UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Using Functional Near-Infrared Spectroscopy to Detect a Fear of Heights Response to a Virtual Reality Environment

Luciënne Angela de With

M.Sc. Thesis
November 2020

Supervisors:
dr. M. Poel
dr. N. Thammasan
prof. dr. D.K.J. Heylen

Human Media Interaction Group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

Over the past decades, virtual reality (VR) technology has gained significant popularity and interest, both in research as well as on the consumer market. One promising application area of VR is virtual reality exposure therapy (VRET), which treats anxiety disorders by gradually exposing the patient to his/her fear using VR. To make VRET safe and effective, it is important to monitor the patient's fear levels during the exposure. Non-invasive neuroimaging can be used to unobtrusively detect fear responses, among which functional near-infrared spectroscopy (fNIRS) technology exhibits the greatest potential for a combination with VR, due to its comparably low susceptibility to motion artifacts. This thesis aims to investigate to what extent the fNIRS signals captured from people with a fear of heights response and people without a fear of heights response during VR exposure differ, and to what extent a person's fear of heights response to a VR environment can be detected using fNIRS data.

Only a very limited amount of work has investigated how fear responses are reflected in fNIRS signals. Furthermore, no previous work on the automatic detection of fear responses using fNIRS data exists. The literature indicates that a combination of VR and fNIRS technology is feasible and that it allows for experiments with greater ecological validity than traditional lab experiments.

An experiment was conducted during which participants with moderate fear of heights (experimental group, $n_e = 14$) and participants with no to little fear of heights (control group, $n_c = 15$) were exposed to VR scenarios involving heights (height condition) and no heights (ground condition). During the experiment, the participants' fNIRS signals were recorded. As an additional measurement, the heart rate (HR) of every participant was extracted from the fNIRS signals. Permutation tests were used to perform between-group statistical analyses and within-group statistical analyses (for the experimental group) on the fNIRS data and HR data. Furthermore, Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) were used to train and test subject-dependent classifiers and subject-independent classifiers on the data of the significant fNIRS channels of the experimental group, in order to detect fear responses.

The between-group statistical analyses show that the fNIRS data of the control group and the experimental group are only significantly different in channel 3, where the grand average Δ[HbO] contrast signal of the experimental group exceeds that of the control group. Furthermore, the HR data of both groups are not significantly different. The within-group statistical analyses show that there are significant differences between the grand average Δ[HbO] values during fear responses and those during no-fear responses, where the Δ[HbO] values of the fear responses were significantly higher than those of the no-fear responses in the channels located towards the frontal part of the pre-frontal cortex. Also, channel 23 was found to be significant for the grand average Δ[HbR] signals. No significant differences were found between the HR data during fear responses or no fear responses of the experimental group. The subject-dependent SVM classifier using 1-second history of the fNIRS signals can detect fear responses at an average accuracy of 72.47% (SD 20.61). The subject-independent SVM classifier using 5-second history of the fNIRS signals can detect fear responses at an average accuracy of 77.29% (SD 10.64). The subject-independent classifiers show potential for usage in online detection scenarios, as they can be trained beforehand on existing fNIRS data and can classify the unseen data of a new person at an average accuracy above 75%.

Acknowledgements

There are some people to whom I would like to express my gratitude for their help throughout this thesis research project. First of all, I would like to thank the members of the supervising committee, Mannes Poel, Nattapong Thammasan, and Dirk Heylen. Thank you for your help, suggestions, and feedback.

Furthermore, I would like to thank the people from the BMS Lab of the University of Twente. Thank you for providing me with a lab space and the required materials to do the experiments. I would like to thank Tenzing Dolmans in particular, for explaining to me how to use the fNIRS hardware and for thinking along with my project.

Of course, I would also like to thank all the 41 people who took the time to participate in my experiment. Without your voluntary participation, I would not have been able to perform this specific research. Next to the participants, I am also very thankful for the help of the people who asked their friends and family to participate in my experiment.

Last but not least, I would like to thank my family and Joep. Thank you for your support and the motivational words whenever I needed it.

List of Figures

List of Tables

List of Acronyms

AQ	Acrophobia Questionnaire
BPM	Beats per minute
BVP	Blood volume pulse
CCN	Cognitive Control Network
dlPFC	Dorsolateral prefrontal cortex
DPF	Differential pathlength factor
DoF	Degrees of freedom
EEG	Electroencephalography
FDR	False discovery rate
fMRI	Functional magnetic resonance imaging
fNIRS	Functional near-infrared spectroscopy
GSR	Galvanic skin response
HbO	Oxygenated hemoglobin
HbR	Deoxygenated hemoglobin
HMD	Head-mounted display
HR	Heart rate
HRV	Heart rate variability
IPQ	IGroup Presence Questionnaire
LDA	Linear Discriminant Analysis
MBLL	Modified Beer-Lambert law
MEG	Magnetoencephalography
NI	Near-infrared
OD	Optical density
OFC	Orbitofrontal cortex
PCA	Principal component analysis
PFC	Prefrontal cortex

RT	Reaction time
SFG	Superior frontal gyrus
SUDS	Subjective Units of Distress Scale
SVM	Support Vector Machine
TDDR	Temporal Derivative Distribution Repair
TPJ	Temporoparietal junction
VE	Virtual environment
vlPFC	Ventrolateral prefrontal cortex
VHI	Visual height intolerance
VR	Virtual reality
VRET	Virtual reality exposure therapy

Contents

Chapter 1

Introduction

This chapter provides an introduction to this thesis research. First, the motivation behind the research is described. Then, the problem statement will be given, including the goals of this research and the research questions. This chapter ends with an outline of the contents of this report.

1.1 Motivation

Over the past decades, virtual reality (VR) technology has gained significant popularity and interest, both in research as well as on the consumer market [1–3]. With the recent advances made in hardware and computer graphics, VR has become more and more realistic and accessible [4]. The increase in realism and accessibility also increased VR's application to certain use cases, including education, training, anxiety therapy, physical therapy, games, entertainment, and pain management [1–10]. Such realistic virtual circumstances can have a significant influence on a person's mental state [8, 11], for example causing mental workload, stress, or feelings of fear.

One promising application area of VR is virtual reality exposure therapy (VRET), a form of therapy that stems from traditional exposure therapy. Exposure therapy treats anxiety disorders by gradually and repeatedly exposing the client to his/her fear [12]. Exposure to fear in the absence of harm activates the fear extinction process, which explains why exposure therapy is an effective intervention [13]. The added value of VRET is that the exposure happens in the virtual world, which makes the exposure setting more controlled, safer, and in some cases also less expensive than traditional exposure therapy [5, 14, 15]. Furthermore, the exposure protocol can be completely standardized when using VRET, which increases the therapist's control over the stimuli and the duration of the exposure, as opposed to traditional in vivo exposure [16]. Despite the greater amount of control that VRET offers to the therapist, it is still common practice that the therapist monitors the fear responses of the client [12]. One important reason to do this is to ensure that the gradual exposure to the fear-eliciting stimuli do not overwhelm the client. Exposure to situations that induce too much fear can, for example, cause panic attacks for the client and might therefore worsen his/her anxiety, instead of treating it [14].

1.2 Problem Statement

Monitoring a person's fear responses whilst using VR can be very challenging. Facial expressions are hard to read when one is wearing a VR head-mounted display (HMD) and people generally find it difficult to verbalize subjective indicators of their current mental state [17]. Additionally, fear responses may change throughout the virtual exposure, while self-reporting on them tends to focus the evaluation on only the last moments of virtual exposure and could interfere with the person's experience in the virtual environment (VE) [18]. Therefore, this research aims to combine VR with non-invasive neuroimaging to unobtrusively detect a person's fear response during virtual exposure.

Not all non-invasive neuroimaging modalities are suitable for a combination with VR. Functional near-infrared spectroscopy (fNIRS) seems to be the most appropriate technique when compared to the other non-invasive methods (electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI)) [7]. The main reason for this is that the ability to move around freely, which is desirable to create realistic VR scenarios, is very limited in the other modalities, due to their high sensitivity to motion artifacts. Furthermore, MEG and fMRI equipment restrain the subject to a very minimal area wherein it is almost impossible, if not undesirable, to move. fNIRS is less sensitive to motion artifacts than the other non-invasive modalities [19], while its portable and lightweight head-caps enable the subject to move to some extent [20]. Therefore, fNIRS technology exhibits the greatest potential among the non-invasive neuroimaging techniques for a combination with VR.

1.2.1 Goals and Research Questions

This research investigates the possibility of inducing and detecting a fear response in VR, using fNIRS data. However, fear responses can be elicited by many different VR stimuli. Examples of VRET applications from the literature were targeted at fear of spiders [21–23], fear of flying [24–26], fear of heights [27–30], fear of driving [31], and even posttraumatic stress disorders [32–35]. Taking the limited time scope of this thesis research into account, it was decided to aim for inducing and detecting a fear of heights response. This decision was made as it was expected that creating a VE that induces a fear of heights response is the least complex and the least time-consuming, as compared to creating a VE that induces any other type of fear.

No previous research has investigated whether the fNIRS data of people with a fear of heights response and people without a fear of heights response are actually different. Therefore, this is the focus of the first research question, which is defined as follows:

> **1** *To what extent do the fNIRS signals captured from people with a fear of heights response and people without a fear of heights response differ?*

In order to answer this question, both people with fear of heights (experimental group) and people without fear of heights (control group) were invited to participate in an experiment, during which they were exposed to virtual heights and virtual ground conditions. It was hypothesized that the virtual heights cause a fear response for the experimental group, whereas it does not cause a fear response for the control group. Furthermore, it was hypothesized that the ground condition does not cause a fear response for any of the groups. Between-group statistical analyses were performed on the fNIRS data of both groups to determine if there are significant differences between the groups.

Furthermore, this research investigates if the fear responses of the experimental group can be detected using machine learning classifiers. Therefore, the second research question is formulated as follows:

> **2** *To what extent can a person's fear of heights response to a virtual reality environment be detected using fNIRS data?*

The answer to this research question is obtained using the fNIRS data of the experimental group, since this group experienced fear responses as well as no-fear responses. Within-group statistical analyses were performed to determine if there are significant differences between the fNIRS data of the experimental group during the ground trials (i.e. "no fear") and during the height trials (i.e. "fear"). Then, subject-dependent and subject-independent classifiers were trained and tested on the data of the experimental group, with the goal to classify between "fear" and "no fear" data. The accuracies of the classifiers serve as an indicator of the performance of the fear detection.

1.3 Report Structure

This report describes the work that was done in order to answer the research questions that were posed in this chapter. First, a review of the literature will be given in Chapter 2. This review consists of definitions of VR and fNIRS, an explanation of fNIRS technology, findings from other works that used fNIRS to detect mental states, related work on the combination of VR and fNIRS and the use of other modalities to detect fear responses induced by VR, and background information on the statistics and classifiers used in this research. Then, Chapter 3 will describe the method that was used to answer the research questions. The methods for collecting the data through the experiment as well as processing it, are described in this chapter. Chapter 4 gives an overview of the results that were generated by the experiment, which can be divided into the results of the statistical analyses and the classification results. After that, a discussion of the results will be given in Chapter 5. Finally, Chapter 6 concludes this thesis research by answering the research questions.

Chapter 2

Literature Review

This chapter contains the literature review. First, relevant background information on VR technology and fNIRS technology is given. Then, the literature on mental states that can be measured with fNIRS is described. Additionally, the previous work on the combination of immersive VR and fNIRS and on the use of physiological signals to measure or detect fear responses in VR is reviewed. Finally, background information on the statistics and classifiers used in this research is given.

2.1 Virtual Reality

Virtual reality (VR) can be described as an advanced human-computer interface which presents a real-time three-dimensional simulation of an environment or situation to the user [1, 3, 5]. Typical VR environments (VEs) allow user interaction [1, 5], enabling the user to see the environment from different angles, to move around in it, and to touch, grab, or manipulate its three-dimensional objects [3]. Often, VR addresses multiple senses of the user, including visual, auditory and sometimes even haptic stimulation [5]. The more senses are addressed in a realistic manner, the more immersive VR the is [3]. An example of an immersive VE is given in Figure 2.1.

Figure 2.1: Immersive VE that shows a 3D simulation of a cockpit, an instructor, and the user's hand in real-time. This VE is used by Airbus for pilot training purposes. Image obtained from [36].

2.1.1 Immersiveness and Presence

Immersive VR systems typically include head-tracking sensors, a head-mounted display (HMD), sound effects, and an input device for user interaction with the environment [10, 37]. The head-tracking sensors are used to compute the user's head position with respect to the VE and to determine the user's vision based on that. The HMD, also called VR glasses or goggles, displays the VE to the user while blocking the user's view of the actual (i.e. physical) world [37]. Figure 2.2 shows an

example of a user wearing an HMD while using a hand-held controller as input device to control a VE.

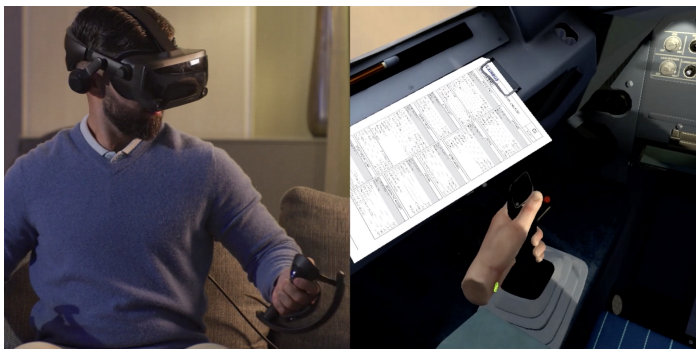


Figure 2.2: A user wearing an HMD and using a controller to interact with the VE (left) and the vision of the user in the VE of the cockpit from Figure 2.1 (right). Image obtained from [36].

