. Classification performances were analyzed after averaging across the two types of classifiers used in the study (SVMs and DTs). Overall the classification yielded a mean accuracy, sensitivity and specificity of 74.18%, 78.39% and 56.43%, respectively. Additionally, PPV (88.54%), NPV (38.00%) and F$_1$-measure (83.02%) were evaluated (Table 3.1 and Figure 3.2). The classification performances were then tested against chance, by comparing them with the results of a permutation test (Table 3.1 and Figure 3.2).

The “real” classification performance was compared with the permutation test results by means of a one-way MANOVA performed on accuracy, sensitivity and specificity. This analysis revealed a significant difference between the performance in the real classification and in the permutation test at the multivariate level (Hotelling’s trace $=4.190$, $F(3,36)=53.886$, $p<0.001$, partial $\eta^2 = 0.818$). Univariate one-way ANOVAs performed on the single performance variables showed that real classification performed significantly better than chance in each of them: accuracy ($F(1,38)=52.525$, $p<0.001$, E.S.
$d=3.847$ with 95% C.I. [2.800 4.893]), sensitivity ($F(1,38)=32.363$, $p<0.001$, E.S. $d=3.646$ with 95% C.I. [2.635 4.658]) and specificity ($F(1,38)=6.922$, $p=0.12$, E.S. $d=0.832$ with 95% C.I. [0.186 1.478]).

![Figure 3.2](image)

**Figure 3.2** fNIRS-based classification compared with permutation test results. Bar plot comparing the (“real”) classification performances with the results obtained in the permutation test (“chance”). The height of the bar represents the mean and the whiskers represent the 95% confidence intervals of the mean.

**Table 3.1** Descriptives of fNIRS-based classification and permutation test results. The “real” classification performance was compared with the results obtained in a permutation test (“chance”), which represented the performance that would be obtained by chance. The table reports the mean and the 95% confidence interval of the mean (LL: lower level; UL: upper level), for both “real” classification and “chance”. It also reports the effect size $d$ and its 95% confidence interval.

(1) Cohen’s $d=(m_{\text{REAL}}-m_{\text{CHANCE}})/s.d_{\text{REAL}}$ where a positive value indicates a larger mean performance in the real classification.

<table>
<thead>
<tr>
<th></th>
<th>REAL</th>
<th>CHANCE</th>
<th>Effect Size $^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>95% C.I. [LL UL]</td>
<td>mean</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>74.18</td>
<td>[70.69 77.67]</td>
<td>56.55</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>78.39</td>
<td>[73.74 83.04]</td>
<td>59.91</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>56.43</td>
<td>[48.66 64.21]</td>
<td>42.14</td>
</tr>
<tr>
<td>PPV (%)</td>
<td>88.54</td>
<td>[86.74 90.33]</td>
<td>81.50</td>
</tr>
<tr>
<td>NPV (%)</td>
<td>38.00</td>
<td>[34.39 41.62]</td>
<td>19.98</td>
</tr>
<tr>
<td>$F_1$-measure (%)</td>
<td>83.02</td>
<td>[79.65 86.39]</td>
<td>68.29</td>
</tr>
</tbody>
</table>
3.4 Discussion

We developed an fNIRS-based approach to the classification between traumatic brain injury subjects and healthy individuals. The classification was performed using an ensemble of classifiers, with each classifier trained on a random subset of fNIRS features (random subspace method). The effectiveness of this strategy was tested against chance with a permutation test. This is the first proof of concept study to investigate fNIRS for the classification between group of individuals with and without cognitive impairments and, as such, it should be regarded as a proof of concept.

The main result of this work is that fNIRS-based classification does perform significantly better than chance, yielding 74.18% accuracy, 78.39% sensitivity and 56.43% specificity on average (see Figure 3.5 and Table 3.1). The effect size for the omnibus test was large, as described by the partial $\eta^2$. The improvement over a chance classification for the single variables was been measured in terms of Cohen’s $d$. For all performance indices, with the exclusion of specificity, the size of the effect was large (>0.8) (see Table 3.1). This corroborates the results of the statistical analysis by confirming that the statistically significant improvements over chance were not trivial in size. In the case of specificity, although the mean of the effect size was 0.832, the 95% confidence interval ranged from 0.186 to 1.478: such high variability does not allow qualifying the effect as large. However, since any effect size measure is greatly influenced by the sample size, the effect size variability encountered for specificity would be reduced in future studies with an increased sample size.
The performance values achieved in this study are overall consistent with those obtained by other groups that applied pattern recognition to fNIRS. In one of the first brain-computer interface studies using fNIRS, Coyle and colleagues (Coyle et al., 2004) employed fNIRS to detect characteristic hemodynamic responses elicited by motor imagery, obtaining an accuracy of approximately 75% (although the details of the specific classification algorithm that was used are not reported). Similar accuracy values (73%-89%) were obtained by Sitaram and colleagues (Sitaram et al., 2007) in another motor imagery application using SVM and hidden Markov models. Other applications of fNIRS to binary brain-computer interfaces involving cognitive or emotional tasks (Ayaz et al., 2009; Butti et al., 2007; Power et al., 2010; Tai and Chau, 2009) achieved mean accuracies ranging between 67% and 94%. However, no clear comparison of the results is possible. First of all, most of the studies do not involve classification among subjects but of different within-subject states. Second, studies that involve motor tasks or motor imagery rely on much stronger signals, thus boosting the performance of classification algorithms. Third, most of the studies report only accuracy values but no information about sensitivity and specificity; one study reporting such values (Butti et al., 2007) achieved 57% sensitivity and 87% specificity.

A random subspace method was used to create the different classifiers employed in the ensemble decision-making strategy. This strategy included all the available fNIRS features and not just the ones significantly different between the TBI and healthy group. In fact features that are not significantly different when comparing the two groups (in a one-dimensional space) can nevertheless improve the separation between the two groups when combined with other features (in a higher dimensional space). The use of all available features as inputs to a single classifier would have requested a much higher
sample size: the number of features used relates in fact to the number of parameters that a classifier needs to ‘learn’, therefore demanding a larger number of samples (a problem known as ‘curse of dimensionality’). A dimensionality reduction approach was thus necessary. In order to identify the subset of fNIRS features yielding the best performance, the exhaustive search (or brute force search) should have been used. This algorithm would systematically test all possible combination of features and pick the best performing one. However, given the 32 fNIRS features, the number of combinations to tests would have been \[ \sum_{m=1}^{M_{fNIRS} - 1} \frac{M_{fNIRS}!}{m! (M_{fNIRS} - m)!} \] \[=4.295*10^9 \]. In such cases, suboptimal searches are usually applied. One additional reason to choose the random subspace method over other (suboptimal) searches is that it can handle missing values in the data. If a particular data instance lacks a feature, the classification will be in that case performed based only on the classifiers that do not use the missing feature. The advantage provided by such approach was indeed noteworthy. In a preliminary evaluation, in fact, only the fNIRS features significantly different between the TBI and HC group were included in the classification and the achieved accuracy and F1-measure were notably lower (for accuracy the mean was 67.8% with 95% confidence intervals [66.1% 69.4%]; for F1-measure the mean was 76.8% with 95% confidence intervals [74.7% 79.8%]). Support vector machines and decision trees were chosen as classifiers because they are suitable for non-normally distributed data and for short training times. Additionally, a preliminary evaluation showed that they performed better than some common linear and non-linear parametric methods (Euclidean classifier and quadratic classifier).
CHAPTER 4: CLASSIFICATION CAPABILITY OF MULTIMODAL APPROACH

4.1 Introduction

Functional neuroimaging techniques, especially positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), have most often been applied to traumatic brain injury (TBI) patients. Functional neuroimaging can provide unique insight into the investigation and objective assessment of the cognitive rehabilitation approaches. As proposed by Strangman and colleagues (Strangman et al., 2005), functional neuroimaging has the potential to help distinguish between restorative effects (i.e. when recovery of specific neuropsychological or regional brain functions are observed) in contrast with compensatory effects (i.e. when alternative strategies are used that rely on other abilities and brain regions to bypass impairments). The possibility to capitalize on the advantage offered by functional neuroimaging modalities has however been so far limited by their main weakness: they rely on expensive technologies and on experimental task that are not ‘ecologically valid’ in reference to real-world functional behaviors.

Functional near infrared spectroscopy (fNIRS) does offer some advantages over the traditional imaging techniques, since it is affordable and noninvasive and it can be implemented in a portable system. Therefore it could be deployed in a variety of experimental settings that could provide information about the brain activation in more
realistic situations (Izzetoglu et al., 2007), a key characteristic for neurorehabilitation studies (Arenth et al., 2007).

However, the connection between brain function and behavior is intrinsically multifaceted, both in normal and in pathological conditions. Therefore an integrated approach with technologies able to investigate brain activity might be of considerable value in providing information on the interplay between the activation of different brain areas, the elicitation of neurophysiologic events of interest and behavior. Additionally, deepening out knowledge of the mechanisms regulating the neurologic sequelae of brain damage might prompt newer and more efficacious therapeutic and rehabilitative strategies for neurologic diseases. There is in fact growing consensus that a multimodal approach to the investigation of cognitive processes holds some potential (Beese et al., 1998; Benar et al., 2007; He and Lian, 2002; Mulert et al., 2004; Schiff, 2006).

In this study we investigated the combination between fNIRS and electroencephalography (EEG) for the investigation of TBI-related cognitive impairments. EEG is in fact a very well established technology, cost-effective and wearable, that has been employed in a few studies on brain injury (Duncan et al., 2005; Hensel et al., 2004; Keren et al., 1998; Lew et al., 2004; Lew et al., 2005; Segalowitz et al., 1997; Solbakk et al., 2002). As in the case of the fNIRS-based classification presented in Chapter 3, this certainly represents a gross classification but it offers the possibility of a basic investigation of the advantages offered by this the multimodal approach. To the best of our knowledge, only two preliminary studies have been published so far (Hirshfield et al., 2009; Merzagora et al., 2009b) that compare the classification abilities of fNIRS and EEG in differentiating between working load
conditions in healthy subjects and only one of them (Merzagora et al., 2009b) has also preliminary investigated the capability of a combined fNIRS-EEG classification. Therefore the fNIRS-EEG multimodal approach to classification between healthy and traumatic brain injury subjects needs an initial proof-of-concept. The knowledge obtained from this basic investigation could be further developed to categorize TBI individuals based on their working memory impairments and to follow the changes induced by neurorehabilitation intervention.

4.2 Methods

4.2.1 Subjects and experimental paradigm

The same subjects pool of Chapter 2 was included. It consisted of a total of 17 adults participated in the study: 11 healthy subjects and 6 traumatic brain injury subjects. All participants were right-handed, with vision correctible to 20/20, denied any history of neurological disorders, psychiatric illness or substance abuse. In order to be included in the study, participants with traumatic brain injury needed to present an injury diagnosis as defined by the NIDRR Traumatic Brain Injury Model Systems National Database (Harrison-Felix et al., 1996). Additional details about the inclusion/exclusion criteria and about the demographics of the two groups are presented in Chapter 2. The subjects performed a visual n-back task with four load conditions (0-back, 1-back, 2-back and 3-back) that incrementally varied the working memory load, as described in Chapter 2.
4.2.2 Data acquisition and processing

The hemodynamic activity of the prefrontal cortex was recorded using a continuous-wave fNIRS device. Changes in light absorption were converted to changes in concentration of oxyhemoglobin (HbO) and deoxyhemoglobin (HHb). From the overall data, epochs locked to the beginning of a single load condition presentation were extracted and the maximum ($H_{max}$) in the single epoch was used as feature describing the hemodynamic response for each variable (HbO, HHb and total hemoglobin Hb$_{TOT}$). Additionally, the maximum ratio between HHb and HbO was considered ($MR$). The regions of interest for this task were identified to be the left and right dorsolateral prefrontal cortex based on previous neuroimaging studies (Cabeza and Nyberg, 2000; Owen et al., 2005). fNIRS data acquisition and processing is presented in further detail in Chapter 2.