The main attribute that distinguishes VR from other human-computer interfaces is the sense of 'presence' that it induces [1], which makes a user feel as if he/she is actually physically present in the VE [11, 37]. This feeling is typically only caused by immersive VEs. When a person feels physically present in the VE, this person will most likely respond in a realistic way to the virtual stimuli [3, 4]. Therefore, experiments, training, and therapy sessions that use realistic immersive VR are able to reach a high level of ecological validity [8], which can be too dangerous, expensive or simply impossible to create otherwise [4, 6].

2.2 Functional Near-Infrared Spectroscopy

Functional Near Infrared Spectroscopy (fNIRS) is a non-invasive neuroimaging modality that utilizes light in the near-infrared (NI) spectrum (650 nm – 1000 nm wavelength) to detect concentration changes of the chromophores oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) [19, 38–43]. fNIRS relies on the principle of neurovascular coupling, which describes the relationship between neural activity and changes in cerebral blood flow because of that activity [43]. Neural activity demands for increased oxygenated blood in the activated cortical area [19, 39, 44]. The supply of oxygen to an activated cortical area exceeds its oxygen consumption rate, causing an increase in HbO concentration and an accompanying decrease in HbR concentration. This phenomenon is also described as the hemodynamic response and is indicative of brain activity [42, 43].

Skin, tissue, and bone are generally transparent to NI light, while HbO and HbR absorb it [38, 41, 43, 46]. The fact that HbO and HbR have different molar absorption coefficients for varying wavelengths of NI light makes it possible to detect the two separately [19, 43]. Figure 2.3 shows the molar absorption coefficients for both HbO and HbR at varying wavelengths. The molar absorption coefficients are identical at around 800 nm wavelength. Therefore, fNIRS systems typically use at least two wavelengths to be able to dissociate between HbO and HbR: one below 800 nm and one above 800 nm [19, 44, 46].

2.2.1 Brain-Signal Acquisition

Brain signals are acquired through emitter-detector pairs that operate at varying wavelengths, often around 780 nm and 830 nm [44]. Every unique emitter-detector pair is a measurement channel, whereas a single emitter or detector can be referred to as an optode [43]. The NI light is distributed in a banana-shaped region between the emitter and the detector [39, 46], as can be seen in Figure 2.4. The depth at which the brain signals are measured is approximately half the distance between emitter and detector [43, 46]. A trade-off exists between measurement depth and signal quality [43]. Emitters and detectors that are placed too close to each other ($\sim$ 1 cm apart) will only measure

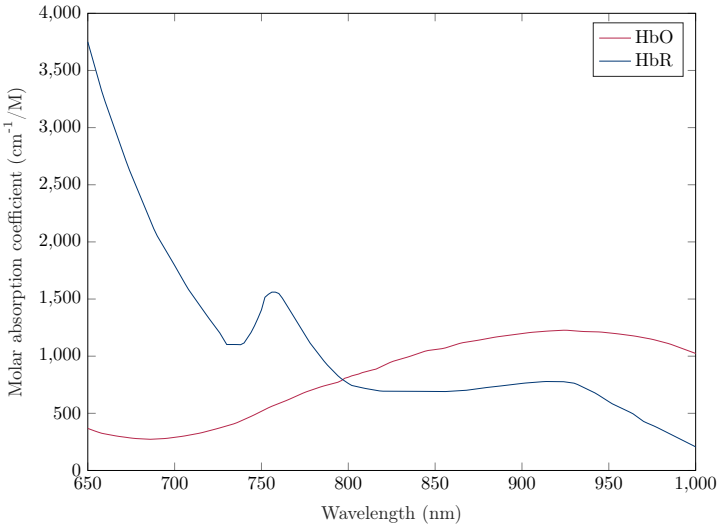


Figure 2.3: Molar absorption coefficients of HbO and HbR for different wavelengths within the NI spectrum, data obtained from [45].

skin, whereas placing them too far apart ($\sim$ 5 cm apart) will weaken the signal [19]. The optimal distance between an emitter and a detector is approximately 3 to 3.5 cm [19, 38, 43, 46]. However, the optimal distance might vary depending on the NI light intensity, the wavelengths, the age of the subject, and the brain area that is measured [38].

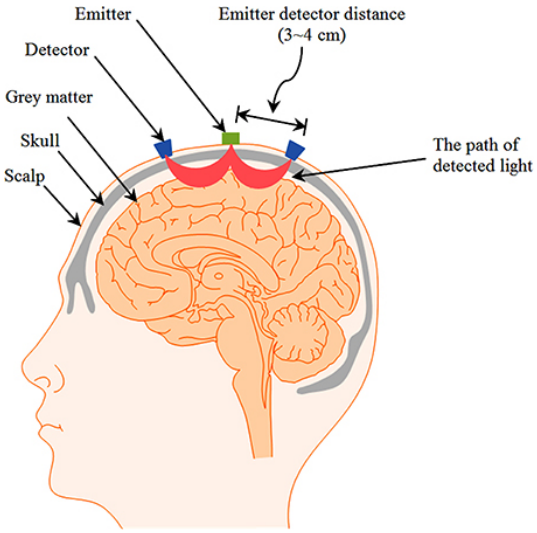


Figure 2.4: Schematic overview of emitter and detector placed on the scalp and the banana-shaped light distribution between them [19].

2.2.2 Deriving Chromophore Concentration Changes

The HbO and HbR concentration changes can be derived based on the Modified Beer-Lambert law (MBLL), which extends the Beer-Lambert law by taking into account the scattering of light with a scattering-dependent light intensity loss parameter (G) [44, 47]. The MBLL describes the

loss of light intensity (optical density, OD) as a function of chromophore concentrations (c), molar extinction coefficients (ϵ), distance between emitter and detector (l), path length of light scattering (differential pathlength factor, DPF) and loss parameter G, see equation 2.1. OD is expressed as the logarithm of the quotient of the detected light intensity (I) and the emitted light intensity (I_0) on the tissue. Chromophores HbO and HbR are expressed by index i. The variables t and λ denote time and wavelength, respectively. Figure 2.5 gives an example plot of OD.

$$OD(t,\lambda) = -\log_{10}\frac{I(t,\lambda)}{I_0(t,\lambda)} = \sum_i \varepsilon_i(\lambda)\cdot c_i(t)\cdot l\cdot DPF(\lambda) + G(\lambda) \tag{2.1}$$

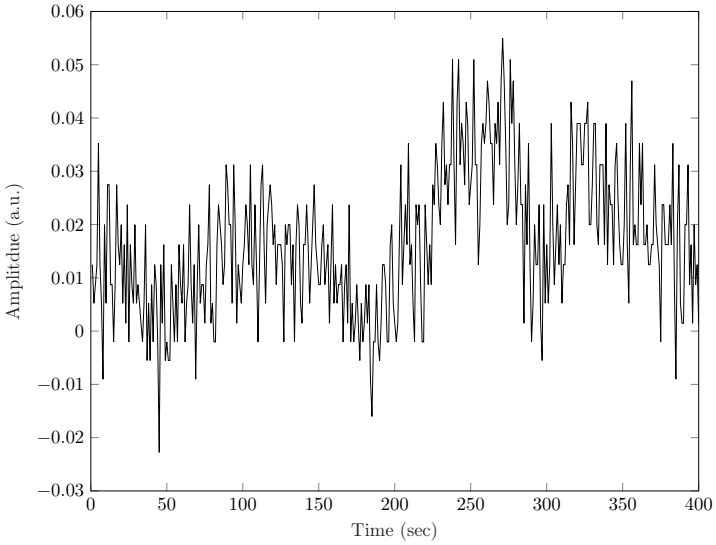


Figure 2.5: An example of a plot of OD. The OD data used in this plot was obtained from [48] and baseline corrected before generating the plot.

The change in optical density $\Delta OD(\Delta t,\lambda) = OD(t_1,\lambda) - OD(t_0,\lambda)$ can be computed under the assumption that there is constant light scattering loss over time, thus eliminating G from equation 2.1, see [44, 46]. Furthermore, it is assumed that the emitted light intensity I_0 is constant as well [44]. This yields the following equation for the change in optical density ΔOD:

$$\Delta OD(\Delta t,\lambda) = -\log_{10}(\frac{I(t_1,\lambda)}{I(t_0,\lambda)}) = \sum_i \varepsilon_i(\lambda)\cdot \Delta c_i\cdot l\cdot DPF(\lambda) \tag{2.2}$$

Solving equation 2.2 for Δc_i at two different wavelengths λ_1 and λ_2 yields equations 2.3 and 2.4 for chromophore concentration changes Δ[HbO] and Δ[HbR], respectively. See Appendix A for a step-by-step approach to deriving these equations.

$$\Delta[HbO] = \frac{\varepsilon_{HbR}(\lambda_2)\cdot\frac{\Delta OD(\Delta t,\lambda_1)}{l\cdot DPF(\lambda_1)} - \varepsilon_{HbR}(\lambda_1)\cdot\frac{\Delta OD(\Delta t,\lambda_2)}{l\cdot DPF(\lambda_2)}}{\varepsilon_{HbO}(\lambda_1)\cdot\varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbO}(\lambda_2)\cdot\varepsilon_{HbR}(\lambda_1)} \tag{2.3}$$

$$\Delta[HbR] = \frac{\varepsilon_{HbO}(\lambda_1)\cdot\frac{\Delta OD(\Delta t,\lambda_2)}{l\cdot DPF(\lambda_2)} - \varepsilon_{HbO}(\lambda_2)\cdot\frac{\Delta OD(\Delta t,\lambda_1)}{l\cdot DPF(\lambda_1)}}{\varepsilon_{HbO}(\lambda_1)\cdot\varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbO}(\lambda_2)\cdot\varepsilon_{HbR}(\lambda_1)} \tag{2.4}$$

2.2.3 Data Pre-Processing and Analysis

According to Pinti et al. [40] and Hocke et al. [49], the data analysis approaches of different fNIRS researches vary significantly. Therefore, it is difficult to define a standard method for the analysis of fNIRS data. In an effort to identify a more general approach, they reviewed the data analysis methods of other fNIRS studies and tested these different methods within their own experiments. The results of their reviews are given below.

A typical first step in the analysis of fNIRS data is to visually inspect the signal and assess its quality. Motion artifacts [39], instrument and environment noise, and poor coupling of optodes on the scalp can significantly degrade the signal quality [19, 40, 42, 49]. Signals that do not show cardiac oscillations should be excluded, because the absence of cardiac oscillations indicates that changes in the signal are not coupled with hemodynamic changes [49], thus making the signal meaningless. Channels with large artifacts, often visible as sudden spikes, can be removed upon visual inspection [40]. However, automated methods, like assessing every channel's coefficient of variation, are less subjective and less time-consuming [49]. Therefore, the usage of such methods is preferred when working with larger datasets and in cases of real-time detection.

The second step is to convert the raw light intensities to changes in optical density and then to HbO and HbR concentration changes using equations 2.3 and 2.4 [40, 46]. The HbO and HbR concentration changes should be compared against a baseline period where no stimulation was present [40, 42, 46]. This can for example be done by subtracting the mean HbO and HbR concentration changes during the baseline period from every HbO and HbR concentration change during stimulation, respectively [50].

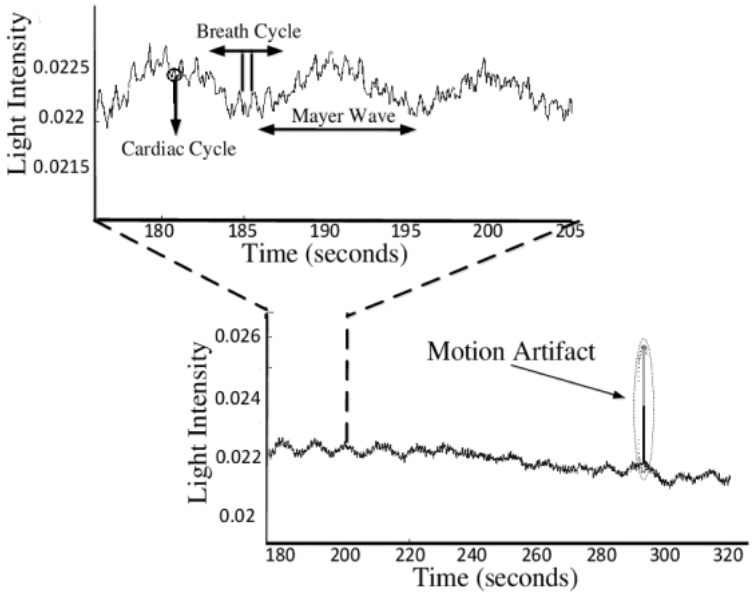


Figure 2.6: Example plot of physiological noises and a motion artifact in an fNIRS signal, figure obtained from [51].

A next step is to filter out the physiological noises that contaminate the fNIRS signal. Sources of physiological noise include breath cycles ($\sim$ 0.2 - 0.3 Hz), cardiac cycles ($\sim$ 1 Hz), and Mayer Waves ($\sim$ 0.1 Hz) [19, 39, 40]. See Figure 2.6 for a visualization of such noise signals. Digital filters (i.e. low-pass filters, band-pass filters or high-pass filters) can be used to reduce the physiological noises in the fNIRS signal. In most fNIRS studies, a Butterworth filter is used [40, 49]. Pinti et al. advise to use a band-pass filter, with a low cut-off frequency of 0.01 Hz and a high cut-off frequency above the stimulation frequency but below the Mayer Waves frequency of approximately 0.1 Hz [40]. This way, the physiological noises, which have frequencies of 0.1 Hz or higher, will be filtered out of the signal, while the important information about the stimulation remains present.