Electroencephalographic data were recorded from surface electrodes placed at International 10-20 System locations and referenced to linked mastoid leads. The locations of interest were the midline electrodes Pz, Cz and Fz, given their particular importance in P300-related studies (Polich and Kok, 1995). Vertical and horizontal electrooculograms (VEOG and HEOG) were monitored through electrodes placed above and below the left eye and at the left and outer canthi respectively. EEG signals were collected using a NuAmp amplifier (Neuroscan Inc., El Paso, TX); all impedances were systematically kept below 10 kΩ and the amplification was set to 50 mV/mm. EEG data were filtered between 0.15 Hz and 100 Hz (-6 dB/octave), using an analog filter, and sampled at 500 samples/s. Eyeblink artifacts were minimized using Jung’s Independent Component Analysis (ICA) approach (Jung et al., 2000a; Jung et al., 2000b). Stimulus-locked ERPs were extracted as 2300 ms epochs, using a 800 ms pre-stimulus baseline
and a 1500 ms post-stimulus response window. Epochs were baseline corrected by subtracting the mean of the baseline window from the full epoch. Epochs containing significant movement or muscle artifacts were discarded and only epochs containing correct responses were included in this analysis. For each subject, average ERPs were calculated in each load condition for the following responses: *true matches (TM)*, defined as the matches that were correctly identified as such by the subject; *correct non-matches (NM)*, defined as the non-matches that were correctly identified as such by the subject. The amplitude of the P300 peak was automatically identified as the amplitude of the largest positive deflection in the 250-600 ms post-stimulus response and then averaged across channels and epochs within the single presentation of each n-back load condition.

First, two confirmatory analyses were performed on the P300 peak amplitudes (averaged across channels) of the healthy subjects: 1) the P300 peak amplitude of TM and NM was compared using a Wilcoxon signed ranks test, after averaging across load condition; 2) the effect of the load condition on the P300 peak amplitude of TM was analyzed with Friedman’s ANOVA for ranks. The post-hoc tests for Friedman’s ANOVA for ranks were performed using separate Wilcoxon signed ranks tests. After these preliminary analyses, the P300 peak amplitude of TM (averaged across load condition) was compared between HC and TBI subjects with a Mann-Whitney U test. The threshold for significance was at $\alpha=0.05$.

### 4.2.3 Classification

Both fNIRS and EEG contributed features to the classification between HC and TBI subjects. As described in Chapter 3, a total of 32 fNIRS-related features was used to
classify between healthy (HC) and TBI subjects. The features consisted in the maximum value $H_{max}$ for the three hemodynamic variables and the maximum ratio $MR$ in the four load conditions at the two regions of interest. The EEG-related features consisted in the P300 amplitude (averaged across channels) for the four load conditions.

The same procedure was used as explained in Chapter 3. For each subject, multiple instances (7) of the proposed features were extracted, one for each presentation of a given load condition. Training and testing were performed using a modified leave-one-out (mLOO) cross-validation (Merzagora et al., 2009a).

Four different classification approaches were evaluated for their ability to discriminate individuals as belonging to either the HC group or the TBI group:
- *fNIRS-based classification*, where only the features extracted from fNIRS recording were used;
- *EEG-based classification*, where only the features extracted from EEG recordings were used;
- *Feature-level fusion*, where the features extracted from both fNIRS and EEG recordings were used and combined in a single feature vector;
- *Decision-level fusion*, where the classification was performed separately using the fNIRS and EEG features and the results were then combined to reach the final decision.

*fNIRS-based classification*: A random subspace method (RSM) approach was used so that all 32 fNIRS-related features could contribute to the classification. For each instance, the estimated label was assigned based on a weighted majority-voting of 1000
classifiers, each based on a random subset of the 32 features (see Appendix A for additional details).

**EEG-based classification:** For the classification between HC and TBI group based on EEG features, a bagging approach was applied (Breiman, 1996): the decision about the label to assign to a given instance is taken by an ensemble of 100 classifiers (each trained on a bootstrapped replica of the training subset) whose decisions (labels) were combined using a weighted majority-voting rule (see Appendix A for additional details).

**Feature-level fusion classification:** In this approach the fNIRS-related features and the EEG-related features were combined in one single feature vector. Thus a total of 36 features were used: 32 from the fNIRS recordings and 4 from the EEG recordings. A procedure similar to the one used for fNIRS-based classification is performed (for details, see Appendix A): an RSM approach was used to create an ensemble of classifiers (based on different combinations of the 36 features) that assigned a label to each instance.

**Decision-level fusion classification:** The same procedures used in fNIRS-based classification and EEG-based classification are applied until the labels assigned by the classifiers of each algorithm are obtained. The final decision on the estimated label for a given instance is obtained by applying a weighted majority-voting rule on the labels of all the classifiers obtained through fNIRS-based and EEG-based approaches (see Appendix A for details).
4.2.3 Characterization and statistical evaluation of classification performance

The classification performances (measured in terms of accuracy, sensitivity and specificity, PPV, NPV and F1-measure, as defined in Chapter 3) were analyzed with single repeated measures ANOVAs. The repeated factor was the Classification approach, with four levels (“fNIRS-based classification”, “EEG-based classification”, “Feature-level fusion classification” and “Decision-level fusion classification”). The Geisser-Greenhouse correction was applied; as follow-up analyses, five paired t-tests were performed between the results of fNIRS-based classification and the other approaches (3 t-tests) and between EEG-based classification and the fusion approaches (2 t-tests). All results were considered significant at p<0.05. The effect sizes were measured using the Cohen’s $d$. Five different effect sizes were calculated for the performance indices. The effect size $d_{\text{fNIRS-EEG}}=(m_{\text{fNIRS}}-m_{\text{EEG}})/s.d._{\text{fNIRS}}$ compared the performance of fNIRS-based classification with the performance of EEG-based classification. The effect size $d_{\text{fNIRS-FL}}=(m_{\text{fNIRS}}-m_{\text{FL}})/s.d._{\text{fNIRS}}$ compared the performance of fNIRS-based classification with the performance of Feature-level fusion (FL) classification. The effect size $d_{\text{fNIRS-EEG}}=(m_{\text{fNIRS}}-m_{\text{EEG}})/s.d._{\text{fNIRS}}$ compared the performance of Feature-level fusion classification with the performance of EEG-based classification. The effect size $d_{\text{DL-fNIRS}}=(m_{\text{fNIRS}}-m_{\text{DL}})/s.d._{\text{DL}}$ compared the performance of Decision-level fusion classification with the performance of fNIRS-based classification. The effect size $d_{\text{DL-EEG}}=(m_{\text{DL}}-m_{\text{EEG}})/s.d._{\text{DL}}$ compared the performance of Decision-level fusion classification with the performance of EEG-based classification.
4.3 Results

4.3.1 EEG results

An initial analysis was performed selectively on the healthy control group to test if the P300 amplitude extracted for TM was higher than the amplitude for NM. The median amplitude for TM was 11.52 µV, with a median absolute deviation (M.A.D.) of 5.01 µV (mean: 13.48 µV). For NM, the median amplitude was 10.36 µV, with a median absolute deviation of 1.28 µV (mean: 9.94 µV). The analysis did confirm that in the control group the P300 amplitude of TM is significantly larger than in NM (Wilcoxon signed rank test $z(10)=-4.344$, $p<0.001$). Additionally the trend of the P300 amplitude for different n-back conditions showed that, with the exclusion of the highest load (3-back), the amplitude decreases with increasing working memory loads. This trend was tested with a Friedman’s ANOVA, which confirmed a significant effect of the working memory load on the P300 amplitude ($\chi^2(2)=8.937$, $p=0.011$). Subsequent contrasts, performed using separate Wilcoxon signed ranks tests, revealed that the P300 amplitude in the 0-back condition was significantly larger than in the 2-back condition ($z(10)=-2.395$, $p=0.017$). When investigating the effect of working memory load on the P300 amplitude in the TBI group, the overall trend of decreasing amplitude for increasing load did not reach significance.

The comparison between the two groups (Table 4.1) revealed that the TBI group elicited significantly smaller P300 peak amplitudes in response to TM ($z(15)=-2.822$, $p=0.005$, E.S. $d=0.288$, 95% C.I. of $d$ [0.149 0.426]). (The effect size was measured as $d=(m_{HC}-m_{TBI})/s.d_{HC}$ where a positive value indicates a larger P300 peak amplitude in the HC group).
Table 4.1 P300 peak amplitude by load condition and overall. Amplitude of the P300 peak recorded in the n-back task, assessed both globally and at the single-load level. The table reports the mean, the median and the median absolute deviation (M.A.D.). All values are in µV.

<table>
<thead>
<tr>
<th></th>
<th>HC mean ± M.A.D.</th>
<th>TBI mean ± M.A.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-back</td>
<td>12.05 ± 5.15</td>
<td>11.79 ± 5.15</td>
</tr>
<tr>
<td>1-back</td>
<td>12.96 ± 5.73</td>
<td>11.33 ± 5.73</td>
</tr>
<tr>
<td>2-back</td>
<td>12.86 ± 3.99</td>
<td>10.75 ± 3.99</td>
</tr>
<tr>
<td>3-back</td>
<td>14.68 ± 4.85</td>
<td>12.08 ± 4.85</td>
</tr>
</tbody>
</table>

4.3.2 Classification results

The classification between healthy and TBI subjects was performed using four different approaches: i) using only fNIRS-related features, ii) using only EEG-related features, iii) using a combination of fNIRS- and EEG-related features or iv) combining the decisions taken based on those features.

The mean accuracy achieved in the classification ranged between 67.12% (EEG-based classification) to 75.41% (Feature-level fusion classification). The mean sensitivity ranged between 64.15% (EEG-based classification) and 79.83% (Feature-level fusion classification), while the mean specificity ranged between 56.43% (fNIRS-based classification) and 79.64% (Feature-level fusion classification). The values for all performance indices, divided by classification approach, are presented in Table 4.2 and Figure 4.1.

As described in Chapter 3, the repeated measure ANOVAs were performed only on accuracy, sensitivity and specificity as a measure to reduce Type I error. Mauchly’s test indicated that the assumption of sphericity had been violated (accuracy:
\( \chi^2(5) = 25.171, \ p < 0.001; \) sensitivity: \( \chi^2(5) = 25.819, \ p < 0.001; \) specificity: \( \chi^2(5) = 21.274, \ p = 0.001 \), therefore \( p \)-values were corrected using the Greenhouse-Geisser adjustment.

The results show that the Classification approach significantly affected accuracy \( F(3,57) = 12.167, \ p < 0.001, \) partial \( \eta^2 = 0.390 \), sensitivity \( F(3,57) = 39.765, \ p < 0.001, \) partial \( \eta^2 = 0.677 \) and specificity \( F(3,57) = 51.283, \ p < 0.001, \) partial \( \eta^2 = 0.730 \).

**Table 4.2** Descriptives of results for different classification approaches.

Four classification approaches were investigated: “fNIRS-based classification”, “EEG-based classification”, “Feature-level fusion classification” and “Decision-level fusion classification”. The table reports the mean and the 95% confidence interval of the mean (LL: lower level; UL: upper level), for each of them.

<table>
<thead>
<tr>
<th></th>
<th>fNIRS-based</th>
<th>EEG-based</th>
<th>Feature-level</th>
<th>Decision-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>74.18</td>
<td>67.12</td>
<td>75.41</td>
<td>71.09</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[72.03, 76.32]</td>
<td>[64.45, 69.79]</td>
<td>[72.78, 78.04]</td>
<td>[68.12, 74.07]</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>78.39</td>
<td>64.15</td>
<td>79.83</td>
<td>71.95</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[76.02, 80.76]</td>
<td>[60.26, 68.05]</td>
<td>[76.61, 83.04]</td>
<td>[67.39, 76.51]</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>56.43</td>
<td>79.64</td>
<td>56.79</td>
<td>67.50</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[48.39, 64.47]</td>
<td>[74.64, 84.64]</td>
<td>[50.61, 62.97]</td>
<td>[62.13, 72.87]</td>
</tr>
<tr>
<td>PPV (%)</td>
<td>88.54</td>
<td>93.24</td>
<td>88.69</td>
<td>90.53</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[86.71, 90.36]</td>
<td>[91.84, 94.63]</td>
<td>[87.28, 90.11]</td>
<td>[89.45, 91.60]</td>
</tr>
<tr>
<td>NPV (%)</td>
<td>38.01</td>
<td>35.11</td>
<td>41.21</td>
<td>37.79</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[34.04, 41.97]</td>
<td>[32.92, 37.30]</td>
<td>[36.48, 45.93]</td>
<td>[34.55, 41.03]</td>
</tr>
<tr>
<td>F1-measure (%)</td>
<td>83.02</td>
<td>75.65</td>
<td>83.87</td>
<td>79.76</td>
</tr>
<tr>
<td>95% C.I. [LL, UL]</td>
<td>[81.49, 84.55]</td>
<td>[73.14, 78.15]</td>
<td>[81.92, 85.82]</td>
<td>[77.13, 82.38]</td>
</tr>
</tbody>
</table>

**Figure 4.1** Comparison of performances obtained by different classification approaches. Mean and 95% confidence intervals obtained by the four classification approaches in accuracy, sensitivity and specificity.
Follow-up paired t-tests between the four classification approaches revealed that:

- in terms of accuracy, EEG-classification achieved performances significantly lower than fNIRS-based classification ($t(19)=3.820$, $p=0.001$, E.S. $d_{\text{fNIRS-EEG}}=1.539$, 95% C.I. of $d$ [0.834 2.246]), Feature-level fusion classification ($t(19)=-4.944$, $p<0.001$, E.S. $d_{\text{FL-EEG}}=1.477$, 95% C.I. of $d$ [0.777 2.176]) and Decision-level fusion classification ($t(19)=-4.781$, $p<0.001$, E.S. $d_{\text{DL-EEG}}=0.624$, 95% C.I. of $d$ [-0.010 1.259]);

- in terms of sensitivity, fNIRS-based classification performed significantly better than either EEG-based classification ($t(19)=7.799$, $p<0.001$, E.S. $d_{\text{fNIRS-EEG}}=2.810$, 95% C.I. of $d$ [1.937 3.684]) or Decision-level fusion classification ($t(19)=3.251$, $p=0.004$, E.S. $d_{\text{DL-fNIRS}}=-0.662$, 95% C.I. of $d$ [-1.298 -0.025]); additionally, EEG-based classification achieved performances significantly lower than either Feature-level fusion ($t(19)=-8.599$, $p<0.001$, E.S. $d_{\text{FL-EEG}}=2.283$, 95% C.I. of $d$ [1.487 3.080]) and Decision-level fusion ($t(19)=-9.210$, $p<0.001$, E.S. $d_{\text{DL-EEG}}=0.801$, 95% C.I. of $d$ [0.157 1.445]);

- in terms of specificity, both EEG-based classification ($t(19)=-7.377$, $p<0.001$, E.S. $d_{\text{fNIRS-EEG}}=-1.352$, 95% C.I. of $d$ [-2.039 -0.665]) and Decision-level fusion classification ($t(19)=4.971$, $p<0.001$, E.S. $d_{\text{DL-fNIRS}}=0.966$ with 95% C.I. [0.311 1.621]) outperformed fNIRS-based classification; additionally, EEG-based classification performed significantly better than Feature-level fusion classification ($t(19)=9.718$, $p<0.001$, E.S. $d_{\text{FL-EEG}}=-1.731$, 95% C.I. of $d$ [-2.457 -1.004]) and Decision-level fusion classification ($t(19)=7.033$, $p<0.001$, E.S. $d_{\text{DL-EEG}}=-1.059$, 95% C.I. of $d$ [-1.721 -0.397]).

The effect sizes comparing all performance indices and calculated as described in section 4.2 are reported in Table 4.3 and Table 4.4.
Table 4.3 Effect sizes comparing single-modality and multimodal classification.
The single-modality classification approaches (fNIRS-based and EEG-based) were
compared with multimodal classification approaches (Feature-level fusion and Decision
level fusion) and the effect sizes were calculated. The table reports the mean and the 95%
confidence interval of the mean (LL: lower level; UL: upper level), for each of them.
(fNIRS: fNIRS-based classification; EEG: EEG-based classification; FL: Feature-level
fusion classification; DL: Decision-level fusion classification).
(1) Cohen’s $d_{FL\rightarrow fNIRS}=(m_{FL}-m_{fNIRS})/s.d._{FL}$ where a positive value indicates a larger mean
performance in Feature-level fusion classification. (2) Cohen’s $d_{FL\rightarrow EEG}=(m_{FL}-
m_{EEG})/s.d._{FL}$ where a positive value indicates a larger mean performance in Feature-level
fusion classification. (3) Cohen’s $d_{DL\rightarrow fNIRS}=(m_{DL}-m_{fNIRS})/s.d._{DL}$ where a positive value indicates a larger mean performance in Decision-level fusion classification. (4) Cohen’s $d_{DL\rightarrow EEG}=(m_{DL}-m_{EEG})/s.d._{DL}$ where a positive value indicates a larger mean performance in Decision-level fusion classification.

<table>
<thead>
<tr>
<th>Effect size $d_{FL\rightarrow fNIRS}^{(1)}$</th>
<th>Effect size $d_{FL\rightarrow EEG}^{(2)}$</th>
<th>Effect size $d_{DL\rightarrow fNIRS}^{(2)}$</th>
<th>Effect size $d_{DL\rightarrow EEG}^{(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean 0.219</td>
<td>1.477 [0.777 2.176]</td>
<td>0.624 [-0.010 1.259]</td>
<td></td>
</tr>
<tr>
<td>95% C.I. [-0.402 0.841]</td>
<td>95% C.I. [-1.113 0.144]</td>
<td>95% C.I. [0.157 1.445]</td>
<td></td>
</tr>
<tr>
<td>LL UL</td>
<td>LL UL</td>
<td>LL UL</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>mean 0.209</td>
<td>2.283 [1.487 3.080]</td>
<td>0.801 [0.157 1.445]</td>
<td></td>
</tr>
<tr>
<td>95% C.I. [-0.412 0.831]</td>
<td>95% C.I. [-1.298 -0.025]</td>
<td>95% C.I. [-1.721 -0.397]</td>
<td></td>
</tr>
<tr>
<td>LL UL</td>
<td>LL UL</td>
<td>LL UL</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4 Effect sizes comparing fNIRS-based and EEG-based classification.
The effect size for the comparison between the single-modality classification approaches
(fNIRS-based and EEG-based) was calculated. The table reports the mean and the 95%
confidence interval of the mean (LL: lower level; UL: upper level) for the effect size.
(fNIRS: fNIRS-based classification; EEG: EEG-based classification).
(1) Cohen’s $d_{fNIRS\rightarrow EEG}=(m_{fNIRS}-m_{EEG})/s.d._{fNIRS}$ where a positive value indicates a larger mean performance in fNIRS-based classification.

<table>
<thead>
<tr>
<th>Effect size $d_{fNIRS\rightarrow EEG}^{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean 1.539 [0.834 2.245]</td>
</tr>
<tr>
<td>95% C.I.</td>
</tr>
<tr>
<td>LL UL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean 2.810 [1.937 3.684]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% C.I.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LL UL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| mean -1.352 [-2.039 -0.665] |
| 95% C.I. |
| LL UL |
4.4 Discussion

In this study we compared the ability to classify between traumatic brain injury subjects and healthy individuals when using either fNIRS-related features alone or EEG-related features alone or a combination of the two modalities.

The comparison between fNIRS data recorded from healthy subjects and TBI subjects was presented in Chapter 2. On EEG, instead, we present here a series of analyses performed to confirm the consistency of our recordings with the existing literature.

First, the P300 peak amplitudes extracted in the healthy group were compared between true matches (infrequent events to which the subjects responded correctly) and correct non-matches (frequent events to which the subjects responded correctly). Based on the existing literature, we expect that infrequent (or target) stimuli elicit a positive deflection occurring about 300ms-500ms after stimulus onset that is larger than that elicited by frequent (or non-target stimuli): this is due to the lower probability of the matching stimuli and to the fact that they correspond to a target item (Naatanen and Picton, 1986; Polich et al., 1985). The P300 peak amplitude in the control group was confirmed to be larger for true matches than for correct non-matches.

Second, we confirmed that for increasing working memory loads the P300 peak amplitude in true matches decreases. This is consistent with the general conceptualization of the P300 amplitude as indexing a pool of processing resources that can be allocated among tasks (Donchin et al., 1986; Kramer and Spinks, 1991). In particular previous studies (McEvoy et al., 1998; Watter et al., 2001) that investigated the effects of varying working memory loads on P300 peak amplitude have suggested that the decrease in P300
amplitude with increasing loads reflects the reallocation of attention and processing capacity away from matching evaluation to cope with the growing working memory demands. In particular Watter and colleagues (Watter et al., 2001) suggested that n-back is by nature a dual-task paradigm, whose distinct components are the memory search (i.e. the identification what was the item presented n stimuli back) and the item matching (i.e. establishing if the currently presented item matches the item of interest). Based on this conceptualization of the n-back task, the P300 peak would mainly reflect the item matching process but its amplitude would be attenuated to the extent that the other process utilizes the same resources. Whereas the trend of decreasing P300 peak amplitude for increasing working memory load was confirmed in the healthy group, it did not reach significance in the TBI group, probably because of the limited number of subjects in the cohort.

Third, we confirmed that the P300 peak amplitude elicited by true matches in the healthy group is larger than that elicited in the TBI group. This is in line with previous studies that investigated P300-eliciting paradigms. Lew and colleagues found that TBI subjects exhibited P300 amplitudes consistently lower than healthy subjects both in visual and auditory oddball paradigms (Lew et al., 2004; Segalowitz et al., 1997; Solbakk et al., 2002) and in a socio-emotional oddball paradigm (Lew et al., 2005). The lower P300 amplitudes found in the TBI individuals have been suggested to represent an impaired organization and categorization of incoming information (Lew et al., 2004; Solbakk et al., 2002).

Therefore, the variations in P300 peak amplitude could provide information about different aspects of working memory and information processing.
The features extracted from fNIRS and EEG were used alone or in combination to classify between the healthy group and the TBI group and their performances were compared.

The direct comparison between fNIRS-based classification and EEG-based classification revealed that fNIRS-based classification achieved significantly better accuracy and sensitivity but it performed significantly lower in terms of specificity. In all three cases the effect size was measured to be large. This pattern shows that each modality offers unique advantages in the classification between healthy and TBI group. Whereas fNIRS is more capable in correctly identifying instances belonging to TBI subjects, EEG seems to perform a better identification of healthy subjects. Therefore an opportunity exists to take advantage of these different strengths in order to improve the classification that can be achieved by using each single modality.

The multimodal classification was performed following two approaches: Feature-level fusion and Decision level fusion. On one side, Feature-level fusion achieved performances comparable to those of fNIRS alone; also the improvements it offered over EEG alone are equivalent to those offered by fNIRS. On the other side, Decision-level fusion does not perform as well as fNIRS in terms of accuracy and sensitivity; however, the variability of the effect size is rather high, so it cannot be concluded that there is an important (although significant) difference between the two approaches from this point of view. A similar remark can be made also about the significant improvement offered by Decision-level fusion over EEG-based classification. In terms of specificity, the pattern is different: Decision-level fusion performs significantly better than fNIRS-based classification though significantly lower than EEG-based classification, in both cases with a large effect size. Therefore, based on the analysis of the effect sizes, we cannot conclude that important (although significant) deviations exist between Decision-level
fusion classification and fNIRS-based classification in terms accuracy and sensitivity or between Decision-level fusion and EEG-based classification in terms of specificity. Nevertheless, effect sizes show a meaningful improvement over the performance of fNIRS-based classification in terms of specificity.

Taken together, these results suggest that there is indeed room for improvement in classification performance with respect to the single modalities. The approach to combine them should therefore take into consideration the specific strengths of each technique. In this particular case, fNIRS provides higher overall accuracy and sensitivity and EEG provides higher specificity. In order to achieve better appreciation of the specific strengths, multiple possible strategies could be applied.