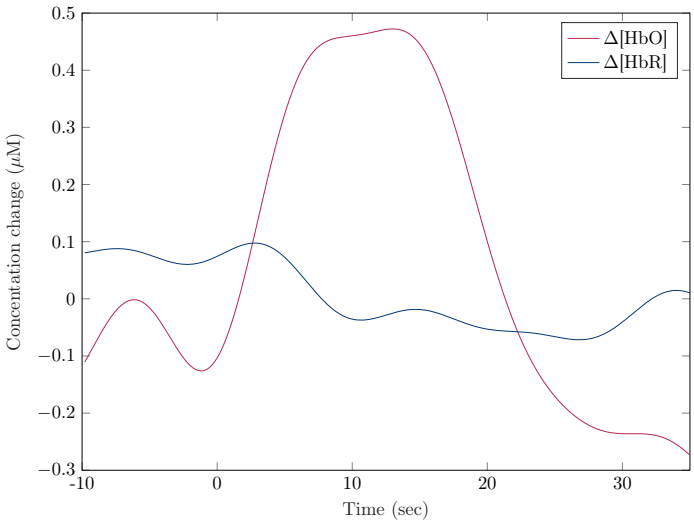


Figure 2.7: An example of a plot of pre-processed Δ[HbO] and Δ[HbR]. The data used in this plot was obtained from [48] and filtered and averaged over trials before generating the plot.

This recommendation was made based on the fact that they achieved the highest performance on signals that were filtered this way, where performance is defined as the amount of influence the filter had on the statistical inference.

Once the filtering step is completed, the Δ[HbO] and Δ[HbR] signals are pre-processed and can be used for statistical analyses. Figure 2.7 gives an example of a typical plot of pre-processed Δ[HbO] and Δ[HbR] signals.

2.2.4 Advantages and Limitations

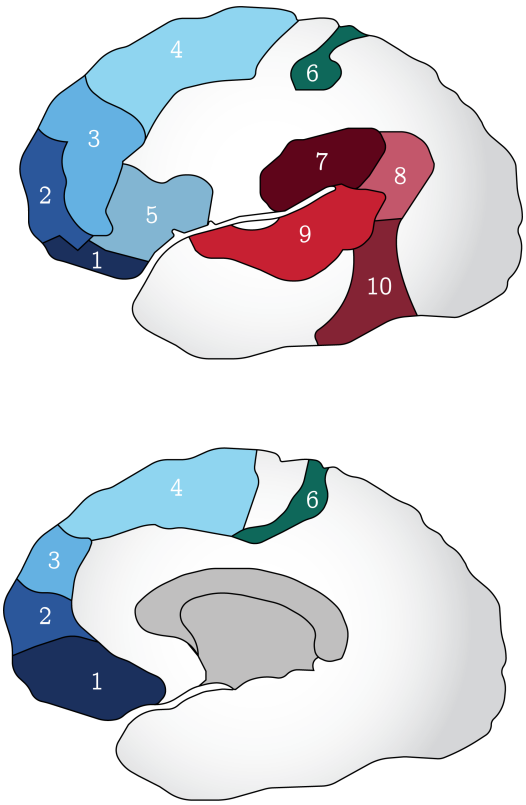
The use of fNIRS has several advantages over other neuroimaging modalities. First of all, fNIRS is completely safe, portable, and equipment costs are moderate to low as opposed to most other neuroimaging modalities [19, 39, 41–43]. Secondly, fNIRS measurements are relatively resistant to movement artifacts as compared to all other non-invasive neuroimaging modalities [19, 42, 43]. This, and the fact that the equipment is portable, allows fNIRS measurements to be taken in naturalistic environments without many movement restrictions for the participant [39, 41]. Therefore, experiments with high ecological validity can be executed. Finally, fNIRS is also compatible with other neuroimaging modalities, such as EEG [39].

Besides the advantages, the use of fNIRS also has its limitations. As explained before, this neuroimaging modality is not capable of measuring activity in the deeper brain regions [39, 42, 43]. Therefore, only the activity in the outer cortical regions can be assessed, which limits experimental designs. Furthermore, hair and dark skin color tend to weaken the NI light [39], which makes it difficult to use fNIRS on certain subjects. Especially thick hair can obstruct the contact between the optodes and the subject's scalp. Also, the spatial resolution of fNIRS is limited as compared to fMRI, although it is superior to that of EEG [42, 43]. On the other hand, the temporal resolution of fNIRS is inferior to that of EEG [43], due to the hemodynamic delay in the signals. When there is an activation in the brain, it takes approximately 5 to 7 seconds before a peak in the hemodynamic response can be observed [43, 52]. Therefore, fNIRS is an inappropriate modality for the observation of instantaneous events.

2.3 Mental State Detection with fNIRS

The effects of a multitude of mental states on fNIRS measurements were investigated in previous work. These states include mental workload [53–65], mental stress [66–73], fear responses [74–

81], affective responses [82–86], attentional state [87–92], deception [93–97], preference [52, 98–100], anticipation [101–103], suspicion [104, 105], and frustration [105–107]. In the following sections, it will be discussed how such mental states are measured using fNIRS. See Appendix B for a complete overview of the mental states and their effects on the oxygenated and deoxygenated hemoglobin concentration changes. Only a small portion of the literature that is under review in this section also focused on the detection of the mental state using machine learning classifiers. Table 2.1 summarizes the mental states that were detected, along with the classification algorithms that were used and the performances (i.e. accuracies) of the classifiers. For an overview of the brain areas that are mentioned in this section, see Figure 2.8.



Brain area	Number
Prefrontal cortex	1 to 5
Orbitofrontal cortex	1 + 2
Anterior PFC	2
Superior frontal gyrus	3 + 4
Dorsolateral prefrontal cortex	3
Ventrolateral prefrontal cortex	5
Sensory association cortex	6
Supramarginal gyrus	7
Temporoparietal junction	8
Superior temporal gyrus	9
Occipitotemporal area	10

Figure 2.8: Rough estimation of the locations of the brain areas where mental states were measured using fNIRS. This figure shows a schematic lateral view (top) and medial view (bottom) of the human brain. Brain areas are denoted by numbers, the names of the areas are given in the table on the right.

2.3.1 Mental Workload

The human brain contains a limited amount of mental resources [58], which determine what a person can or cannot do. Mental workload can be defined as the portion of those limited mental resources that are demanded by a task [53, 54, 56, 60]. When a task demands more mental resources than a person has available, the person's performance generally decreases [53, 54], leading to slower task performance and human errors [58, 60]. Furthermore, mental overload can cause cognitive tunneling, which can be defined as a person's inability to redistribute his/her attention from one task to another. [54, 58].

A great body of fNIRS research is dedicated to measuring the effects of mental workload on the hemodynamic activity, often focusing on the differentiation between diverse levels of mental workload based on n-back tasks [53, 55–57, 61, 63, 65]. Studies that aim at measuring mental workload effects on fNIRS signals generally measure the hemodynamic response over the prefrontal cortex (PFC),

which is a logical choice as this region has a functional relationship with working memory [54]. Such studies often report a positive relation between mental workload and HbO concentration changes [53–56, 58, 59, 61–64]. However, some studies focus on the HbR concentration changes instead. These studies complementarily report a negative relation between HbR concentration changes and mental workload [57, 60, 63, 65].

Whereas some studies only mention that they measured cortical activations over the PFC [55, 56, 60, 61, 63, 65], others specify the areas with significantly higher or lower concentration changes in more detail. Areas that showed significantly higher HbO concentration changes differ per study, and include the dorsolateral PFC (dlPFC) [53, 54, 62], the left dlPFC [58], the left anterior PFC [64], and the left PFC in general [53, 59]. Regarding the HbR concentration changes, the right hemisphere was reported as an area with significant concentration changes [57].

The work of Aghajani et al. [56] focused on the detection of mental workload from fNIRS data. To this end, they trained and tested a linear Support Vector Machine (SVM) on the data of 17 participants who performed n-back tasks. The linear SVM performed at an average accuracy of 74.8% in the case of a binary classification task (rest versus 3-back task) over a 5-second window. The features that were used in the classification consisted of the amplitude, slope, standard deviation, kurtosis and skewness of the HbO and HbR concentration changes.

2.3.2 Mental Stress

Mental stress can be defined as the state in which a person believes that what is expected from him/her exceeds their abilities [66, 67]. Both the body and mind respond to stress. The hypothalamus-pituitary-adrenocortical axis and the sympathetic nervous system are both activated by stress, which causes an increase in the cortisol production in the body [66–69]. Next to cortical activity, stress can be measured by heart rate variability (HRV), blood pressure, and galvanic skin response (GSR) [67, 68, 70–72].

The literature on the effects of stress on fNIRS signals shows mixed results. Some studies mention that in stress conditions, the concentration change of HbO decreases as compared to control situations. This effect was observed over the right PFC [66, 68] and the ventrolateral PFC (vlPFC) [73]. One of those studies hypothesizes that the lowered HbO concentration changes could be due to task disengagement [73]. Other studies show contradictory results, which indicate that the HbO concentration changes during stress situations are higher as compared to control situations. The significant brain regions in those cases include the right PFC [69] with electrode position FP2 mentioned specifically [67], the right dlPFC [70, 72], the left vlPFC, and the sensory association cortex [72].

Parent et al. [71] used the Naive Bayes classifier to discriminate between stress and no stress, based on the fNIRS data of 17 participants. The averages and slopes of the HbO and HbR concentration changes were used as features in the classification. Their classifier performed at an average accuracy of 63%.

2.3.3 Fear Response

The fear circuit includes multiple brain areas, which are related to emotion and managing attention and cognitive control [74–76]. The latter is also called the Cognitive Control Network (CCN), and comprises of the dlPFC, the vlPFC and the angular gyrus [74]. Due to the fNIRS measurement capabilities, the literature on fear responses measured with fNIRS is mainly about activities in the CCN, which can be measured over the PFC area. The PFC is connected to both the induction and regulation of emotions, such as fear responses [76], and therefore plays an important role in the mediation of fear responses [75].

The majority of fNIRS studies about cortical responses to fear-invoking stimuli report an increase in cortical activations in the parietal cortex [77, 78] or the PFC [74, 76, 79–81] during fearful stimulation. The studies that found activations in the parietal cortex presented subjects to fearful and neutral sounds. Decreased HbR concentration changes [78] and higher HbO concentration changes [77] were found when subjects were listening to fearful sounds as compared to neutral

sounds. The areas with significant activations include the (right) supramarginal gyrus and the right superior temporal gyrus (STG).

The studies that found an increased cortical activation in the PFC exposed their subjects to spiders [74], fearful faces [76, 79], a fear learning experiment based on shocks [80] or virtual heights [81]. All those studies measured increased HbO concentration changes in the PFC when subjects were exposed to the fearful stimuli as compared to the control situations. The complementary decrease in HbR concentration changes were only reported in one case [76]. PFC areas where significant activations were found include the left PFC [80], dlPFC, anterior PFC [81], left dlPFC, and left vlPFC [74]. One of the studies recorded cortical responses to fearful stimuli over multiple sessions and reported decreased activation of the PFC over sessions, along with a decrease in fear symptoms [74].

All of the above studies were conducted with healthy participants. This is important, as studies conducted on patients with anxiety disorders display contradictory results. Next to the studies mentioned above, other (non-fNIRS) studies also reported that fearful responses in healthy subjects lead to increased cortical activity in the PFC [108, 109], which is inversely related to the activity in the amygdala [109]. On the contrary, it was found that patients with anxiety disorder show decreased activity in the PFC in response to fearful stimuli and increased activity in the amygdala instead [110–112]. A similar effect was observed in an fNIRS study that used a cave automatic virtual environment system to expose subjects with moderate acrophobia to artificial heights [75]. Their subjects displayed decreased HbO concentration changes in the dlPFC and anterior PFC during the first exposure session. However, towards the third exposure session, significant increases in HbO concentration changes were detected in the dlPFC and anterior PFC, accompanied by significant decreases in HbR concentration changes in the right dlPFC. Based on this observation, the authors hypothesize that subjects learned how to manage their fear responses better.

2.3.4 Affective Responses

The induction and regulation of emotional responses cause cortical activations [75]. Next to fear responses, the fNIRS field also studied multiple other affective responses. Such fNIRS studies investigated the cognitive evaluation of threatening stimuli [82], neural correlates of affective responses to robot interlocutors [83], cortical activations caused by emotional stimuli [84], and the effect of negative mood on prefrontal activations during working memory tasks [85, 86].

Some of those studies interpret the cortical activity that they measured as related to emotion regulation. Those studies found increased activation in the ventrolateral PFC (vlPFC) during the labeling of threatening visual stimuli [82] and increased HbO concentration changes in the PFC when people were responding aversively to a robot [83]. Another study is more related to the induction of emotion and focuses on distinguishing between emotional and neutral audio-visual stimuli. The results suggest that exposure to stimuli from the emotional classes, which varied in valence and arousal, resulted in increased HbO and decreased HbR concentration changes over the PFC, whereas the opposite effect was observed for the neutral stimuli [84]. However, it was not possible to distinguish between the emotional classes based on the hemodynamic responses. Finally, the effects of negative and positive mood on activity in the PFC during working memory tasks were also studied. The results show that negative moods significantly correlated with decreased HbO concentration changes in the left dlPFC during working memory tasks [85, 86].

The detection of affective responses was investigated by Heger et al. [84]. Using SVMs with radial basis function kernels, they were able to train a binary classifier that predicts between neutral states and low valence-high arousal states at an average accuracy of 67.9%, based on the data of 8 participants. The average HbO and HbR concentration changes over 5-second windows were used as features, because the usage of other time-domain fNIRS features did not significantly improve the average classification performance.