In the case of the feature-level fusion classification, for example, the feature vector that was used when applying the random subspace method contained 32 fNIRS-related features and 4 EEG-related features. A possible approach to improve the overall performances could come from using a different feature sampling strategy. The method we used assigned to all features the same probability of being selected. The sampling probability could however be modified so that comparable number of features from fNIRS and EEG will be represented. It is therefore presumable that both signals would contribute similarly to the classification performances.

In the case of decision-level fusion classification, the final decision of the ensemble was based on a weighted majority-voting (where EEG-based decisions and fNIRS-based decision weighted equally), with weights assigned based on the accuracy achieved during training. A possible strategy to improve the overall performances could be to assign weights in the majority-voting based on a different “metric” that takes into account also
sensitivity and specificity. In this way, the advantages offered by the single modalities could be integrated to improve the final classification.

As a final note, a post-hoc power analysis was performed in order to estimate how many repetitions would be needed in the mLOO in order for other studies to reach adequate power with a comparable effect size. In order to be conservative, the post-hoc power analysis was performed starting from accuracy, since it offered the lowest effect size (measured in terms of partial $\eta^2$). Additionally, a conservative approach was used in the estimation of the population effect size based on the sample effect size. Based on this analysis, 15 mLOO repetitions would be needed in order to reach adequate power (0.8) when the effect size is medium (0.47) and 25 mLOO repetitions would be needed when the effect size is medium-to-small (0.35) (analyses performed using GPower 3).
CHAPTER 5: DEMONSTRATION OF EXPLORATORY APPLICATIONS

5.1 Introduction

The results presented in Chapter 2, 3 and 4 lay the ground for further investigations and new potential applications. In this chapter we present two exploratory studies that have been developed taking advantages of the results presented in the previous chapters and that could represent examples of future directions.

In the first application we suggest how the knowledge obtained from the use of functional near infrared spectroscopy (fNIRS) in the traumatic brain injury population (TBI) could be employed in the development of a therapeutic tool. The results presented in Chapter 2, in fact, suggest an impaired pattern of hemodynamic and metabolic activation (during a task involving working memory), in particular in the left dorsolateral prefrontal cortex (DLPFC). Therefore, fNIRS could be used to monitor directly the physiological effects of a therapeutic intervention that targets the local metabolic level. A possible technology that could be investigated for the treatment of such hemodynamic and metabolic impairments following TBI is transcranial direct current stimulation (tDCS). tDCS has the potential to improve working memory performance (Fregni et al., 2005; Marshall et al., 2005) and it is known to modulate the cortical excitability (Paulus, 2004; Priori, 2003) and, as a consequence, the regional blood flow and metabolism. Here we present evidence that fNIRS can be effectively used to monitor these physiological changes induced by tDCS (Merzagora et al., 2010).
In the second application we present a preliminary proof-of-concept of the fact that fNIRS and EEG could be employed to assess cognitive performance within group. In Chapter 3 and 4 we presented the use of fNIRS and EEG to categorize between healthy subjects and subjects with TBI. While this is a gross classification, it represented the first step needed towards a classification of levels of impairments. Here we present preliminary evidence that fNIRS and EEG can be used to assess the performance of healthy subjects in a working memory task (Merzagora et al., 2009b). A more detailed investigation could therefore extend this to the TBI population, bringing a physiologically-based approach to the assessment of a continuum of impairments and providing new means for the monitoring of changes induced by neurorehabilitation.

5.2 Effects of anodal tDCS on prefrontal hemodynamic activity

5.2.1 Introduction

The majority of human data on working memory function has been derived initially from lesion studies and neuropsychological investigations. In the last decades, different neuroimaging approaches have provided additional information on the working memory processes (Cabeza and Nyberg, 2000; Smith and Jonides, 1999). More recently, neuromodulation techniques such as transcranial direct current stimulation (tDCS) and transcranial magnetic stimulation (TMS) have been proven capable of ‘interfering’ with working memory processes (Desmond et al., 2005; Fregni et al., 2005; Marshall et al., 2005; Mottaghy, 2006; Mull and Seyal, 2001; Oliveri et al., 2001; Osaka et al., 2007). On
one side, TMS relies on the application of a large but brief currents through an insulated coil placed on the subject’s scalp. These brief currents flowing in the coil induce a magnetic pulse that penetrates the skull and in turn generates small currents in the brain cortex. Because of the need to generate pulses of large currents and to handle magnetic coils, the TMS devices are usually expensive and of limited portability. On the other side, instead, tDCS is a noninvasive neuromodulation technique that is quite inexpensive and has been developed in a compact fashion. tDCS relies on the delivery of weak constant currents (0.5-2 mA) for a prolonged period of time by applying two surface electrodes on the subject’s head (Sparing and Mottaghy, 2008). Application of tDCS to the cortex has been shown to shift the membrane potential of superficial neurons in a de- or hyperpolarizing direction, and to modulate spontaneous neuronal activity as well as the processing of afferent signals (Paulus, 2004; Priori, 2003). If tDCS is continuously applied for 5 minutes or more, it can induce sustained changes in neuronal firing rates that last for many hours after the current is switched off (Bindman et al., 1962; Lang et al., 2004; Nitsche et al., 2003; Nitsche and Paulus, 2001). tDCS modulates excitability in a polarity-specific manner: anodal polarization increases excitability measures of the motor and visual cortex (Edwards et al., 1993; Paulus, 2004), whereas cathodal stimulation produces opposite effects, i.e. excitability is reduced and some functions worsen (Been et al., 2007).

Changes in cortical excitability are associated with changes of the underlying cortical neuronal activity and with subsequent changes in the regional cerebral blood flow (rCBF). For this reason, it is conceivable that a measurable output of the after-effects of tDCS – virtually in any cortical area – could be obtained by measuring the rCBF. This aspect is not trivial, as so far it has been possible to measure the after-effects of tDCS only in the motor and visual cortex (where stimulation can be evaluated by motor evoked
potentials and phosphene threshold to transcranial magnetic stimulation, respectively). However, it is not possible to assume that every different cortical area responds to tDCS in the same way. This is an important aspect as tDCS has been proposed as a possible neuromodulation treatment for many neurological and psychiatric disorders (Boggio et al., 2006; Boggio et al., 2007; Boggio et al., 2008; Ferrucci et al., 2008; Nitsche et al., 2009; Ohn et al., 2008). Both anodal and cathodal tDCS have been reported to increase the metabolism of the cortex underlying the stimulation electrodes (Lang et al., 2005).

In this study we investigated functional near infrared spectroscopy (fNIRS) as a method to obtain an objective and easily applicable measure of such neurophysiologic and metabolic effects triggered by tDCS. Compared to other traditional functional neuroimaging modalities, in fact, fNIRS is more portable, easy to apply and more cost-effective, therefore representing a good potential match to tDCS in clinical applications.

5.2.2 Methods

Subjects and experimental protocol

To evaluate whether the hemodynamic effects of anodal tDCS can be detected using fNIRS, twelve healthy volunteers (6 females and 6 males, aged 24-39 y.o., mean±S.D. 29.5±3.9) were recruited for the study. Of the twelve participants, ten (5 males and 5 females; aged 24-39 y.o., mean±S.D. 29.8±4.3) also participated in the sham stimulation condition. The participants were screened for history of hormonal, metabolic, circulatory, psychiatric and neurological disorders, and were medication-free at the time of the study. The participants were seated comfortably in a semi-darkened room, were instructed to refrain from speaking and to remain awake while in a calm, relaxed state.
All participants gave their informed consent; the procedures had the approval of the hospital ethics committee and were conducted in accordance with the declaration of Helsinki.

The experimental set up is reported in Figure 5.1. Subjects were tested under two conditions: real stimulation (12 subjects) and sham stimulation (10 subjects). After the end of the stimulation, the effects of the tDCS on the rCBF in the prefrontal cortex were monitored with a continuous-wave fNIRS system. The fNIRS system does not measure the absolute concentration value of HbO\textsubscript{2} and HHb, but their deviations from a control value. In this experimental protocol, the control value was chosen to be at the end of the fNIRS data collection. The after-effects produced by the tDCS tend in fact to decay with time (Nitsche and Paulus, 2001), and fNIRS data were collected for a sufficiently long time, such that the final hemodynamic conditions could be considered with reasonable confidence as representing the baseline state.

**Transcranial direct current stimulation (tDCS)**

The experimental protocol consisted of two conditions: real stimulation (12 subjects) and sham stimulation (10 subjects, see above). Sessions were separated by an interval of at least one week. tDCS stimulation was delivered by a battery-driven electrical stimulator (Eldith DC-Stimulator, Germany) connected to a pair of thick (0.3 cm) saline-soaked synthetic surface sponge electrodes (surface area: 35 cm\textsuperscript{2} each) placed on the scalp. Electrodes were applied bilaterally at two prefrontal locations, lateral to Fp1 (anode) and Fp2 (cathode) of the International 10-20 System for EEG electrodes placement.
Figure 5.1 Schematic representation of tDCS experimental setup. (A) The tDCS electrodes were placed bilaterally at two prefrontal locations, with the anode lateral to Fp1 and cathode lateral to Fp2. (B) After tDCS stimulation, the fNIRS probe was placed on the prefrontal area so as to cover the two stimulated locations. (C) The arrangement of the 4 light sources and 10 photodetectors on the fNIRS probe and the configuration used for data acquisition yielded 16 active optodes, each monitoring different areas of the prefrontal cortex. (D) The experiment consisted of two possible conditions: real tDCS and sham tDCS. In both cases, the intervention lasted for 10 minutes, after which the fNIRS probe was positioned and the fNIRS data collection started, lasting for 21 minutes. For both real and sham tDCS, the placement of the tDCS electrodes and of the fNIRS probe was the same.

We chose this location as it is optimal for positioning the fNIRS probe (Emir et al., 2005). The active condition (real stimulation) consisted of a 10-minute anodal stimulation with a constant current of 1mA (8 second phase in/phase out for a total stimulation time of 616 seconds), with total current density of 0.02857 mA/cm². We chose this duration as it produces – when applied over the motor cortex – a stable effect on most subjects for about 15 minutes (Nitsche and Paulus, 2001). Sham stimulation involved the same electrode placement and duration as the active condition; however, the 1mA constant current was delivered for only 30 seconds, with the same phase in/phase out time (total stimulation of 46 seconds). This stimulation in the sham condition was used to induce the slight tingling or burning sensations that some subjects report they experience during
tDCS stimulation, in order to further blind the participants as to which type of stimulation they were receiving (real or sham). As a control condition to evaluate the effects of lengthening the tDCS, in three subjects we also evaluated a tDCS duration of 15 minutes.

fNIRS data acquisition and processing

Upon the conclusion of the tDCS sessions, the effects of tDCS on rCBF in the prefrontal cortex were monitored using the continuous-wave fNIRS system described in Chapter 2. The modified Beer-Lambert law (mBLL) (see Chapter 2 for additional details) was used to convert light intensity readings into changes in oxyhemoglobin (HbO₂) and deoxyhemoglobin (HHb) concentration and the control condition was the average attenuation value recorded during the last minute of data acquisition. The values for changes in HbO₂ and HHb over time were then calculated for the other 20 minutes, yielding a total of 2400 points for each optode and each variable. Since the variations in concentration of HbO₂ and HHb induced by the tDCS are expected to occur on a much slower time scale, we reduced the number of points per optode and per variable to twenty: each point represented the average over 1 minute. Additionally, the fNIRS probe held two rows of optodes: the bottom row closer to the orbitofrontal area and the top row covering more caudal regions of the prefrontal cortex, but both spanning the entire length of the fNIRS probe. Given the position of the tDCS electrodes, there was no reason to expect the optodes in the top and bottom row to differ. Therefore, in order to further reduce the dimensionality of the data, we averaged together the signals from pairs of optodes that shared the same degree of laterality but belonged to different rows. This operation led to a total of 8 channels. Going from the one at the far left to the one to the far right, the channels were named L4, L3, L2, L1, R1, R2, R3 and R4 (these channels
were obtained by averaging respectively optodes 1-2, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14 and 15-16).

After real tDCS, the measurements of changes in HbO$_2$ and HHb from the 8 channels were analyzed as independent variables with an initial two-way MANOVA whose factors were: Hemodynamic variable, with two levels: “HbO$_2$” and “HHb”; Time, with 20 levels (one for each minute of the recording);

The effects of tDCS were observed only for HbO$_2$ and not for HHb, so this last variable was no further analyzed. The measurements of changes in HbO$_2$ from the 8 channels were then analyzed as independent variables with a two-way MANOVA whose factors were: Stimulation condition, with two levels: “real stimulation” and “sham stimulation”; Time, with 20 levels (one for each minute of the recording);

This was followed by a series of 8 univariate two-way repeated measure ANOVAs, with Stimulation condition as between-subject factor (with two levels) and Time as within-subject factor (with 20 levels). All post-hoc comparisons were performed with Tukey’s honest significance test. All results were considered significant at p<0.05.

5.2.3 Results

Overall, real tDCS stimulation (with anode over the left prefrontal cortex) was performed on a total of 12 subjects. The stimulation produced an increase in HbO$_2$, more markedly in the areas under the anode. HbO$_2$ reached a peak in about 3-6 minutes after the start of the recording and then slowly decayed, reaching again a baseline level after the effect of the stimulation vanished. On the right prefrontal area, where the cathode was placed, the time course appeared similar, though the amplitude of the increase was more
limited, probably denoting a border effect (Figure 5.2A). In contrast to the marked increases of HbO₂, the stimulation produced negligible changes in HHb (Figure 5.2B). These qualitative observations were confirmed by the subsequent statistical analyses. MANOVA revealed a significant effect for the 8 latent variables (HbO₂ and HHb values obtained from the 8 channels) as a group in relation to the Hemodynamic variable (Roy’s Largest Root=0.461; $F(8, 593)=34.15$, $p<0.0001$) and the Time factor (Roy’s Largest Root= 0.424; $F(19, 599)=13.38$, $p<0.0001$). Also, the interaction between the two factors was significant (Roy’s Largest Root= 0.122; $F(38, 599)=1.92$, $p=0.001$), which is consistent with the observed differences in the Time factor being due to HbO₂.

Figure 5.2 Spatio-temporal hemodynamic patterns following real tDCS. Spatio-temporal representation of HbO₂ (A) and HHb (B) in 12 subjects that received real tDCS. The x-axis represents the time (in minutes) of the fNIRS data collection and the y-axis represents the distribution of the 8 channels across the forehead. The HbO₂ and HHb values (in μM) are color-coded and the scale is defined by the colorbar at the right of the graphs.
Figure 5.3 Spatio-temporal HbO$_2$ patterns following real and sham tDCS. Spatio-temporal representation of HbO$_2$ obtained for the two conditions: real stimulation (A) and sham stimulation (B). Only the 10 subjects that received both real and sham stimulation have been included in these graphs. The x-axis represents the time (in minutes) of the fNIRS data collection and the y-axis represents the distribution of the 8 channels across the forehead. The HbO$_2$ values (in μM) are color-coded and the scale is defined by the colorbar at the right of the graphs. The bottom graph (C) shows the evolution of HbO$_2$ following the real and the sham stimulation (respectively the thick solid red line and the thick solid blue line). The time courses were obtained considering only the two channels (L3 and L4) that showed a significant difference between the two stimulation conditions. The thinner dotted lines represent instead the 95% confidence intervals for HbO$_2$. 
A control condition was also tested on 10 subjects who received sham stimulation. In these subjects, the spatiotemporal course of HbO₂ obtained for the real stimulation sessions (Figure 5.3A) was compared with the course obtained for the sham sessions (Figure 5.3B). The sham stimulation produced no detectable changes in HbO₂. Again, these qualitative observations were confirmed by the subsequent statistical analyses. The initial MANOVA revealed a significant effect for the 8 latent variables (HbO₂ values obtained from the 8 channels) as a group in relation to the Stimulation condition (Roy’s Largest Root=0.436; \( F(8, 332)=18.09, \ p<0.0001 \)) and the Time factor (Roy’s Largest Root= 0.230; \( F(19, 339)=4.11, \ p<0.0001 \)). Also, the interaction between the two factors was significant (Roy’s Largest Root= 0.134; \( F(19, 339)=2.39, \ p=0.001 \)), which is consistent with the observed differences in the Time factor being due to the real stimulation.

As a follow-up, a repeated-measure ANOVA was conducted on each single channel. Findings revealed markedly significant differences between the real stimulation and the sham condition in channels L3 (\( F(1,9)=7.89, \ p=0.020 \)) and L4 (\( F(1,8)=5.64, \ p=0.044 \)). Channel L3 showed a significant interaction between the stimulation condition and the time factor (\( F(1,19)=2.01, \ p=0.010 \)), confirming the differential effect of the stimulation on the time course of HbO₂. Tukey’s post hoc tests revealed that during the real stimulation session the HbO₂ for the minutes 2 to 8 was consistently higher than the concentration measured at the end of the recordings and higher than the concentration measured during sham stimulation. In channel L4, the other location for which the stimulation condition was significant, the time factor revealed a statistically significant difference (\( F(19,10)=3.57, \ p<0.0001 \)), with again a HbO₂ peak occurring at minutes 4 to 6. Figure 5.3C shows the mean HbO₂ and its 95% confidence interval obtained for the
real stimulation and the sham session when considering only the channels for which the stimulation effect was significant.

In a subgroup of 3 subjects, real tDCS was also delivered for 15 minutes. As expected, the hemodynamic effects of 15 min tDCS lasted longer compared to 10 min tDCS. Figure 5.4 shows the time course of HbO₂ after the 15 min tDCS.

![Figure 5.4 Temporal HbO₂ course following 15 min real tDCS.](image)

The graph shows the time course of HbO₂ following 15 min tDCS. This experiment was conducted in 3 subjects. Time courses were obtained considering only the channels L3 and L4). Longer tDCS produced longer and more stable increase of HbO₂. The thinner dotted lines represent instead the 95% confidence intervals for HbO₂.

5.2.4 Discussion

The main result of this study is that weak anodal tDCS produces a local increase of the concentration of HbO₂ in the underlying brain tissue. Moreover we demonstrated that this effect is relatively focal. In fact, the difference between the stimulation and sham condition on HbO₂ was strongly significant only for a few channels. In particular, it was
significant at the channels on the left side (L3 and L4). These channels correspond to the area stimulated with anodal tDCS, thus suggesting that mainly anodal stimulation acts on HbO₂. The presence of significant variations over time confirms that the effects of the anodal stimulation are localized in time and, with 10min tDCS, last up to 8-10 minutes (6-8 minutes from the beginning of the recording) after the end of the stimulation, with a peak effect at six minutes after the end of stimulation. Moreover, longer stimulation session (15 minutes) produced longer effect on the hemodynamic response.

This effect is unlikely to be due to a general alteration in arousal of subjects or other non-specific causes for several reasons: 1) sham stimulation would have produced a similar effect; 2) an effect of general arousal would have been expected to involve both anodal (left) and cathodal (right) sites of stimulation; 3) similar effects would have been expected at all channels and not in a focal manner. The effect must have been cerebral in origin because the stimulation settings for the tDCS have been proven to alter the cortical excitability (Marshall et al., 2005; Nitsche and Paulus, 2001); additionally, the optode spacing used in the fNIRS device ensured readings sensitive to the hemodynamic activity in the first millimeters of cortical gray matter (Chance et al., 1998; Firbank et al., 1998).

In this study we observed that cathodal stimulation has a negligible effect. This could simply be due to a different threshold of cathodal stimulation on the hemodynamic response. Lang et al. (2005) reported that both anodal and cathodal tDCS increased the metabolism of the cortex underlying the stimulation electrodes, with cathodal stimulation producing a much less effective increase of the metabolic response under the electrodes. In any case it is well known that cathodal and anodal tDCS affect cortex in a different manner (Lang et al., 2004; Nitsche et al., 2003). Another plausible explanation is that we evaluated the brain hemodynamic response at rest (the subjects were not involved in any task and were completely relaxed). It is possible that the lack of effect of cathodal
stimulation is due to a “floor effect”, and that it is not possible to further reduce the oxygen availability – determined by the hemodynamic link between neural activity and regional blood flow – in a rest condition. We cannot exclude also that there are side differences as we applied cathodal stimulation always on the right prefrontal cortex and anodal on the left prefrontal cortex. However, the experience of similar excitability findings in right and left motor cortex makes this possible difference unlikely.

In conclusion, the present results provide direct evidence that the hemodynamic response of the brain is substantially different when anodal or cathodal tDCS were used. The possibility to induce hemodynamic changes in the brain using tDCS and the possibility to monitor these effects by fNIRS can have very interesting clinical and scientific applications. Changes in the rCBF can be used as a measurable output when cortices without any measurable output are modulated by tDCS and other neuromodulation techniques. The availability of such ‘window’ to monitor the tDCS effects on the cortical neurophysiology and metabolism can promote the development of tDCS-based treatments for a variety of mental and neurological diseases. For example, anodal tDCS of the dorsolateral prefrontal cortex (DLPFC) is known to have effects on working memory. Based on this study there is evidence that anodal stimulation of DLPFC at rest increases the availability of metabolic resources. Therefore, tDCS could be investigated as a potential neurorehabilitation tool in traumatic brain injury and other disorders that impair working memory and indicate compromised metabolism of DLPFC. Moreover, monitoring of the effects on rCBF can be used to calculate the “normal” hemodynamic response to manipulation of the cortical excitability (by tDCS or other neuromodulation techniques). This would be useful, for example in chronic cerebrovascular disorders, to...
identify patients with a reduced capability to increase the blood flow on demand. The fact that tDCS can modulate focally the rCBF can be used to increase the oxygen availability or to facilitate the elimination of “neurotoxic” substances in stroke patients and in degenerative disorders. In summary, these findings open new avenues to new diagnostic and therapeutic options of tDCS and fNIRS.

5.3 Performance evaluation using fNIRS and EEG in healthy subjects

5.3.1 Introduction

The last 20 years have seen a rapid advance in neuroimaging technologies that are now widely used for non-invasive investigation of human brain functions. Application of these technologies to the fields of basic and clinical neuroscience has greatly expanded our knowledge about brain activity associated with perceptual, cognitive, emotional and behavioral processes, in health (Sergent, 1994; Stufflebeam and Rosen, 2007) and disease (Eliassen et al., 2008; Malhi and Lagopoulos, 2008). In particular, neuroimaging techniques have contributed to the investigation of the specialization and integration of different cerebral areas in the normal brain and to the study of brain dysfunction in varying disorders (Calvert, 2001; Lalanne and Lorenceau, 2004). Nonetheless, the current understanding of the relation between brain activity and behavior is still limited. One of the restricting factors is the inherent complexity of the system to be investigated. In fact, most task designs in neuroimaging aim at probing or manipulating one cognitive domain at a time, but human behavior results from the
interaction of multiple components (e.g., attention, orienting response, planning or short-term memory). Additionally, the macroscopic brain activity is a multifaceted process and the combined use of multiple neuroimaging technologies could capture different aspects of this process. Therefore, given the complexity of the investigated processes and the wide range of characteristics for the different imaging technologies, the use of multimodal approaches is gaining the interest of the scientific community (George et al., 1995; He and Lian, 2002; Ritter and Villringer, 2006; Schiff, 2006). The underlying principle is that all neuroimaging techniques provide in vivo measures of brain function but each has its own set of assets and drawbacks. Hence, the combination of multiple imaging modalities with complementing strengths can partially overcome the limitations encountered by each individual modality.

We provide here an example of multimodal imaging approach using fNIRS and EEG. The aim is evaluate the performance level of subjects during a task with high working memory load. This was pursued by using measures from fNIRS or EEG individually or in a combination, the results of which were then compared. The rationale for using these modalities among others is two-fold. First, there are indications that the oxygenation changes recorded by fNIRS in working memory tasks are related to the task load and at the same time are affected by the performance level of the subject (Izzetoglu et al., 2007). Second, EEG has been extensively applied in working memory research. In particular, many studies focused on the P300 component, a peak occurring about 300 ms after a target stimulus presentation and reflecting the demand on attentional resources (Kok, 1997). Based on its neuropsychological interpretation, the P300 amplitude is expected to increase with increasing task complexity (Johnson, 1993), but studies have shown a decline when the stimulus is objectively harder to discriminate or when the subject is less confident in its discrimination (Kok, 2001). Hence, the combined use of
fNIRS and EEG can provide insight into the different mechanisms underlying the observed low performance on a working memory task.