2.3.5 Attentional State

Attention can be defined as a person's ability to remain focused and alert during a cognitively demanding task [87–90]. Since attention performance is dependent on the availability of mental resources [90], it is also related to mental workload.

Traditionally, reaction times (RTs) are used to measure a person's attentional state, as increasing RTs indicate attention losses [89, 91]. It was observed that RTs correlate with the time at which the HbO concentration change peaks in the PFC and parietal cortex, with longer RTs resulting in later peaks [89]. Furthermore, the general trend seems to be that HbO concentration changes measured over the PFC increase during the performance of tasks that require attentional resources. Such effects were measured over the dlPFC [88] and over the right PFC [87, 91], which is in accordance with the claim that right lateralization is related to attention [89, 90]. On the contrary, one study that focussed on distinguishing between 'drowsy state' and 'alert state' during a driving task, found that the mean HbO concentration change over the right PFC during the drowsy state is higher as compared to the alert state [92].

The detection of attentional state based on fNIRS signals was investigated using SVM [88, 90–92] and Linear Discriminant Analysis (LDA) [92] classifiers. Harrivel et al. [88] were able to discriminate between rest and task periods using an SVM at an average accuracy of 83.8% over 7 participants. As features, the HbO and HbR concentration changes of the optodes that showed the highest task discrimination based on F-scores were used. A similar average accuracy was obtained by Khan et al. [92], who used SVMs and LDAs to decode alert versus drowsy attentional states. They obtained the highest average accuracies when using the mean HbO concentration changes, the signal peaks, and the sum of peaks over 5-second windows as features. The LDA classifier reached an average accuracy of 83.1% over the 13 participants, whereas the average accuracy of the SVM classifier was 84.4%. Similarly, Zhang et al. [90] used an SVM classifier to distinguish between the attentional states of easy and hard tasks based on the fNIRS data of 15 participants. The selected features were the mean, signal slope, power spectrum, and approximate entropy of the HbO and HbR concentration changes over a 10-second window. Their binary classifier that discriminated between attentional states during easy and hard tasks performed the best, at an average accuracy of 81.53% over participants. They also implemented a multi-class classifier in order to discriminate between the attentional states during easy, medium, and hard tasks. The average accuracy of this classifier was 57.04%. Furthermore, Derosière et al. [91] discriminated between full attentional states and decremented attentional states using an SVM. As features, they used the HbO and HbR concentration changes averaged for each 1-second epoch duration. The average accuracy over their 7 participants was highest when using both the HbO and the HbR features over the PFC and the right parietal area, which resulted into an average accuracy of 90.7%.

2.3.6 Deception

Deception, the act of deliberately concealing the truth, is a mentally demanding task [93] which seems to gain increasing interest in the neuroimaging field. A number of fNIRS studies measured the effect of deception on fNIRS measurements. In general, these studies report increased HbO concentration changes over the PFC during deception-related tasks as compared to the neutral control tasks. The locations within the PFC where such activations were most significantly present include the left PFC [93, 94], with the left superior frontal gyrus (SFG) mentioned specifically [95], the right anterior PFC [93], the right SFG [94, 96], and the bilateral dlPFC [97]. The differences in lateralization could indicate that there is a collaboration between the left and the right PFC during deception [93]. Some studies also reported complementary decreases of HbR concentration changes [93, 94]. However, those concentration changes were not significant as compared to the baseline.

Hu et al. [93] were able to detect deception using a binary SVM classifier at average accuracies of 83.44% (using radial basis function kernels) and 81.14% (using linear kernels) over 7 participants. The HbO and HbR concentration changes were used as features, along with their short histories over different time windows: 1 second, 3 seconds, and 5 seconds. The best accuracies were obtained using the 3-second window.

2.3.7 Preference

Some fNIRS studies investigated how preference can be measured using fNIRS signals. Since preference is a subjective rating, these studies employed user evaluations of the presented stimuli to determine what a person's actual preferred option is. The brain activities related to the preferred options were measured over the PFC. The preferred stimuli caused an increase in HbO concentration changes in the orbitofrontal cortex (OFC) [98, 99], the anterior PFC [100], and the right PFC [52] as compared to neutral stimuli. A simultaneous decrease in HbR concentration change was only measured by one study [99]. Interestingly, Hosseini et al. [98] detected an increase in HbR concentration change along with the increase in HbO concentration change. This phenomenon of simultaneous peaks in both HbO and HbR concentration changes has not been reported in any other study that is part of this literature review and seems to be inconsistent with the principle of neurovascular coupling, as explained in section 2.2.

Hosseini et al. [98] also investigated the decoding of attractive and unattractive visual stimuli based on fNIRS signals. Using a linear SVM, they decoded attractive versus other stimuli at an average accuracy of 72.9% over 5 participants. The average accuracy of the detection of unattractive versus other stimuli was 68.3%. As features they computed the average HbO and HbR concentration changes over 4-second windows (from 1 to 5 seconds post-stimulus onset) for every channel. Principal component analysis (PCA) was used to reduce the dimensionality of the data while keeping 99% of the variance.

2.3.8 Anticipation

Anticipation, the mental preparation for a certain event, is a mental state that was investigated by only a few fNIRS studies. These studies investigated the anticipation of a mentally demanding task [101], positive emotion [102], and a walking task [103]. All of those studies found increased HbO concentration changes under the anticipatory conditions. The HbR concentration change signal was excluded from their analyses due to its relatively low sensitivity and signal-to-noise ratio.

Both the anticipation of a mentally demanding task and the anticipation of positive emotion cause increased HbO concentration changes in the dlPFC [101, 102], with a left lateralization in the latter case. Such activations were significantly less for the anticipation of an 'easy' mental task or the anticipation of neutral or negative emotion. The anticipation and execution of a walking task elicited increased HbO concentration changes in the PFC and the premotor cortex, which were significantly less present for participants who were not anticipating the walking task [103].

2.3.9 Suspicion

Suspicion can be described as a demanding mental state that induces uncertainty and concern about the trustworthiness of certain information [104, 105]. It is important to note that only a very limited body of fNIRS research was dedicated to this mental state and that the two research papers that were found about this topic were partially written by the same authors.

Both studies used surveys to let the participants self-report on their emotions, cognitive load, and feelings of trust and distrust. Those results were used to identify the cases in which subjects were suspicious. In the first study, higher HbO concentration changes were found in the SFG for suspicious subjects as compared to non-suspicious subjects [105]. The second study, however, showed different results. In this case, higher levels of HbO concentration changes were reported in the OFC (Brodmann Areas 10 and 11) and some areas that are part of the left and right temporoparietal junction (TPJ) [104]. The activations in the OFC are in accordance with several fMRI studies, which report that the OFC is activated during decision-making in risky and uncertain situations [113, 114]. As uncertainty is a characteristic of suspicion, it seems logical that this mental state activates the OFC.

2.3.10 Frustration

Frustration is a negative mental state that is caused when goal-oriented actions are obstructed [106]. This mental state was also investigated by only a very limited number of fNIRS studies. One of those studies had the participants self-report on their feelings of frustration during a computer task [105], whereas the others constructed simulated driving scenarios that were labeled as either 'frustrating' or 'non-frustrating' [106, 107]. The results of these studies indicate that frustrating scenarios cause increased HbO concentrations in various cortical areas, including the dlPFC [105], the vlPFC, and the occipitotemporal area [106, 107].

The detection of frustration was investigated by Ihme et al. [106]. They used multivariate logistic regression to distinguish between the fNIRS measurements of frustrated and non-frustrated trials during a driving experiment. The fNIRS data of a total of 12 participants was collected and an average detection accuracy of 78.1% over participants was obtained. The normalized values of the pre-processed HbO and HbR concentration changes were used as features in the classification model.

Table 2.1: Overview of the previous work on the detection of mental states. This table contains the mental state that was detected, the classifier that was used, the number of participants in the study (N), the average detection accuracy, and the features used in the model.

Detection	Classifier	N	Accuracy	Features	Ref
Mental workload versus rest	SVM	17	74.8%	Amplitude, slope, standard deviation, kurtosis and skewness of Δ[HbO] and Δ[HbR]	[56]
Stress versus no stress	Naive Bayes	17	63%	Averages and slopes of Δ[HbO] and Δ[HbR]	[71]
Low valence-high arousal versus neutral	SVM	8	67.9%	Average Δ[HbO] and Δ[HbR] over 5-second windows	[84]
Attention versus rest	SVM	7	83.8%	Δ[HbO] and Δ[HbR] with highest F-scores	[88]
Attention during easy versus hard tasks	SVM	15	81.53%	Mean, signal slope, power spectrum and entropy of Δ[HbO] and Δ[HbR] over 10-second windows	[90]
Attention during easy versus medium versus hard tasks			57.04%		
Decremented versus full attention	SVM	7	90.7%	Δ[HbO] and Δ[HbR] averaged over 1-second epochs	[91]
Alert versus drowsy state	SVM LDA	13	84.4% 83.1%	Mean Δ[HbO], signal peaks and sum of peaks over 5-second windows	[92]
Deception versus truth telling	SVM	7	83.44%	Δ[HbO] and Δ[HbR] short histories over 3-second windows	[93]
Attractive versus other stimuli	SVM	5	72.9%	Average Δ[HbO] and Δ[HbR] over 4-second windows	[98]
Unattractive versus other stimuli			68.3%		
Frustration versus no frustration	Logistic regression	12	78.1%	Normalized Δ[HbO] and Δ[HbR]	[106]

2.3.11 Discussion

The literature described throughout this section implies that fNIRS can be used to measure the effects of mental workload, mental stress, fear responses, affective responses, attentional state, deception,

preference, anticipation, suspicion, and frustration. Although the different studies claim that they were able to measure different mental states, the brain signals that were indicative of those mental states have similar characteristics. The majority of the reviewed studies reported increased HbO concentration changes and/or decreased HbR concentration changes over the PFC under a certain condition as compared to a control condition or the baseline. Furthermore, the brain areas with the most significant activation for a certain mental state seem to differ per study. This implies that it is very difficult, if not impossible, to distinguish between different mental states based solely on fNIRS data. Therefore, the context of the experiment and potential other measurements (i.e. behavioral data, self-reports or other physiological signals) seem important to be able to claim which mental state was measured or detected using fNIRS.

2.4 Immersive VR and fNIRS

The combination of immersive VR and fNIRS for research purposes seems to be a novel one, based on the very limited amount of literature available on this topic. Previous studies focused on a virtual line bisection task [115], the assessment of prospective memory [116, 117], the processing of racial stereotypes [118], performance monitoring during training [119], and a neurofeedback system to help people focus their attention [120]. Each of these studies are described below, highlighting their main findings and how they combined the fNIRS measurements with a VR HMD.

Seraglia et al. [115] already investigated the combination of immersive VR and fNIRS in 2011. In order to do so, they assembled their own custom-made VR helmet from a bike helmet and the LCD screens of another HMD, see Figure 2.9. However, their fNIRS measurements were limited to the occipital area, because the helmet did not leave enough room for measurements over the PFC. Furthermore, their helmet was not adjustable in size, which caused problems with the measurements as head circumferences differ among people. Their experiment investigated cortical activations over the occipital area during a virtual line bisection task in peripersonal space (i.e. close to the subject's body) and extrapersonal space (i.e. further away from the subject's body). The fNIRS measurements of both conditions were not significantly different. However, they did find significant activity during the conditions as compared to the baseline period, over the right parietal and occipital lobes.

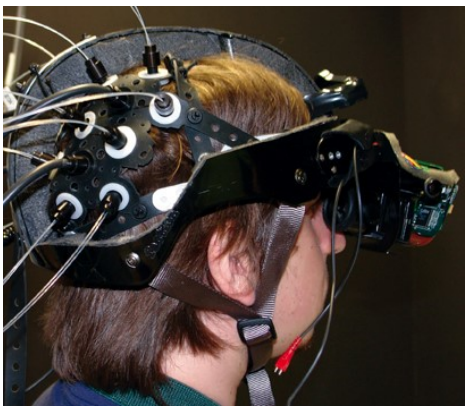


Figure 2.9: The custom-made helmet of [115], consisting of a bike helmet, fNIRS optodes, and LCD screens from another VR HMD.

The combination between immersive VR and fNIRS was also used to conduct two experiments on prospective memory [116, 117]. During the experiments, the participant was located in a virtual city environment. In this environment, the participant got a shopping list with items to collect and actions to undertake. This is referred to as the 'prospective memory' component. Furthermore, there was an 'ongoing' component that asked the participant to press a button every time he/she passed a store. Their results indicate that the hemodynamic activity over the anterior PFC is significantly greater during the prospective memory component than it is during the ongoing component [116]. In

a follow-up experiment, they compared the fNIRS data generated by the immersive VR experience to that generated by PowerPoint slides. Their results show that the hemodynamic response during the VR-based task was greater than during the slide-based task [117]. This could potentially indicate that higher task engagement can be achieved by VR experiments. The researchers also pointed to some challenges cause by the simultaneous use of fNIRS and the immersive VR HMD that they used, the Oculus Rift. First of all, the VR HMD makes it difficult to move the hair aside underneath the fNIRS optodes, hence the fNIRS signal quality gets degraded. Furthermore, the VR HMD and the fNIRS headcap had cables attached to them, which made it difficult for the participant to move freely.

Another study focused on the cortical processing of racial stereotypes using immersive VR and fNIRS [118]. During the experiment, participants were seated and exposed to a racially-charged VR scene and a non-racially-charged VR scene. The VR scenes were presented to the participants via the HTC Vive. A custom-made fNIRS probe arrangement was used, to measure over the medial and lateral PFC, see Figure 2.10. The results of the experiment show that there is significant activation over the right lateral PFC during the racially-charged exposure, which is absent during the non-racially-charged exposure.

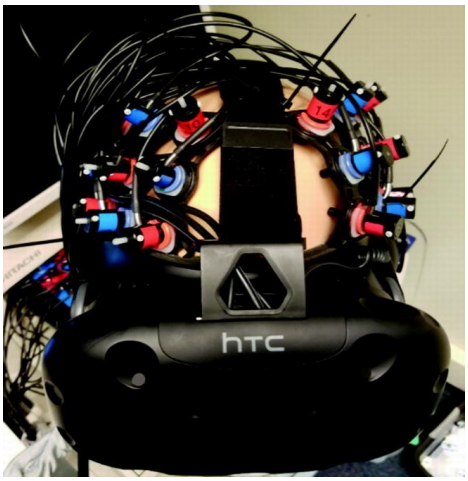


Figure 2.10: The HTC Vive HMD and a custom-made fNIRS probe arrangement [118].

Hudak et al. [119] investigated whether immersive VR training performance can be monitored using fNIRS measurements. They measured the cortical hemodynamic responses over the PFC area using an fNIRS sensor pad. This sensor pad fitted underneath the HTC Vive, which was their HMD of choice. Participants underwent a virtual tutorial during which they followed basic life support training. After the tutorial, there were two VR scenarios where the participants had to apply their newly acquired knowledge in the form of a serious game. The serious game performance of the participants was compared to their fNIRS measurements. The results show a negative correlation between performance and PFC activation, hence learning the contents of the training reduced the amount of mental resources needed to perform the tasks.

Additionally, the combination of immersive VR and fNIRS measurements was used to develop a neurofeedback intervention in which participants had to control room lighting with their dlPFC activity [120]. The Oculus Rift HMD was used in this research. After 8 training sessions over the course of two weeks, participants significantly increased their dlPFC activity on a go/no-go task, which indicates that they learned how to activate the dlPFC. The authors mention that immersive VR has the potential to improve the ecological validity of neurofeedback training situations.

2.5 Physiology of Fear in VR

A limited amount of research was conducted on measuring fear in VR scenarios based on physiology. Previous works collected data from various physiological signals, including HR [15, 121–124], GSR [15, 121, 122, 124–126], EEG [15, 122, 127], skin temperature [121, 126], blood volume pulse (BVP) [126, 128], and salivary cortisol levels [124]. These studies reported on the change in physiological signals during fear responses or the use of physiological signals to detect fear responses using machine learning classifiers. The VR scenarios that elicited these fear responses were focused on fear of flying [121], fear of heights [15, 122–125, 127], fear of public speaking [126], and social anxiety [128].

2.5.1 Physiological Effects of VR-Induced Fear

Several studies found that HR significantly increased between virtual ground conditions and virtual heights [122, 124], with a positive correlation between HR and self-reported fear of heights [123]. One study also reported an additional increase in HRV [122]. However, in two studies the significant increases were not only reported for the experimental group who suffered from fear of heights, but also for the control group who did not have fear of heights [123, 124]. On the contrary, another study that measured HR during fear of heights responses in VR reported no significant changes between the HRs of ground conditions and height conditions [125]. Furthermore, a study that focused on the HR measurements taken during VR flying scenarios did not find significant differences between the HRs of participants with flight phobia and healthy controls [121]. The authors suggest that more sensitive measures, like HRV, might have the potential to unravel differences between phobics and healthy controls.

Additionally, significant differences in GSR were measured during VR exposure scenarios. Several studies measured a significant increase in GSR during virtual heights conditions as compared to virtual ground conditions [122, 124]. However, this difference was not only measured for participants with fear of heights, but also for the control group who did not have a fear of heights [124]. Also, a positive correlation between GSR and the participants' self-reported fear was found [125]. Furthermore, in the case of VR exposure to flying scenarios, a significant difference in GSR between participants with flight phobia and healthy controls was found [121].

The literature suggests that skin temperature and salivary cortisol levels have less potential to discriminate between fear responses and non-fear responses, although the amount of findings are very limited. The skin temperatures of participants with flight phobia and healthy controls during an airplane flight in VR were not significantly different [121]. Also, no significant differences in salivary cortisol levels were measured between virtual ground and virtual height conditions [124].

2.5.2 Detecting Fear in VR

Several other studies used physiological signals acquired during VR exposure to detect fear responses using machine learning classifiers. Handouzi et al. [128] collected the BVP data of 7 participants who were exposed to VR scenarios related to social anxiety. Before, during, and after every exposure scenario, the participants indicated their perceived level of fear on the Subjective Units of Distress Scale (SUDS), which is an 11-point Likert scale. The SUDS scores served as the ground truth labels for the classifier. An SVM classifier was trained to discriminate between data from calm and anxious episodes. Their classifier performed at an accuracy of 76%.

Similar work was done by Salkevicius et al. [126], who trained an SVM classifier on BVP, GSR and skin temperature data to discriminate between different fear levels related to public speaking anxiety. They collected the data from 30 participants during a VR scenario in which the participants had to perform a public speaking assignment. There were multiple public speaking assignments, and the participants' SUDS scores were taken directly after the assignments and at baseline. The SUDS scores served as the ground truth labels for four different levels of fear: low, mild, moderate, and high. Using leave-one-subject-out-cross-validation, their 4-class classifier performed at an accuracy of 80.1%.

Another study conducted by Balan et al. [15] investigated the performance of multiple classifiers on the discrimination of two classes of fear (fear versus relaxation) and four classes of fear (relaxation, low fear, medium fear, and high fear). They collected GSR, HR, and EEG data from 8 participants who scored sufficiently high on the Visual Height Intolerance (VHI) questionnaire to be characterized as having fear of heights. During the experiment, the participants were exposed to virtual heights and virtual ground conditions, while their measurements were taken. Again, SUDS scores after every VR trial served as the ground truth labels for the classifiers. Their best subject-dependent classifiers performed at accuracies of 89.5% (two-class SVM) and 42.5% (four-class SVM). Their best subject-independent classifiers performed at accuracies of 79.12% (two-class deep neural network) and 52.75% (four-class k-Nearest Neighbors).

A similar experiment was conducted by Hu et al. [127], who collected EEG data from 60 participants during virtual heights and virtual ground conditions. After the experiment, participants indicated their level of fear of heights using the VHI questionnaire. The VHI scores served as the ground truth labels for four classes of fear: not strong, moderately strong, quite strong, and very strong. Using 10-fold cross validation, their 16-layer deep convolutional neural network was able to distinguish between the four classes at an accuracy of 88.77%.

2.6 Statistics and Classifiers used in this Research

This section provides the relevant background information for the statistical tests and classification algorithms that were used in this research. To this end, permutation testing, false discovery rate correction, linear discriminant analysis and support vector machines will be discussed in the following sub-sections.

2.6.1 Statistics

2.6.1.1 Permutation Test

The permutation test is a non-parametric test that can be used on small sample sizes and makes minimal assumptions about the distribution of the data, unlike the commonly used t-test or ANOVA [129]. Instead, it is a data-driven approach that utilizes all the possible values of the test statistic under random rearrangements (permutations) of the observed data, to obtain the distribution of the test statistic under the null hypothesis [130].

Let's say a given dataset consists of a list of observations that fall into one of two possible classes: class A and class B. Under the null hypothesis, the observations of class A and class B are assumed to follow the same distribution. If the null hypothesis is true, the observations can be exchanged from one class to the other while the value of the test statistic remains the same. The permutation test tests this null hypothesis by permuting the data and calculating the permutation test statistic T_{perm} for every permutation. See Figure 2.11 of an example.

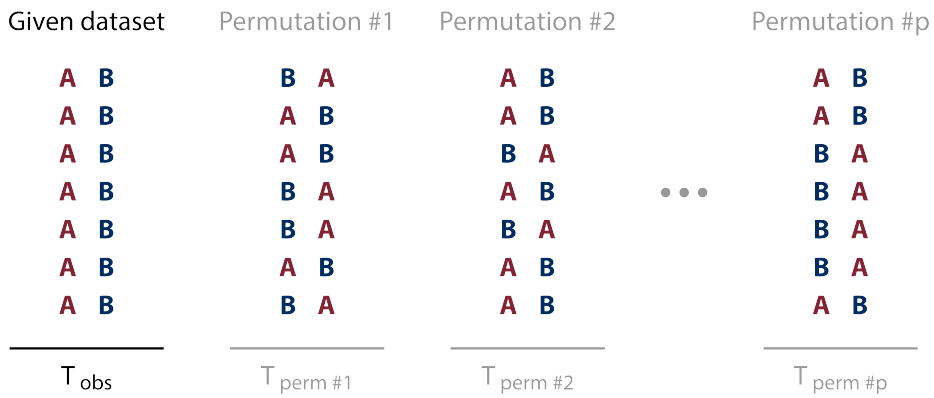


Figure 2.11: Comprehensive overview of the procedure of the permutation test.

The values of T_{perm} can be used to construct the permutation distribution. A comprehensive example is given in Figure 2.12. Based on this, the resulting p-value of the test is calculated by taking the proportion of permutations where T_{perm} is larger than the observed test statistic T_{obs}. The obtained p-value should be compared against a significance level α to determine if the null hypothesis can be rejected ($p < \alpha$) or not ($p \geq \alpha$).

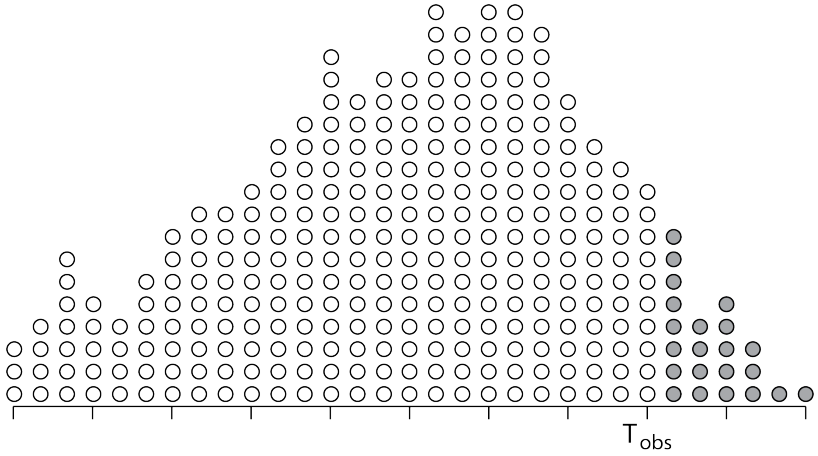


Figure 2.12: An example of a possible permutation distribution. The gray dots indicate the values of T_{perm} that are larger than the value of T_{obs}.

The number of possible permutations n_p equals the factorial of the number of observations n_{obs}, hence $n_p = n_{obs}!$. Even with a rather small dataset, this can easily lead to more than millions of possible permutations. In practice, however, a smaller number of permutations is often chosen, in order to reduce computation times. A minimum of 5,000 (at $\alpha = 0.05$) to 10,000 (at $\alpha = 0.01$) permutations is enough to approximate the distribution of the null hypothesis [131].

2.6.1.2 False Discovery Rate

When multiple hypothesis tests are executed simultaneously, the amount of p-values below the significance level due to chance increases. Wrongful rejection of the null hypothesis causes "false positive" findings, also referred to as type I errors [132]. The False Discovery Rate (FDR) correction is a way to correct for type I errors. It can be used with any statistical test for which a p-value is generated [133]. The FDR is defined as follows:

$$\text{FDR} = \frac{FP}{R} = \frac{FP}{FP + TP} \tag{2.5}$$

Where FP is the number of false positives (i.e. the number of times the null hypothesis is wrongly rejected), TP is the number of true positives (i.e. the number of times the null hypothesis is correctly rejected), and R is the total number of times that the null hypothesis is rejected. The FDR correction procedure for neuroimaging data as suggested by Genovese, Lazar and Nichols [133] is as follows:

1. Specify the rate q between 0 and 1, such that the FDR does not exceed this rate. If $q = 0.05$, it means that in only 5% of the cases a false discovery is made (on average).
2. Sort the p-values in ascending order: $p_1 \leq p_2 \leq ... \leq p_t$. Hypothesis test t_i corresponds to the p-value p_i.
3. Let r be the largest i for which the following holds:

$$p_i \leq \frac{i}{T} \cdot q \tag{2.6}$$

   Where T is the total number of hypothesis tests that were conducted.

4. Hypothesis tests $t_1, ..., t_r$ survive the FDR correction, hence it is concluded that their null hypotheses are correctly rejected and that a significant result is found.