5.3.2 Methods

Subjects and experimental paradigm

Six subjects (3 males and 3 females) were selected from a larger pool of healthy participants. All subjects were right-handed, with vision correctible to 20/20. Participants denied any history of neurological disorders, psychiatric illness, substance abuse or being on any current medication. The experimental protocol was approved by the Institutional Review Board at Drexel University and all participants gave their informed consent. The mean age of the participants was 24.3 years (standard deviation=5.5 years).

The subjects performed a visual n-back task with four load conditions (0-back, 1-back, 2-back and 3-back) that incrementally varied the working memory load, as described in Chapter 2.

Data acquisition and processing

EEG activity was recorded from 12 Ag/AgCl electrodes placed at frontal, central, parietal and occipital locations according to the International 10-20 System (F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, and Oz). All electrodes were referenced to linked mastoid leads. Vertical and horizontal electrooculograms (VEOG and HEOG) were monitored via electrodes placed above and below the left eye, and at the left and right outer canthi, respectively. EEG signals were collected using NuAmp amplifier (Neuroscan Inc., El Paso, TX); all impedances were systematically kept below 10 kΩ and the amplification
was set to 50 mV/mm. EEG signals were filtered between 0.15 and 100 Hz and sampled at 500 samples/second.

The hemodynamic activity of the prefrontal cortex was recorded using a continuous-wave fNIRS device. Changes in light absorption were converted to changes in concentration of oxyhemoglobin (HbO₂) and deoxyhemoglobin (HHb).

Information about the behavioral performance in the task was recorded for all subjects. The percentage of correct responses (% Correct) was calculated separately for the four working memory loads and for the overall test. Out of the total pool of subjects, 3 were randomly selected from the group with an overall performance higher than the median (“high performing” group) and 3 were randomly selected from the group with an overall performance below the median (“low performing” group). Table 5.1 summarizes the behavioral performance for the overall group of subjects.

Table 5.1 Behavioral performance in n-back.

The behavioral performance of the subjects in the n-back task was evaluated in terms of % Correct, defined as the percentage of stimuli that received a correct response. Behavioral performance was assessed both globally and at the single-load level. The table reports the mean, the 95% confidence interval of the mean (LL: lower level; UL: upper level) and the median for the overall performance.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>95% C.I. of mean [LL UL]</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Correct – 0-back</td>
<td>90.07</td>
<td>[81.29  98.84]</td>
<td></td>
</tr>
<tr>
<td>% Correct – 1-back</td>
<td>92.06</td>
<td>[85.63  98.47]</td>
<td></td>
</tr>
<tr>
<td>% Correct – 2-back</td>
<td>95.33</td>
<td>[79.23  91.42]</td>
<td></td>
</tr>
<tr>
<td>% Correct – 3-back</td>
<td>78.07</td>
<td>[73.76  82.37]</td>
<td></td>
</tr>
<tr>
<td>% Correct – OVERALL</td>
<td>86.38</td>
<td>[80.48  92.28]</td>
<td>89.70</td>
</tr>
</tbody>
</table>
fNIRS recordings – fNIRS data were divided into blocks locked to the repeated presentations of the four working memory conditions. Each block lasted 70 s and a 5 s rest baseline was included. The raw data about light absorption acquired by the fNIRS device were low-pass filtered and were converted to changes in concentration of oxyhemoglobin (HbO₂) and deoxyhemoglobin (HHb) using the modified Beer-Lambert law (see Chapter 2 for additional details). The baseline condition used in the modified Beer-Lambert law was the rest period immediately preceding each block. For each of the seven presentations of the 3-back condition, the mean change in HbO₂ concentration was extracted. In particular, the channels of most interest were those monitoring the rostral portion of the superior and middle frontal gyri in the left hemisphere, since they have been previously demonstrated to be significantly activated by the n-back task (Cabeza and Nyberg, 2000; Owen et al., 2005). Therefore, for each subject, average HbO₂ values for each of the seven presentations of the 3-back condition were extracted from optodes 4, 5 and 6 and used as features in the subsequent classification. Only the 3-back condition was investigated, based on earlier evidence that the oxygenation values recorded during the 3-back condition are affected by the performance level (Izzetoglu et al., 2007).

EEG recordings – Independent component analysis was used to minimize ocular artifacts in the EEG recordings (Jung et al., 2000a; Jung et al., 2000b). Stimulus-locked event-related potentials (ERPs) were extracted from channels Cz and Pz for target stimuli presented in the 3-back condition. A 150 ms pre-stimulus baseline window and a 700 ms post-stimulus response window were used. All epochs were baseline corrected by subtracting the mean of the baseline window from the full epoch. Epochs containing significant movement or muscle artifacts were discarded. The P300 peak was automatically identified at each of the two channels as the largest positive deflection in
the 250-600 ms post-stimulus response. For each subject, the average amplitude values of the P300 peak at Cz and Pz were obtained for each of the seven presentations of the 3-back condition and used as features in the classification.

Classification – The classification between “high performing” and “low performing” subjects was performed using five different features: two features were obtained from the EEG recordings (the amplitude of the P300 peak at channels Cz and Pz) and three were obtained from the fNIRS recordings (mean change in HbO₂ concentration at optodes 4, 5 and 6).

For each subject multiple instances of these features were extracted, one for each presentation of the 3-back condition. Each instance \(x_i\) was associated with a label \(y_i\) that stated the group of the subject from which the instance was collected (\(y_i = \text{“low performing”}\) or \(y_i = \text{“high performing”}\)). The total number of instances \(x_i\) was 38 (7 blocks presented to 5 subjects + 3 blocks presented to 1 subject) and constituted the overall set \(S\) of available instances: \(S = [x_i, y_i]\).

Four different approaches were evaluated to determine their ability to identify “high performing” or “low performing” individuals:

1. EEG-based classification: only the features extracted from EEG recordings were used; the feature vector consisted of two elements: the P300 amplitude at Cz and Pz.
2. fNIRS-based classification: only the features extracted from fNIRS recordings were used; the feature vector consisted of three elements: the mean change in HbO₂ concentration at optodes 4, 5 and 6.
3. Feature-level fusion: features extracted from both EEG and fNIRS recordings were used and combined in a single feature vector of five elements.
4. Decision-level fusion: the classification was performed separately using the EEG and fNIRS features, whose results were then combined to reach the final decision.

Two types of classifiers were investigated to be used in the above-mentioned approaches: the Mahalanobis discriminant (MD) and the quadratic classifier (QC). The MD is equivalent to the optimum Bayes classifier if the data are normally distributed with identical (although arbitrary) covariance matrices for all classes. In QC, instances are labeled using a Bayesian error minimization approach, under the more general hypothesis that the covariance matrices for all classes can assume any arbitrary value (Duda et al., 2001).

For training and testing, a modified k-fold (k=5) cross-validation was implemented. In such an approach, the set S is partitioned into k blocks, each one representing the two groups (“low performing” and “high performing”) in an approximately balanced way. Each of the k blocks was in turn held out for testing (S(k)), while the other k-1 blocks (S(k-1)) were used for training using a bagging procedure (Kuncheva, 2004). In bagging, an ensemble of classifiers is created: in this study the ensemble was comprised of 10 classifiers all sharing the same architecture but trained on different randomly generated subsets (rS^{(r-1)}, r=1,2,...,10) of S^{(k-1)}. A label \hat{y}_{(r)} is assigned to each instance in the testing set S(k) by each of the 10 classifiers, which are then combined using a majority voting decision rule. In our implementation, this entire process – of generating 10 training subsets and training 10 corresponding classifiers – was repeated 5 times, each time holding out a different subset S^{(k)} for testing. For each of these 5 repetitions, the accuracy, defined as the probability of correctly classifying an instance, was calculated.
### 5.3.3 Results

Table 5.2 summarizes the behavioral performance in the four n-back conditions for the 3 subjects in the “high performing” group and for the 3 subjects in the “low performing” group. The difference in behavioral performance between the two groups is evident in the overall percentage of correctly identified stimuli and in each of the three n-back conditions.

The distribution, in the features space, of the instances collected from the two groups of individuals is presented in Figure 5.5A and Figure 5.5B. These figures show the features extracted respectively from the EEG recordings (P300 amplitude at Pz and Cz) and from the fNIRS recordings (change in HbO\textsubscript{2} concentration at channels 4, 5 and 6); in both spaces the two classes are substantially overlapping. A separate analysis confirmed also a dissociation in the HbO\textsubscript{2} values during the 3-back condition between the two groups: whereas for the “high performing” group the HbO\textsubscript{2} values were increasing with the working memory load (1-back condition: 0.0086 mM; 2-back condition: 0.0096 mM; 3-back condition: 0.0216 mM), for the “low performing” group this relation was lost and the 3-back condition saw a decrease in the HbO\textsubscript{2} values (1-back condition: -0.0213 mM; 2-back condition: 0.0263 mM; 3-back condition: -0.0245 mM).

#### Table 5.2 Behavioral performance in ‘Low perfoming’ and ‘High performing’ groups.

The behavioral performance of the subjects in the n-back task was evaluated in terms of % Correct separately for the ‘Low performing’ and ‘High performing’ groups. Behavioral performance was assessed both globally and at the single-load level. The table reports the mean of the % Correct.

<table>
<thead>
<tr>
<th></th>
<th>Low performing</th>
<th>High performing</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Correct – 0-back</td>
<td>74.95</td>
<td>94.87</td>
</tr>
<tr>
<td>% Correct – 1-back</td>
<td>80.83</td>
<td>95.22</td>
</tr>
<tr>
<td>% Correct – 2-back</td>
<td>77.87</td>
<td>88.38</td>
</tr>
<tr>
<td>% Correct – 3-back</td>
<td>70.56</td>
<td>82.03</td>
</tr>
<tr>
<td>% Correct – OVERALL</td>
<td>76.05</td>
<td>90.12</td>
</tr>
</tbody>
</table>
Figure 5.5 EEG and fNIRS features ‘Low performing’ and ‘High performing’ groups. (A) Scatterplot of instances of EEG-related features, the P300 amplitude at Pz and the P300 amplitude at Cz. (B) Scatterplot of instances of fNIRS-related features, the mean changes in HbO\textsubscript{2} concentration at optodes 4, 5 and 6 (ΔHbO\textsubscript{2} chn 4, ΔHbO\textsubscript{2} chn 5 and ΔHbO\textsubscript{2} chn 6, respectively). In both part A and B, instances are categorized as belonging to ‘High performing’ (filled triangle) or ‘Low performing’ (empty circles) individuals. From each individual, multiple instances were collected, one for each presentation of the 3-back condition.

Table 5.3 presents the accuracy obtained by the different classification strategies using each of the two classifiers (QD and MD) as base classifiers. In general, the results showed an enhancement in classification performance when features from both EEG and fNIRS are used compared to the results obtained when using them separately. In fact, the mean accuracy for the Feature-level fusion and Fecision-level fusion strategies were overall higher than the mean accuracy for the EEG-based and fNIRS-based classifications.

Additionally, the two fusion approaches (and in particular the Fecision-level fusion, in agreement with (Parikh and Polikar, 2007)) provided an increase, albeit small, in the generalizing ability of the classifiers, as measured by a decrement in the accuracy standard deviation.
Table 5.3 Classification accuracy for different classification approaches.

Four classification approaches were investigated: “fNIRS-based classification”, “EEG-based classification”, “Feature-level fusion classification” and “Decision-level fusion classification”. The table reports the mean and standard deviation for each of them, separately for the two classifiers that were examined: quadratic classifier (QC) and Mahalanobis discriminant (MD).