2.6.2 Classifiers

2.6.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a probabilistic classification method that aims to separate the data of different classes using a linear decision boundary [134, 135]. Consider a two-class problem with dataset $\{\mathbf{x}_n, l_n\}$, where $\mathbf{x}_n$ are the observations (consisting of multiple features f such that $\mathbf{x}_n = \{x_{n1}, x_{n2}, ..., x_{nf}\}$) and l_n are the class labels, $l_n \in \{0, 1\}$. There are two different class labels, thus there are two classes, $C \in \{c_1, c_2\}$. The observations $\mathbf{x}_n$ are real numbers, hence $\mathbf{x}_n \in \mathbb{R}^D$. Following Bayes' theorem, the posterior probability that an observation x_n belongs to a certain class c is defined as follows [134, 136, 137]:

$$P(c|x_n) = \frac{P(x_n|c)P(c)}{P(x_n)} \tag{2.7}$$

LDA assumes that the data from each separate class comes from a multivariate Gaussian distribution with identical covariance matrices $\Sigma_{c_1} = \Sigma_{c_2} = \Sigma$ [134, 137]. The probability density function is then defined as follows:

$$P(x_n|c) = \frac{1}{\sqrt{2\pi^f|\Sigma|}}\exp(-\frac{1}{2}(x_n - \mu_c)^T\Sigma^{-1}(x_n - \mu_c)) \tag{2.8}$$

Where μ_c denotes the class mean of class c and Σ denotes the shared covariance matrix. These two parameters can be estimated as follows [134, 136]:

$$\mu_c = \frac{1}{N_c}\sum_{n:l_n=c} x_n \tag{2.9}$$

$$\Sigma = \frac{1}{N - C}\sum_{c=1}^{C}\sum_{n:l_n=c}(x_n - \mu_c)(x_n - \mu_c)^T \tag{2.10}$$

Where N_c is the number of datapoints that belong to class c and N is the total number of datapoints of the different classes.

The prior probability of class c, also referred to as π_c, is the proportion of the data samples that belong to class c, hence:

$$P(c) = \frac{N_c}{N} = \pi_c \tag{2.11}$$

The term $P(x_n)$ is identical for each class, hence it does not have an influence on the posterior probability and can therefore be canceled from equation 2.7. Now the posterior probability that an observation x_n belongs to class c can be calculated as follows:

$$P(c|x_n) = \frac{\pi_c}{\sqrt{2\pi^f|\Sigma|}}\exp(-\frac{1}{2}(x_n - \mu_c)^T\Sigma_c^{-1}(x_n - \mu_c)) \tag{2.12}$$

After simplification, the discriminant function of class c can be written as:

$$\delta_c(x_n) = x_n^T\Sigma^{-1}\mu_c - \frac{1}{2}\mu_c^T\Sigma^{-1}\mu_c + \log(\pi_c) \tag{2.13}$$

Every datapoint will be classified as belonging to the class for which the value of the discriminant function is highest [134], hence the decision boundary can be defined as the line where $\delta_{c_1}(x_n) = \delta_{c_2}(x_n)$, see Figure 2.13.

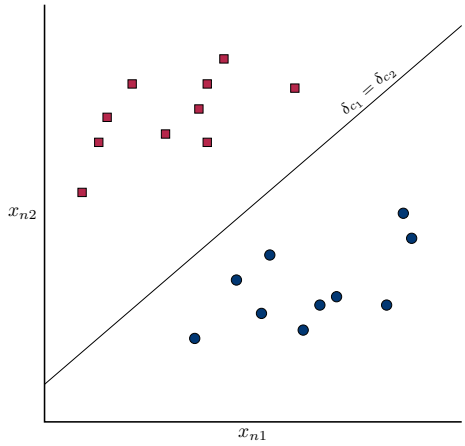


Figure 2.13: Example of a decision boundary made by LDA. Datapoints of two classes are shown: red rectangles and blue circles.

2.6.2.2 Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning algorithm that can be used for data classification and regression [136, 138]. SVMs are non-probabilistic binary linear classifiers that aim to find a linear hyperplane that separates the data into two different classes [136]. The hyperplane that creates the largest distance between itself and the nearest datapoints in the training set (i.e. the support vectors), is the one that yields the best separation. This distance between the hyperplane and the support vectors is also referred to as the margin. See Figure 2.14 for an example. The larger the margin, the less errors due to generalization by the classifier, hence the SVM tries to maximize the margin [134, 136, 138, 139].

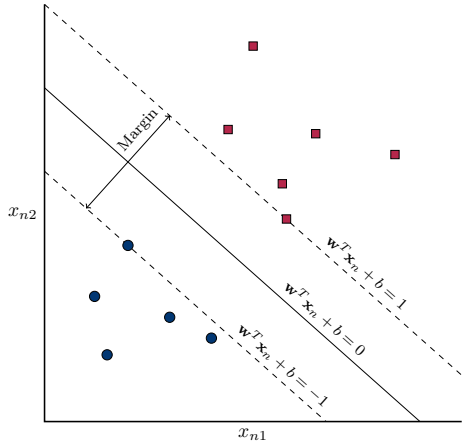


Figure 2.14: Example of the separating hyperplane (the black solid line) and the margin optimized by the SVM. Datapoints of two classes are shown: red rectangles and blue circles. The datapoints that touch the dashed lines are the support vectors.

Let's consider the two-class problem with the dataset from the previous section again. The dataset $\{\mathbf{x}_n, l_n\}$ consists of the real-valued observations $\mathbf{x}_n = \{x_{n1}, x_{n2}, ..., x_{nf}\}$ and labels $l_n \in \{0, 1\}$, hence $\mathbf{x} \in \mathbb{R}^D$. The SVM aims to find a linear separating hyperplane such that the margin is maximized. The linear separating hyperplane $y(x)$ is defined as follows [140]:

$$y(x) = \mathbf{w}^T\mathbf{x}_n + b \tag{2.14}$$

Where b is a constant for the offset of the hyperplane. The conditions for correct classification are:

$$y(\mathbf{x}_n) = \begin{cases} 1, & \text{if } \mathbf{w}^T\mathbf{x}_n + b \geq 1 \\ -1, & \text{if } \mathbf{w}^T\mathbf{x}_n + b \leq -1 \end{cases} \tag{2.15}$$

These conditions can be rewritten to:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \tag{2.16}$$

In this case, the margin equals $\frac{2}{||\mathbf{w}||}$. The goal is to maximize this margin, which is equivalent to minimizing

$$L(\mathbf{w}) = \frac{||\mathbf{w}||^2}{2} \tag{2.17}$$

Subject to the constraint given in equation 2.16. The Lagrange multiplier method can be used to solve this constrained optimization problem.

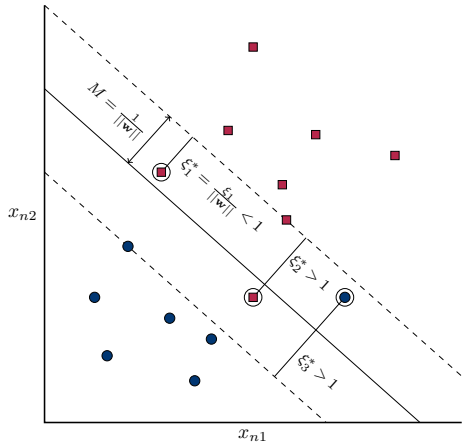


Figure 2.15: Example where the data are not linearly separable. The slack parameter is denoted by ξ.

So far, the only case considered is the one where the data are linearly separable. However, in many real-world applications, the data are not linearly separable. In such a case, one can introduce a slack parameter ξ in order to relax the constraints [136, 140], see Figure 2.15. The slack parameter relaxes the constraint such that most (instead of all) datapoints will be classified correctly, to avoid underfitting of the data. The slack variable relaxes the constraint as follows:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n,\ \xi_n \geq 0 \tag{2.18}$$

In order to maximize the margin, the following equation has to be minimized, subject to the relaxed constraint given in equation 2.18:

$$L(\mathbf{w}) = \frac{||\mathbf{w}||^2}{2} + C\sum_n^N \xi_n \tag{2.19}$$

Where C is a parameter that regulates the trade-off between the width of the margin and the number of datapoints that are missclassified during training [136].

Until now the decision boundary (hyperplane) was considered to be linear. However, there are also cases where a linear decision boundary is not applicable, an example is given in Figure 2.16a. In such cases, the datapoints can be mapped onto a higher dimensional feature space, where the data becomes linearly separable [141]: $\phi : \mathbb{R}^D \to \mathbb{R}^M$, and therefore $\phi(x) \in \mathbb{R}^M$. See Figure 2.16b for an example. Computation of the mapping function itself can be avoided by using the kernel trick.

The kernel function K is used to perform this operation, by taking the inner product between two observations x_i and x_j [140]:

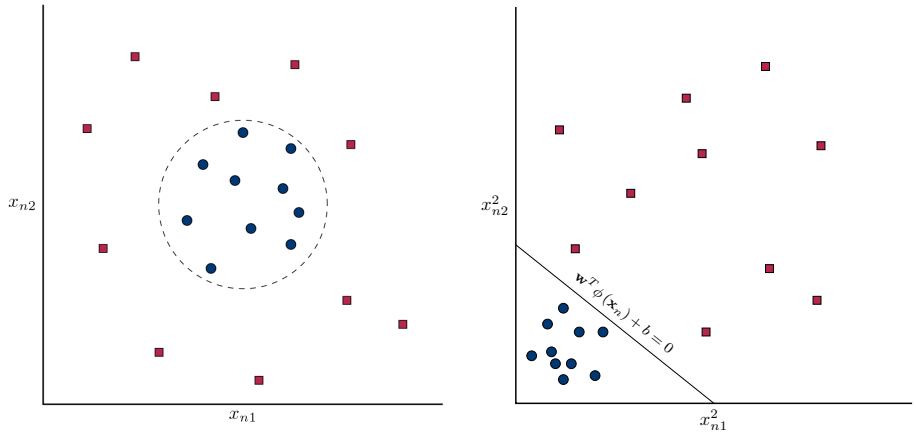
$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{2.20}$$

In the case of a linear SVM, the linear kernel function is used:

$$K(x_i, x_j) = x_i^T x_j + b \tag{2.21}$$

Hence the optimization problem changes into:

$$\begin{aligned} &\underset{w}{\operatorname{argmin}} \frac{||\mathbf{w}||^2}{2} + C \sum_{n}^{N} \xi_n \\ &\text{subject to } y_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n,\ \xi_n \geq 0 \end{aligned} \tag{2.22}$$



(a) Data are not linearly separable.

(b) Data become linearly separable when mapped onto a higher dimensional feature space.

Figure 2.16: Example non-linearly separable data (a) and how mapping the datapoints to a higher dimensional feature space can make the data linearly separable (b).

2.7 Preliminary Conclusions

The findings of the literature review allow for some preliminary conclusions to take into account for this thesis research. These preliminary conclusions will be given in this section.

First of all, it can be concluded that only a very limited amount of work has investigated how fear responses are reflected in fNIRS signals. Especially the use of fNIRS to measure fear responses to virtual heights was researched to a very limited extent.

Additionally, no previous research has investigated the use of fNIRS signals to detect fear responses using machine learning classifiers. However, there are previous works that focus on the detection of other mental states using fNIRS signals. From these studies, it can be concluded that the SVM classifier is often used for the detection of mental states based on fNIRS data. However, similar performances were also obtained using the LDA algorithm. The features and windows that were used to train and test the classifiers vary widely among the reviewed studies. However, in most of the cases the average Δ[HbO] and Δ[HbR] signals over windows ranging from 1 to 10 seconds were used as features.

Furthermore, previous work on the simultaneous use of immersive VR and fNIRS during experiments indicates that the combination is feasible, although the VR HMD sometimes limits the

fNIRS measurement possibilities. These studies were able to create experiments with high ecological validity, because the VR HMD can show a simulation of the real world to the participant.

Finally, previous work on the use of physiological signals to measure or detect fear responses to VR environments indicate the potential of this direction. Especially the detection of fear responses based on physiological signals shows promising results, with accuracies ranging from 76% to even 89.5% for two-class classification, and 42.5% to 88.77% for four-class classification.

Chapter 3

Method

This chapter describes the methods of data collection and data processing that were carried out in order to answer the research questions of this thesis. The first part of this chapter describes the experiment that was conducted in order to create a dataset. In the second part of this chapter, it is explained how the data generated with the experiment was (pre-)processed in such a way that results could be generated from it.

3.1 Data Collection

An experiment was conducted to collect fNIRS data from people with fear of heights and people without fear of heights, while they were experiencing virtual ground situations and virtual height situations. The goal of the experiment was to construct a dataset that contains fNIRS data measured from people with virtually induced fear of heights responses and people without those responses, as such a dataset could not be found in the available fNIRS literature.

3.1.1 Participants

The experiment was approved by the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science at the University of Twente (reference number: RP 2020-76). Every participant gave written informed consent before the experiment started. Two different groups of participants were recruited: participants with fear of heights (experimental group) and participants without fear of heights (control group). Potential participants who were interested in participating in the experiment were asked to fill out the Acrophobia Questionnaire (AQ), which is a self-report questionnaire about situations involving heights [142, 143]. The AQ consists of 20 items that should be rated on a seven-point Likert scale, ranging from *not anxious at all* to *extremely anxious*, see Appendix C. Similar to other works, the AQ scores were used to assess the potential participants' fear of heights [27–30, 75]. Participants were recruited as part of the experimental group if their AQ score was 35 or higher, and recruited as part of the control group if their AQ score was 20 or lower, similar to [30].

Table 3.1: Overview of the participant demographics of the experimental group and the control group.

Demographic	Experimental group	Control group
Females	9	9
Males	11	12
Mean AQ score (± SD)	52.40 (± 11.47)	9.71 (± 5.89)
Mean age (± SD)	26.10 (± 10.47)	22.95 (± 2.11)

A total of 20 participants were recruited to be part of the experimental group ($n_e = 20$) and 21 participants were recruited to be part of the control group ($n_c = 21$). Table 3.1 contains the demographic information of both groups. None of the participants suffered from anxiety disorders.

3.1.2 Instruments

3.1.2.1 VR

The VEs that were developed for this experiment were presented to the participants via the Oculus Rift S [144]. The Oculus Rift S is a VR HMD with 6 degrees-of-freedom (DoF) inside-out tracking. This means that the HMD tracks a person's head rotations and translations (forward/backward, left/right, up/down). Therefore, participants were able to look around in the VEs by simply rotating their head and to walk around by moving their body in the physical world. The movement possibilities are clarified by Figure 3.1.

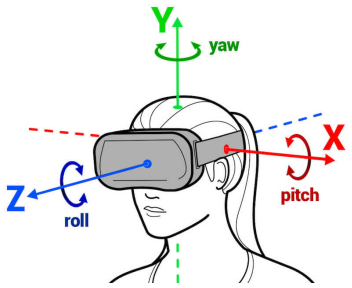


Figure 3.1: The movement possibilities offered by a 6 DoF HMD: translation and rotation along 3 axes [145].

3.1.2.2 fNIRS

Changes in HbO and HbR concentrations were measured using the Artinis Brite 24 [20]. The Brite is a wireless continuous wave fNIRS device that can measure up to 27 channels. The NI light is emitted at two nominal wavelenghts: 760 nm and 850 nm. Every participant's cortical hemodynamic response was measured at a sampling rate of 10 Hz. The optodes were arranged such that a large part of the PFC was covered, including the dlPFC, anterior PFC and part of the vlPFC. This choice was based on the outcomes of the literature review on detecting fear responses with fNIRS. Every emitter-detector pair had a maximum distance of 3 cm between the optodes. Figure 3.2 shows the positioning of the optodes and channels on the scalp, with an overview of the 10-20 system next to it as a reference.

3.1.3 Stimuli

The participants were alternately exposed to two different types of VEs, which are the stimuli of the experiment. The VEs were created using the Unity development platform [146]. In the first type of VEs, the participant is standing on a sidewalk or square in the middle of a city (ground condition), whereas the other type of VEs place the participant on the rooftop of a high building (height condition). All the VEs were accompanied by city sounds, to make the experience more immersive. The same audio was used for every VE. See Figure 3.3 for an overview of all the VEs that were presented to the participant.

It is hypothesized that the distinct types of VEs elicit different responses for the participants of the different groups. For the experimental group, it is hypothesized that the height condition evokes a fear response, while the ground condition would not evoke a fear response. This hypothesis is based on the assumption that the participants who are part of the experimental group have a moderate fear of heights. For the control group, it is hypothesized that the ground condition nor

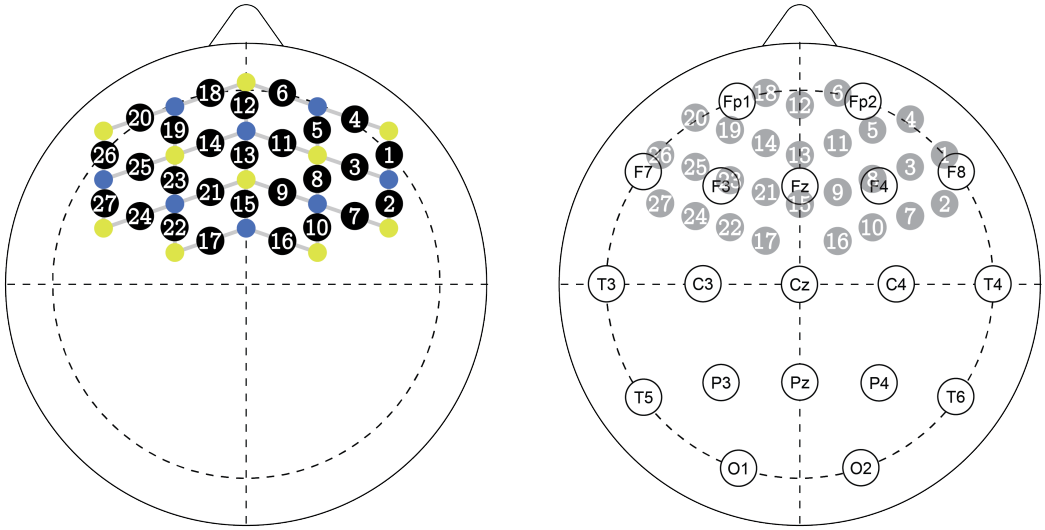


Figure 3.2: The positioning of the optodes on the scalp during the experiment. On the left are the detectors (blue) and emitters (yellow) with the channels indicated by a circle with a number in it. On the right, the channels are projected on the layout of the 10-20 system, as a reference.

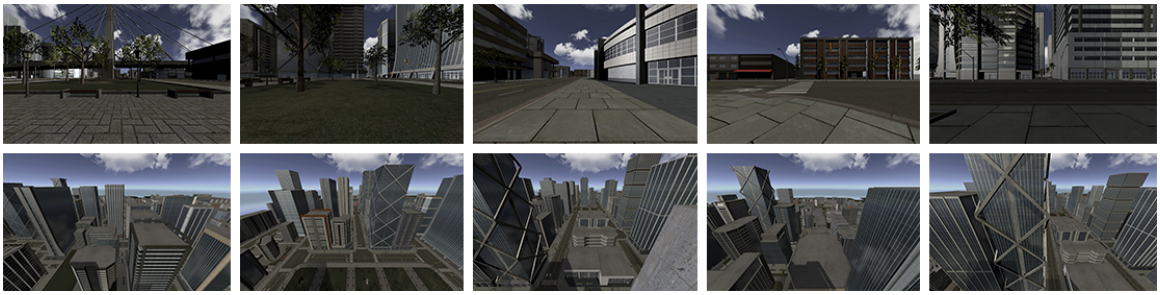


Figure 3.3: The VEs of the ground condition (top) and height condition (bottom).

the height condition evoke a fear response, as the participants who are part of this group have little to no fear of heights.

3.1.4 Experimental Procedure

Before the experiment started, the participant was asked if he/she has a fear of heights. This was done as an extra step, next to the AQ scores, to verify that the participants of the experimental group had a fear of heights and that the participants of the control group did not have a fear of heights.

After the participant answered the question, the participant was instructed about the usage of the Oculus Rift S. The researcher demonstrated how the Oculus Rift S should be adjusted to fit the head and how the hand-held controllers should be held. Participants could not use the hand-held controllers to interact with the VR application. However, the controllers were needed to make the tracking of the Oculus Rift S more reliable, hence every participant was asked to hold the controllers during the experiment. Once the explanation was given, there was a practice round during which the participant saw an example VE, which was similar to the ground type VE. This was done to make the participant familiar with the VE, the HMD, and the hand-held controllers. The practice round ended when the participant indicated that he/she felt familiar enough.

After the practice round, the Oculus Rift S was removed from the participant's head and the fNIRS cap was put on. The researcher used a narrow, oblong tool to move the participant's hair to

Figure 3.4: Participant wearing the fNIRS headcap and the VR HMD during the experiment.

the side in case it got in between an optode and the participant's skin. This was done to prevent the hair from absorbing the NI light. When the raw NI light intensities were in the appropriate range, the researcher checked if there was a heartbeat visible in the signals. This is a way to verify that the optodes make good contact with the scalp and that they are measuring hemodynamic activity [49]. As soon as this final check was performed successfully, the participant was asked to stand in a designated place and to put the Oculus Rift S on his/her head. In order to do so, the straps of the Oculus Rift S were loosened as much as possible, such that they would not interfere with the optodes of the headcap while putting the HMD on the participant's head. Figure 3.4 shows a participant wearing both the fNIRS headcap and the VR HMD.

When everything was set up correctly, the experiment started. As said before, there were two conditions: the ground condition and the height condition. The experiment consisted of 5 trials for each condition, which makes 10 trials in total. Every trial lasted for 30 seconds. Although the Oculus Rift S provides the possibility to walk around in the VE, the participant was instructed not to do so, to prevent severe motion artifacts. Instead, the participant was asked to gently look around in the VE, while preventing large head movements. Additionally, the participant was allowed to bend forward slightly during the height condition, if he/she felt comfortable enough to do so. The participant was also asked to step back into his/her original position after bending forward, to prevent problems due to lack of space in the experimental room. The ground condition and height condition were alternately shown to the participant, with every time a baseline period in between. During the baseline period, there was no stimulus, hence the VE turned entirely black and the audio stopped playing. The participant was instructed to relax and not think about anything during the baseline period. See Figure 3.5 for an overview of the experimental design.

After the experiment, the participant was asked to rate his/her perceived feelings of distress or fear using the Subjective Units of Distress Scale (SUDS), similar to [15, 75, 126, 128]. The SUDS is a self-report 11-point Likert scale ranging from 0 (*no distress/anxiety*) to 100 (*worst distress/anxiety that you have ever felt*) [147] that is often used to assess exposure settings during cognitive behavioral treatment [148]. The participant was asked to give two SUDS ratings: one for the ground condition VEs and one for the height condition VEs. The participant was asked to give the ratings after the

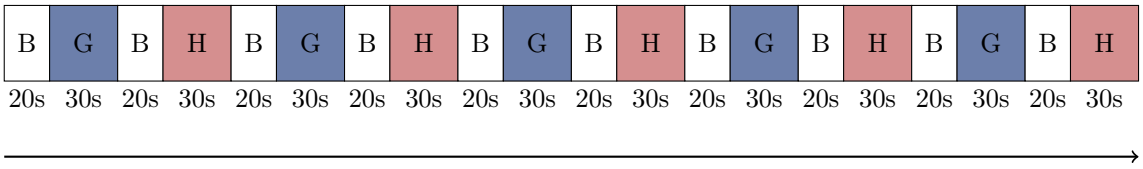


Figure 3.5: The experimental design. B = baseline, G = ground condition, H = height condition.

experiment to ensure that he/she would not be distracted from the VEs during the experiment. See Appendix C for the SUDS.

Additionally, the participant was asked to fill out the IGroup Presence Questionnaire (IPQ), which measures a person's sense of presence in VR [149]. The IPQ consists of 14 items that should be rated on a seven-point Likert scale. The items cover three subscales (*spatial presence*, *involvement*, and *realism*) and there is one additional item (*sense of being there*) that does not belong to a subscale. Table 3.2 gives an overview of the subscales and what they measure. Because presence and fear responses in VR are often positively correlated [150, 151], it is expected that a fear response can only be evoked in VEs where the person feels present. Therefore, the IPQ scores were taken to test if the participants felt present enough in the VEs for a fear response to occur.

Table 3.2: The IPQ subscales and what they measure [152].

Subscale	Measures	Items
Spatial presence	The sense of being physically present in the VE	2 - 6
Involvement	The attention devoted to the VE and the involvement experienced	7 - 10
Realism	The subjective experience of realism in the VE	11 - 14

After filling out the IPQ, the participant was asked to fill out the AQ once more. This was done to test if the participant's AQ scores prior to and after the experiment were similar. Since the AQ has adequate test-retest reliability (median r = 0.82) [142], it would be expected that every participant's pre-experiment and post-experiment AQ scores are similar. The post-experiment AQ scores were taken to check if participants would still belong to their respective group (experimental group or control group) after the experience of the experiment.

Finally, a structured interview was held to ask every participant about their experiences during the experiment, see Appendix D. During the interview, the participant was asked if he/she felt any feelings of fear during the experiment and, if yes, at which moments. Additionally, the participant was asked if he/she felt any other emotions throughout the experiment and again, at which moments. The main goal of the interview was to have an extra measure of every participant's fear of heights, next to their AQ scores.

3.2 Data Processing

The data that was collected with the experiment was used to answer the research questions of this thesis. Participants with fear of heights responses (experimental group) and participants without fear of heights responses (control group) were selected based on the questionnaire data that was collected during the experiment. The fNIRS data of these participants was pre-processed and the HR was extracted from it, which served as an additional physiological measurement. Statistical analyses and classification of fear responses were performed using the pre-processed fNIRS and HR data.

3.2.1 Questionnaires

For every participant, the AQ scores were computed by taking the sum of every item. Higher AQ scores indicate more severe fear of heights. Based on the pre-experiment AQ scores, potential

participants who belonged to the control group (AQ score < 20) or experimental group (AQ score ≥ 35) were invited to participate in the experiment. The post-experiment AQ scores were computed in the same manner and compared against the threshold values for both groups, to determine if every participant still belonged to their respective group after experiencing the experiment.

The SUDS scores were taken as the absolute values that the participants indicated. Similar to the research of Balan et al. [15], the SUDS scores between 0 and 30 were labeled as "relaxation" and the SUDS scores above 30 were labeled as "fear". Based on the SUDS scores and their accompanying labels, it was decided whether a participant felt relaxed or anxious during the ground trials and the height trials.

Before the IPQ scores were computed, individual scores on items 3, 4, 9, and 11 had to be inverted, to ensure that every item ranged from *negative/not present* to *positive/present*. Once this was done, the IPQ scores were computed by taking the averages of every item. Both the overall averages as well as the averages per subscale were computed. Since a score of 3 is the neutral option, an average score above 3 indicates a stronger feeling of presence. The higher the average score, the more present the participant felt.

3.2.2 Participant Selection

Before the experiment started, participants were selected to be part of either the control group (AQ score < 20) or the experimental group (AQ score ≥ 35), based on their pre-experiment AQ scores. The pre-experiment AQ scores therefore served as an indicator of a person's fear of heights and expected fear responses during the experiment. However, the results of the questionnaires that were taken during the experiment were used to determine in a more reliable way if participants were actually experiencing a fear of heights response (experimental group) or not (control group). To this end, the scores of the post-experiment AQ, the SUDS, and the IPQ were used.

The post-experiment AQ scores were used to determine if participants would still fall within their respective group after experiencing the experiment. Therefore, the threshold values for the post-experiment AQ scores are identical to those of the pre-experiment AQ scores. The SUDS scores of the height trials were used to determine if a participant felt relaxed or scared during the height trials. Participants who felt relaxed (no fear response) will be part of the control group, whereas participants who felt scared (fear response) will be part of the experimental group. Finally, the IPQ presence score of every participant should be 3 or higher to be part of one of the groups. IPQ presence scores lower than 3 indicate that the participant did not feel present in the VEs, hence no realistic fear or relaxation responses could be induced in that case. Therefore, participants who did not feel present will be excluded from the analysis, to avoid possible discrepancies due to this matter. Table 3.3 gives a complete overview of the post-experiment selection criteria that were used to determine if a participant suited the control group or the experimental group. If a participant did not suit any of the groups, the data of the participant was disregarded from the analyses.