<table>
<thead>
<tr>
<th>Classification Approach</th>
<th>QC</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>fNIRS-based classification</td>
<td>62.53 ± 21.78</td>
<td>51.74 ± 7.41</td>
</tr>
<tr>
<td>EEG-based classification</td>
<td>60.31 ± 18.27</td>
<td>62.53 ± 16.44</td>
</tr>
<tr>
<td>Feature-level fusion classification</td>
<td>65.39 ± 16.51</td>
<td>58.09 ± 18.56</td>
</tr>
<tr>
<td>Decision-level fusion classification</td>
<td>71.11 ± 10.67</td>
<td>71.11 ± 10.67</td>
</tr>
</tbody>
</table>

5.3.4 Discussion

Overall, the classification of instances collected from “high performing” and “low performing” subjects benefited from the use of combined fNIRS and EEG features. We acknowledge that the results of our statistical analyses cannot be considered conclusive at this time due to the limited data of 3 subjects (from each class) that were available to us. A larger pool of subjects, and therefore a higher number of instances available for training and testing, would allow a better estimation of the accuracy index. Nonetheless, the combination of fNIRS and EEG improved classification accuracy, even if two relatively simple classifiers were used: a linear parametric classifier (MD) and a nonlinear parametric classifier (QC). It is reasonable to expect that higher accuracies can be obtained using more sophisticated nonlinear classifiers (such as neural networks or support vector machines), which are not bounded by assumptions about the features distributions. Similarly, the relative performance of other fusion algorithms could be investigated, ranging from the simple majority voting (presented in this paper) to multinomial methods, to the fusion of discriminant scores.
We have investigated the feasibility and performance of fNIRS and EEG data fusion for the evaluation of the behavioral performance of six healthy adults in a working memory task. Although fNIRS and EEG have been co-registered in previous studies (Butti et al., 2006; Kennan et al., 2002), this is the first attempt at their integration by using them together in a pattern recognition application, together with that presented in Chapter 4. For this study, the fNIRS-EEG fusion took advantage of both the spatial information about the hemodynamic activity (by including only the channels monitoring the rostral portion of the superior and middle frontal gyri in the left hemisphere) and the fast temporal dynamic of the cognitive processes of interest (by including information about the P300 amplitude). A similar approach can also help explain the mechanisms underlying low task performance in case of neurological disorders such as traumatic brain injury or multiple sclerosis, hence providing some physiological evidence important for the choice of a proper neurorehabilitation and pharmacological intervention. fNIRS-EEG fusion may further be applied to the study of other cognitive domains, in particular taking advantage of the flexibility in task designs allowed by fNIRS and EEG.
CHAPTER 6: CONCLUSION

The aim of this thesis was to investigate the applicability of functional near infrared spectroscopy (fNIRS), and its integration with electroencephalography (EEG), to assessment of working memory after traumatic brain injury (TBI).

The results from the research presented in this thesis provide the first evidence of the ability of fNIRS to reveal differences between the TBI and healthy subjects in the working memory tasks. These differences suggest a dysfunction that reflects a mismatch between cognitive demands and cortical resources. This suggests that this dysfunction underlies working memory impairments in TBI subjects. Moreover, this study has demonstrated that fNIRS measures can distinguish between the two groups and these data have been investigated relative to the results obtained by EEG alone or by the ‘composite measure’ based on the combination of these two technologies. Based on this investigation, each of the two modalities revealed unique strengths that can contribute to the classification between the two groups. Therefore the successful combination of fNIRS and EEG in the context of this application takes advantage of synergistic strengths between the two modalities in order to improve the classification between healthy and TBI cases.

Studies undertaken in this thesis proved the potential of fNIRS individually and of the composite fNIRS/EEG measure to constitute the basis of clinical indices to
evaluate the efficacy of neurorehabilitation interventions. Building on these steps, future work may need to address the following points:

1- demonstration of the proposed fNIRS or the combined fNIRS/EEG approach in the characterization of the neurophysiological underpinnings underlying the process of recovery from TBI;

2- identification or design of interventions that target those neurophysiological processes;

3- demonstration of the proposed measure in the monitoring or guidance of the interventions.

A similar approach could also been applied to other TBI-related cognitive impairments that imply dysfunction of the prefrontal cortex. Potential applications to a variety of clinical populations that suffer from similar cognitive disabilities, such as multiple sclerosis, could also be anticipated.
APPENDIX A: CLASSIFICATION ALGORITHMS

A.1 fNIRS-based classification

A total of 32 fNIRS-related features was used to classify between healthy (HC) and TBI subjects. The features consisted in the maximum value $H_{max}$ for the three hemodynamic variables ($HbO_2$, $HHb$ and $Hb_{TOT}$) and the maximum ratio $MR$ in the four load conditions (0-back, 1-back, 2-back and 3-back) at the two regions of interest (left and right DLPFC) ($3 \times 4 \times 2$ for $H_{max} + 4 \times 2$ for $MR = 32$ features = $M_{fNIRS}$). For each subject, multiple (=7) instances of these features were extracted, one for each presentation of a given load condition. Each instance ($x_i$) was associated with a label that stated the group of the subject from which the instance was collected ($y_i$ = “TBI group” or $y_i$ = “HC group”). The available instances constituted the overall set $S = \{(x_i, y_i)\}$.

The selection of relevant features was performed using the random subspace method (RSM) (Ho, 1998). Based on the RSM, instead of a single classifier, the decision about the label to assign to a given instance $x_i$ is taken by a random subspace method ensemble of $T$ classifiers. Each $t$-th classifier in the ensemble assigns a label $\hat{y}^{fNIRS(t)}$ to the instance $x_i$ based on a subset of features of predefined size $d$ and randomly chosen out of the $M_{fNIRS}$ (=32) available features ($d$<$M_{fNIRS}$) (Figure A.1). If $T$ is large enough (i.e. if the number of classifiers in the ensemble is large enough), all features are represented (with high probability) in the ensemble and in a variety of combinations. Though some combinations might have limited generalization performances, the large number of
classifiers contributing to the selection reduced the risk of misclassification. For this application $T=1000$ and $d=8$ were chosen. The combination of the decisions (labels) of the single $t$-th classifier was performed using a weighted majority-voting rule: the weight of the single classifier was assigned based on its performance on the training data.

![Diagram](image)

**Figure A.1** Random subspace method applied to fNIRS-based classification.

The label $\hat{y}_{\text{fNIRS}}^{(i)}$ for the test data point $x_i$ is predicted based on a weighted majority voting of $T$ classifiers that comprise the ensemble. Each of the $T$ classifiers is trained on a subset of $d$ random features sampled from the $M_{\text{fNIRS}} (=32)$ available ones, producing a different decision boundary $g^{(t)}_{\text{fNIRS}}(x)$. The overall decision boundary for the ensemble is $g_{\text{fNIRS}}^{(i)}(x)$ and is used to predict the label $\hat{y}_{\text{fNIRS}}^{(i)}$ for the test data point $x_i$.

Training and testing were performed using a modified leave-one-out (mLOO) cross-validation (Merzagora et al., 2009a), which incorporates bootstrapping in the well-known leave-one-out technique (Figure A.2). One instance $(x_i, y_i)$ from the available $N$ instances in the set $S$ was removed to be used as a test data point. The remaining $N-1$ instances formed the subset $S^{(i)}$. From $S^{(i)}$, 50 instances were randomly selected (with replacement) to serve as training data – 25 representing the “TBI group” and 25 representing the “HC
group” – forming the training subset $\text{TS}^{(i)}$. One ensemble of classifiers was trained on this training dataset. This process was repeated 10 times, in each case randomly choosing a different set of 50 instances out of the $N$-1 available, creating 10 training sets $\text{TS}^{(r)}$, $r=1,2,…10$ and corresponding 10 ensemble of classifiers with slightly different decision boundaries $\text{g}_{\text{fNIRS}}^{(i)}$, $r=1,2,…10$. These 10 ensembles of classifiers were then evaluated on the one test data point $(x_i, y_i)$ that was previously left out. This entire process – generating 10 training data subsets of 50 instances and training 10 corresponding ensembles – was repeated a total of $N$ times, once for each data point to be used as test data. Hence, the $N$ available instances in the set $S$ were each classified 10 times using the mLOO, allowing a statistical characterization of performance indices.

The investigated classifiers ($C$) were two: support vector machines and decision trees. Support vector machine (SVM): Support vector machines are binary classifiers that use a non-linear mapping kernel function to transform the given data into a higher dimensional space, where the data is believed to be linearly separable. Classification is then performed in the new space by finding the optimal hyperplane that offers maximum separating margin between the closest samples of the two classes (Boser et al., 1992; Scholkopf et al., 1995). The performance of a given SVM depends also on a trade-off parameter $c$: the $c$ parameter balances the relative importance of minimizing the training error and maximizing the margins between the classes, which directly affects the classifier’s generalization ability. In this work a Gaussian radial basis function was used as the kernel. The spread $\sigma$ of the kernel and the $c$ parameter were selected using a $k$-fold cross-validation approach ($k=3$ was used in this study). In such approach, the dataset is
Figure A.2 Pseudo-code for modified leave-one-out (mLOO) algorithm.

Each instance \((x_i, y_i)\) was in turn removed to be used as a test data point. From the remaining instances, \(R\) random subsets were then created to be used for training of \(R\) different ensembles. Each ensemble was comprised of \(T\) classifiers, each trained on a different subset of features. The label \(\hat{y}_{\text{NIRS}}^{(j)}\) predicted for the test data point \(x_i\) was obtained through weighted majority voting among the \(T\) classifiers in the ensemble and the \(R\) repetitions were used to obtain a bootstrapped estimation of the classification performances.
partitioned into \( k \) blocks; multiple SVMs with different values of \( \sigma \) and \( c \) are trained on \( k-1 \) blocks and tested on the remaining \( k \)-th block. The values of \( \sigma \) and \( c \) that provide the best performance over \( k \) trials is then chosen. \emph{Decision tree (DT)}: Decision trees are classification methods that use a non-metric approach, i.e. they are not based on any concept of distance between feature vectors. An instance \( x_i \) is classified by testing its describing features with a series of mutually distinct and exhaustive questions or queries on their values (Brodley and Utgoff, 1995).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure_a3}
\caption{SVM \( k \)-fold calibration for fNIRS-based classification.}
\end{figure}

Isolevel plots of accuracy, sensitivity and specificity obtained by SVM in a \( k \)-fold validation with varying values of kernel \( \sigma \) and margin-error trade-off \( c \) for the fNIRS-based classification. The results of the \( k \)-fold validation were used to set the two free parameters in the classification algorithm. In order to optimize the classification performances of the SVM, \( \sigma=3 \) and \( c=0.1 \) were chosen.
The results of the \( k \)-fold cross-validation approach used to select the kernel spread \( \sigma \) and margin-error trade-off \( c \) for the SVM are presented in Figure A.3. Based on this preliminary information, the Gaussian kernel spread \( \sigma \) was set to 3 and the margin-error trade-off parameter \( c \) was set to 0.1. This choice allowed for a balanced classification performance: it offered good sensitivity, though without sacrificing accuracy and specificity. In fact, although smaller values for the kernel spread would increase accuracy and specificity, the values of sensitivity would be greatly affected.

A.2 EEG-based classification

For the classification between HC and TBI group based on EEG, a total of 4 EEG-related features was used to classify between healthy (HC) and TBI subjects (\( M_{EEG}=4 \)). The features consisted in the P300 peak amplitude elicited by true matches in the four load conditions (0-back, 1-back, 2-back and 3-back). For each subject, multiple (=7) instances of these features were extracted, one for each presentation of a given load condition. Similarly to the fNIRS-based classification, each instance \((x_i)\) was associated with a label that stated the group of the subject from which the instance was collected \((y_i = \text{“TBI group”} \text{ or } y_i = \text{“HC group”})\). The available instances constituted the overall set \( S = [(x_i , y_i)] \).

In this case a bagging approach was applied (Breiman, 1996), where the decision about the label to assign to a given instance \( x_i \) is taken by an ensemble of \( B \) classifiers: each \( b \)-th classifier in the ensemble assigns a label to the instance \( x_i \) based on a bootstrapped
replica $\text{TS}_{(r)}^{(i_b)}$ of the training subset (Figure A.4 and Figure A.5). $\text{TS}_{(r)}^{(i_b)}$ is created by drawing (with replacement) $q$ random instances from the entire training set $\text{TS}_{(r)}$. For this application $B=100$ and $q=25$. The final label $\hat{y}_{\text{EEG}}^{(i)}$ assigned to the instance $x_i$ was the combination of the decisions (labels) of the single $b$-th classifiers; the combination was performed using a weighted majority-voting rule: the weight of the single classifier was assigned based on its performance on the training data.

Similarly to what described for the fNIRS-classification, the parameters for the SVM classifier were selected based on the results of a $k$-fold validation. Based on these results (Figure A.6), $\sigma=3$ and $c=0.1$ were chosen.

**Figure A.4** Bagging applied to EEG-based classification.

The label $\hat{y}_{\text{EEG}}^{(i)}$ for the test data point $x_i$ is predicted based on a weighted majority voting of $B$ classifiers that comprise the ensemble. Each of the $B$ classifiers is trained on a subset of $q$ random instances sampled from the ones in the training set $\text{TS}_{\text{EEG}}^{(i)}$, producing a different decision boundary $g_{\text{EEG}}^{(i_b)}$. The overall decision boundary for the ensemble is $g_{\text{EEG}}^{(i)}$ and is used to predict the label $\hat{y}_{\text{EEG}}^{(i)}$ for the test data point $x_i$. 
**Input:**
- Set of instances $S = \{(x_i, y_i), \ i=1,2,...,N\}$ with correct labels $y_i \in \{"TBI group", "HC group"\}$. Each instance $x_i$ is represented by a $(1\times MEEG)$ vector containing $MEEG$ features.
- Classification algorithm $C$.
- Integer $B$ specifying the number of classifiers in the bagging ensemble.
- Integer $q$ specifying the number of features to be randomly selected out of the $MEEG$ available features.
- Integer $R$ representing the number of bootstrapping iterations.

**Procedure:**
1. **Do** for each $i=1,2,...,N$:
   a. Obtain a subset $S^{(i)}$ by removing the instance $(x_i, y_i)$ from set $S$:
      $$S^{(i)} = S - (x_i, y_i)$$
   b. **Do** for $r=1,2,...,R$:
      i. Obtain a training subset $TS^{(r)}$ by drawing 25 TBI instances and 25 HC instances from $S^{(i)}$.
      ii. **Do** for $b=1,2,...,B$:
         1. Create a training subset $TS^{(r,b)}$ by resampling $q$ instances from $TS^{(r)}$.
         2. Train classifier $C$ on $TS^{(r,b)}$, obtaining decision boundary $g_{EEG}^{(r,b)}$; the accuracy of $C$ on $TS^{(r,b)}$ is assigned as weight $w(b)$.
         3. Based on $g_{EEG}^{(r,b)}$, determine the predicted label $\hat{y}_{EEG}^{(r,b)}$ for instance $(x_i, y_i)$.
   c. **Create** the matrix $\hat{Y}$ of all predicted labels $\hat{y}_{EEG}^{(i)}$ for all instances in $S$:
      $$\hat{Y} = \begin{bmatrix}
      \hat{y}_{EEG}^{(1)} & \cdots & \hat{y}_{EEG}^{(N)} \\
      \vdots & \ddots & \vdots \\
      \hat{y}_{EEG}^{(R)} & \cdots & \hat{y}_{EEG}^{(N)}
      \end{bmatrix}$$
   d. **Compute** the performance indices for each row of $\hat{Y}$.

**Output:** Statistical characterization of performance indices.

**Figure A.5** Pseudo-code for mLOO algorithm for EEG-based classification. Each instance $(x_i, y_i)$ was in turn removed to be used as a test data point. From the remaining instances, $R$ random subsets were then created to be used for training of $R$ different ensembles. Each ensemble was comprised of $B$ classifiers, each trained on a different subset of instances. The label $\hat{y}_{EEG}^{(i)}$ predicted for the test data point $x_i$ was obtained through weighted majority voting among the $B$ classifiers in the ensemble and the $R$ repetitions were used to obtain a bootstrapped estimation of the classification performances.
Figure A.6 SVM k-fold calibration for EEG-based classification. Isolevel plots of accuracy, sensitivity and specificity obtained by SVM in a k-fold validation with varying values of kernel $\sigma$ and margin-error trade-off $c$ for the EEG-based classification. The results of the k-fold validation were used to set the two free parameters in the classification algorithm. In order to optimize the classification performances of the SVM, $\sigma=3$ and $c=0.1$ were chosen.
A.3 Feature-level fusion classification

In this approach the fNIRS-related features and the EEG-related features were combined in one single feature vector. Thus a total of 36 features were used: 32 from the fNIRS recordings and 4 from the EEG recordings \((M = M_{fNIRS} + M_{EEG} = 36)\). A RSM approach was used to create an ensemble of classifiers (based on different combination of the 36 features) that assigned a label \(\hat{y}_{FL}^{(i)}\) to each instance \(x_i\) (Figure A.7 and Figure A.8).

**Figure A.7** Random subspace method applied to feature-level fusion classification. The label \(\hat{y}_{FL}^{(i)}\) for the test data point \(x_i\) is predicted based on a weighted majority voting of \(T\) classifiers that comprise the ensemble. Each of the \(T\) classifiers is trained on a subset of \(d\) random features sampled from the \(M(=M_{fNIRS}+M_{EEG})\) available ones, producing a different decision boundary \(g_{FL}^{(r)(i)}\). The overall decision boundary for the ensemble is \(g_{FL}^{(r)}\) and is used to predict the label \(\hat{y}_{FL}^{(i)}\) for the test data point \(x_i\).
Figure A.8 Pseudo-code for mLOO algorithm for feature-level fusion classification. Each instance \((x_i, y_i)\) was in turn removed to be used as a test data point. From the remaining instances, \(R\) random subsets were then created to be used for training of \(R\) different ensembles. Each ensemble was comprised of \(T\) classifiers, each trained on a different subset of features. The label \(\hat{y}_{FL(i)}\) predicted for the test data point \(x_i\) was obtained through weighted majority voting among the \(T\) classifiers in the ensemble and the \(R\) repetitions were used to obtain a bootstrapped estimation of the classification performances.
A.4 Decision-level fusion classification

For each instance $x_i$, the final decision on the label to assign to it (“TBI group” or “HC group”) was based on an ensemble of $B+T$ classifiers ($B=100$, $T=1000$). While $B$ classifiers assigned their labels $\hat{y}^{EEG}_{i(b)}$ (with $b=1,2,\ldots,B$) based on EEG-related features following the bagging approach previously described, $T$ classifiers assigned their labels $\hat{y}^{fNIRS}_{i(t)}$ (with $t=1,2,\ldots,T$) based on fNIRS-related features following the RSM approach (Figure A.9 and Figure A.10). The final label $\hat{y}_{DL(i)}$ assigned to the instance $x_i$ was the combination of the decisions of the single classifiers in the ensemble; the combination was performed using a weighted majority-voting rule: the weights of the single classifiers were such that EEG-based labels and fNIRS-based labels were equally weighted.

**Figure A.9** Random subspace method and bagging for decision-level fusion.

The label $\hat{y}_{DL(i)}$ for the test data point $x_i$ is predicted based on a weighted majority voting of $B+T$ classifiers that comprise the ensemble. Each of the $B$ classifiers is trained on a subset of $q$ random instances sampled from the ones in the training subset and each of the $T$ classifiers is trained on a subset of $d$ random features sampled form the $M_{fNIRS}$ available ones. The overall decision boundary for the ensemble is $g_{DL(i)}$ and is used to predict the label $\hat{y}_{DL(i)}$ for the test data point $x_i$. 
**Input:**
- Set of instances $S = [(x_i, y_i)]$, $i = 1, 2, \ldots, N$, with correct labels $y_i \in \{"TBI group", \"HC group\}\}. Each instance $x_i$ is represented by a $(1 \times M\text{NIRS})$ vector containing $M\text{NIRS}$ features and by a $(1 \times M\text{EEG})$ vector containing $M\text{EEG}$ features.
- Classification algorithm $C$.
- Integer $T$ specifying the number of classifiers in the random subspace method ensemble.
- Integer $d$ specifying the number of features to be randomly selected out of the $M\text{NIRS}$ available features.
- Integer $B$ specifying the number of classifiers in the bagging ensemble.
- Integer $q$ specifying the number of features to be randomly selected out of the $M\text{EEG}$ available features.
- Integer $R$ representing the number of bootstrapping iterations.

**Procedure:**
1. **Do** for each $i = 1, 2, \ldots, N$
   - Obtain a subset $S^{(0)}$ by removing the instance $(x_i, y_i)$ from set $S$:
     $$S^{(0)} = S - (x_i, y_i)$$
   - **Do** for $r = 1, 2, \ldots, R$
     - Obtain a training subset $TS^{(r)}_i$ by drawing 25 TBI instances and 25 HC instances from $S^{(0)}$.
     - **Do** for $t = 1, 2, \ldots, T$
       - Create a training subset $TS^{(r)}_{(t)}$ by retaining $d$ random features (of the $M\text{NIRS}$ available ones) for each of the instances in $TS^{(r)}_i$.
       - Train classifier $C$ on $TS^{(r)}_{(t)}$, obtaining decision boundary $g^{M\text{NIRS}}_{(r)}(\cdot)$.
       - Based on $g^{M\text{NIRS}}_{(r)}$, determine the predicted label $\hat{y}^{M\text{NIRS}}_{(r)}$ for instance $(x_i, y_i)$.
     - **Do** for $b = 1, 2, \ldots, B$
       - Create a training subset $TS^{(r)}_{(b)}$ by resampling $q$ instances from $TS^{(r)}_i$.
       - Train classifier $C$ on $TS^{(r)}_{(b)}$, obtaining decision boundary $g^{M\text{EEG}}_{(r)}(\cdot)$.
       - Based on $g^{M\text{EEG}}_{(r)}$, determine the predicted label $\hat{y}^{M\text{EEG}}_{(r)}$ for instance $(x_i, y_i)$.
   - Determine the predicted label $\hat{y}^{(r)}_i$ using a weighted majority voting rule (so that labels assigned by EEG-based classifiers and labels assigned by fNIRS-based classifiers are equally weighted).
   - **Create** the matrix $\hat{Y}$ of all predicted labels $\hat{y}^{(r)}_i$ for all instances in $S$.
   - **Compute** the performance indices for each row of $\hat{Y}$.

**Output:** Statistical characterization of performance indices.

---

**Figure A.10** Pseudo-code for mLOO algorithm for decision-level fusion classification. Each instance $(x_i, y_i)$ was in turn removed to be used as a test data point. From the remaining instances, $R$ random subsets were then created to be used for training of $R$ different ensembles. Each ensemble was comprised of $B$ classifiers trained on a different subset of instances and $T$ classifiers trained on different subsets of features. The label $\hat{y}^{(r)}_i$ predicted for the test data point $x_i$ was obtained through weighted majority voting among the $B+T$ classifiers in the ensemble and the $R$ repetitions were used to obtain a bootstrapped estimation of the classification performances.


Ferrucci, R., Mame...
rate correlates of abstract reasoning and attention studied with positron emission tomography. Intelligence 12, 199-217.


VITA

Anna Caterina Merzagora

Education
Institution: Politecnico di Milano, Milan (Italy); Degree: B.S./M.S.; Year: 2004; Field: Biomedical engineering

Positions and Employment
Graduate Assistant and Research Assistant, School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia PA (USA).

Other Experience

Teaching Assistant for: Biosimulation I; Biosimulation II; Body Synthetic; Biomedical Signal Processing; Advanced Signal Processing; Senior Design I; Senior Design II; Product Development for Medical Applications; Medical Sciences II.

Visiting researcher at the Fundación Hospital Nacional de Parapléjicos para la Investigación y la Integración in Toledo (Spain) under the supervision of Drs. Guglielmo Foffani and Antonio Oliviero (April/May 2008).

Visiting researcher at the Center for Brain Injury Rehabilitation in Sevilla (Spain) under the supervision of Dr. José León-Carrión (May 2009).

Honors
Teaching Assistant Excellence Award for the School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia PA (USA) (2005).
Research Assistant Excellence Award for the School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia PA (USA) (2008).

Selected publications

Patents
 Provisional patent application filed on February 26th, 2010 (application number 61308426).