Table 3.3: The post-experiment selection criteria for the control group and the experimental group, based on the post-experiment AQ scores, SUDS scores, and IPQ scores.

Control group	Experimental group
Post-experiment AQ < 20	Post-experiment AQ ≥ 35
SUDS height trials ≤ 30 ("relaxation")	SUDS height trials > 30 ("fear")
Average IPQ ≥ 3	Average IPQ ≥ 3

3.2.3 fNIRS Pre-Processing

A schematic overview of the fNIRS pre-processing pipeline is given in Figure 3.6. The first step is to convert the raw time series to Δ[HbO] and Δ[HbR] signals. The fNIRS data was recorded using the Artinis Oxysoft software, after which the *Oxysoft2Matlab* script was used to export the

Δ[HbO] and Δ[HbR] signals to Matlab. Before the time series were pre-processed further, they were visually inspected. Channels with severe motion artifacts and channels that did not show cardiac oscillations were excluded from further analysis. Motion correction was applied to the remainder of the channels, using the Temporal Derivative Distribution Repair (TDDR) procedure of Fishburn et al. [153]. TDDR takes the temporal derivative of the time series to compute the fluctuations in the signals. It estimates observation weights for each time point and leverages this to remove spikes and baseline shifts from the signal. The TDDR procedure was validated on simulations and on an empirical experiment, and yielded better activation detection performance than five other methods that are often used for motion correction, which are Correlation-Based Signal Improvement, Movement Artifact Reduction Algorithm, Targeted PCA, Kurtosis Wavelet, and Spline Savitzky-Golay. See Appendix E for a step-by-step description of the algorithm and an example of a signal that was corrected using TDDR.

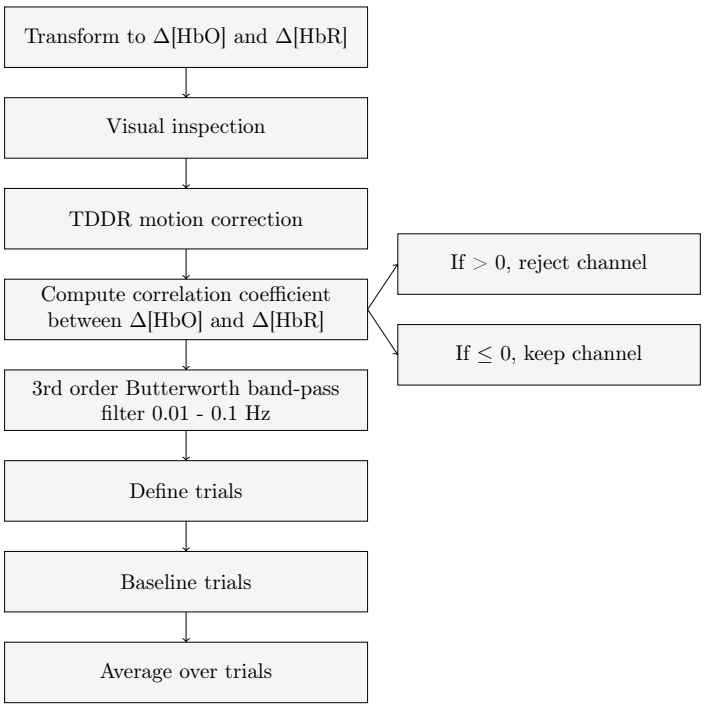


Figure 3.6: The fNIRS pre-processing pipeline.

After motion correction, the correlation coefficients of every channel's Δ[HbO] and Δ[HbR] signal were calculated. Channels with a positive correlation coefficient (correlation coefficient > 0) were removed. This decision is based on the work of Cui et al. [154], who found that Δ[HbO] and Δ[HbR] are negatively correlated when the amount of motion artifacts in the signals is low. The more motion artifacts are introduced in the signals, the more positive the correlation becomes. Therefore, the positive correlation between the Δ[HbO] and Δ[HbR] signals can be used to automatically identify bad channels.

Once the bad channels were identified and excluded from the dataset, the signals were filtered. A 3rd order Butterworth band-pass filter with low cut-off frequency 0.01 Hz and high cut-off frequency 0.1 Hz was used, as it is one of the most frequently used filters for fNIRS data [19, 40]. The cut-off frequencies were chosen such that the physiological noise arising from breath cycles ($\sim$ 0.2 - 0.3 Hz), cardiac cycles ($\sim$ 1 Hz), and Mayer Waves ($\sim$ 0.1 Hz) is mostly removed.

The filtered signals were epoched and baselined. The epoch duration was set from 0 to 30 seconds after the trial onset, such that one epoch contains the data of an entire trial. The 5 seconds before the epoch were used as a baseline period. A total of 10 epochs of equal length were extracted: 5 for the ground condition and 5 for the height condition. For every participant, the grand averages over the epochs for each condition were computed, for every channel separately. This resulted into a

grand average ground condition signal and a grand average height condition signal per channel per participant.

3.2.4 Heart Rate Extraction

Since the fNIRS data is contaminated by cardiac cycles, it is possible to extract the HR from the Δ[HbO] signals [155–158]. HR extracted from fNIRS signals was shown to correlate to a high degree with the more traditional HR extracted from electrocardiography ($r > 0.90$) [155, 157].

As a first step, the motion-corrected Δ[HbO] time series were filtered using a Butterworth band-pass filter between 1 Hz and 1.9 Hz, to remove the components from the signal that are not related to the cardiac cycles [157, 158]. Then, the HR peaks in the signal are detected using Matlab's *findpeaks()* peak detection algorithm [155]. See Figure 3.7 for an example of the filtered signal of a trial with the detected HR peaks in it.

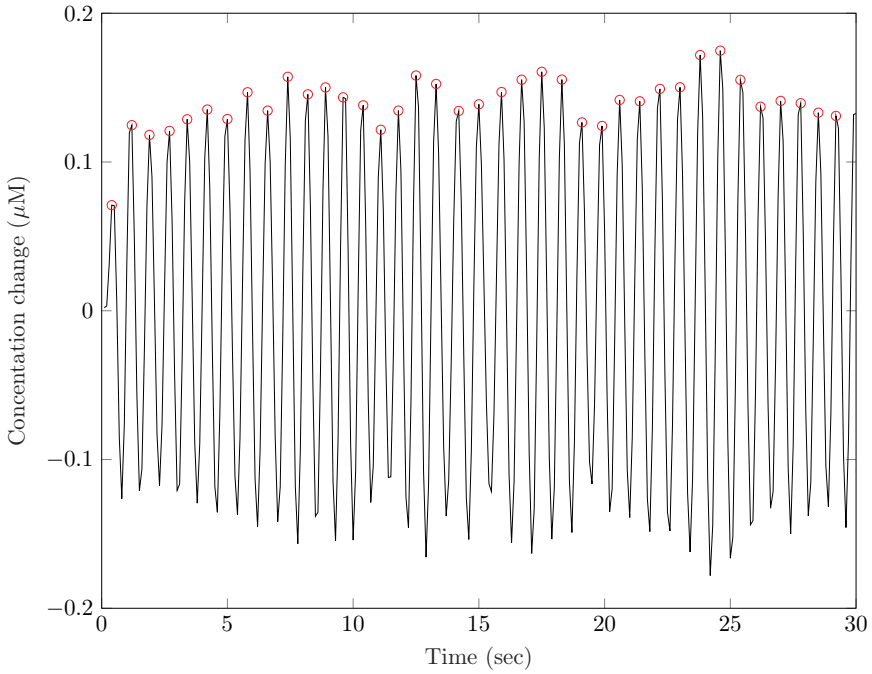


Figure 3.7: Example of the filtered signal (black) and the detected HR peaks (red circles) over a time interval of 30 seconds.

The HR in beats per minute (BPM) over a certain time interval was calculated using the total number of peaks detected during that time interval:

$$\text{BPM} = \frac{p_{num}}{t/60} \tag{3.1}$$

Where p_{num} is the number of detected peaks and t is the total time of the time interval over which the peaks are detected in seconds. Since there is a Δ[HbO] signal for every measurement channel, the calculated HRs were averaged over the channels, such that a final HR value was obtained [157]. This entire procedure was executed per trial, such that an average HR in BPM was computed for every trial of every participant.

3.2.5 Statistical Analysis

Statistical tests were performed on the fNIRS data and the extracted HR data to determine whether there are significant differences between the data of the control group and the experimental group (between-group analysis) and to determine whether there are significant differences between the data of the ground trials and the height trials of the experimental group (within-group analysis). The goal of the between-group analysis is to investigate if the physiological measures taken from participants who had a fear response are indeed significantly different from those taken from participants who did not have a fear response. The goal of the within-group analysis is to determine whether the physiological measures taken during fear responses (i.e. during height trials) are indeed significantly different from those taken during relaxation (i.e. during ground trials).

3.2.5.1 Between-Group Analysis

The results of the fNIRS pre-processing phase (see section 3.2.3) contain the grand average ground condition signals per channel per participant and the grand average height condition signals per channel per participant. In order to simplify the between-group statistical analysis and to deal with the inter-personal differences in the fNIRS signals, the contrast between the ground condition and the height condition grand average Δ[HbO] signals and Δ[HbR] signals for all channels and all participants were computed. The contrast was computed by subtracting the grand average ground condition signal from the grand average height condition signal. This resulted into $p \cdot c$ grand average contrast signals, where p is the amount of participants and c the amount of channels. For all of these signals, the mean over the window from 3 to 15 seconds post-stimulus onset was computed. The 3-15 second window was chosen because the hemodynamic response only starts to become visible after 3 seconds (2.8-second lag was found [159]), and because it was found that the hemodynamic response is most intense in the first 5 to 17 seconds after the stimulus onset [160]. Therefore, this window seems to be the most appropriate in order to analyze if there is a significant difference between the grand average contrast signals of both groups. For every channel, a permutation test with 50,000 permutations was used to test for significant differences between the contrast signal means over the 3-15 second window of the control group and the experimental group, at the significance level $\alpha = 0.05$. The sample means were used as the test statistic. These analyses were executed once for the Δ[HbO] signals and once for the Δ[HbR] signals.

Another between-group statistical analysis was performed on the HR data. The results of the HR extraction procedure (see section 3.2.4) contain the average HR of every ground trial and of every height trial per participant. From these values, the average HR over the ground condition and average HR over the height condition were computed for every participant, by taking the average of the HR over the different trials of the condition of interest. Similar to the fNIRS between-group analyses, the contrast in the HR values between the ground condition and the height condition were computed. This was done by subtracting the ground condition HR value from the height condition HR value, for every participant. This resulted into a list of contrast HR values for both groups. A permutation test with 50,000 permutations was used to test for significant differences between the contrast HR values of both groups, at the significance level $\alpha = 0.05$. The sample means were used as the test statistic.

3.2.5.2 Within-Group Analyses

The within-group statistical analysis of the fNIRS data was performed on the data of the ground trials (i.e. relaxation trials) and the height trials (i.e. fear trials) of the experimental group. Similar to the between-group analysis, the grand average ground condition Δ[HbO] and Δ[HbR] signals and the grand average height condition Δ[HbO] and Δ[HbR] signals were averaged over the 3-15 second window. For every channel, a permutation test with 50,000 permutations was used to test for significant differences between the ground trial means and the height trial means over the 3-15 second window, at the significance level $\alpha = 0.05$. The sample means were used as the test statistic. Also this analysis was performed once for the Δ[HbO] signals and once for the Δ[HbR] signals.

Another within-group statistical analysis was performed on the HR data of the experimental group, which consists of the average HR during every ground trial and every height trial per participant. It must be noted that HR is a very personal measurement and that inter-personal differences in resting state HR can be as large as 70 BPM [161]. This could impact the statistical analysis of the differences in average HR between conditions within the same group. In order to minimize the effect of inter-personal differences in resting state HR, the average HR values of every trial were baseline-corrected by subtracting the average HR over the 10-second pre-stimulus baseline period. From these baseline-corrected HRs, the averages over the ground condition and height condition were computed. A permutation test with 50,000 permutations was used to test for significant differences between the baseline-corrected ground condition HRs and the baseline-corrected height condition HRs of the participants of the experimental group, at the significance level $\alpha = 0.05$. Again, the sample means were used as the test statistic.

3.2.5.3 Correction for Multiple Comparisons

The fNIRS data that was collected with the experiment contains measurements of 27 channels per participant. Therefore, the permutation tests of a single chromophore are executed a total of 27 times on the fNIRS data, once for every channel. As was explained in sections 3.2.5.1 and 3.2.5.2, a total of 4 statistical analyses were executed on the fNIRS data (between-group analysis on the Δ[HbO] data, between-group analysis of the Δ[HbR] data, within-group analysis of the Δ[HbO] data, and within-group analysis of the Δ[HbR] data). This results into a total of $4 \cdot 27 = 108$ hypothesis tests. Two additional hypothesis tests were executed on the HR data, one for the between-group analysis and one for the within-group analysis, which makes a total of 110 hypothesis tests. This causes the multiple comparisons problem, as was explained in section 2.6.1.2. Therefore, FDR correction was executed on the 110 p-values that resulted from the statistical analyses to correct for multiple comparisons. The rate q was set to $q = 0.05$.

3.2.6 Classification

This section describes the classification of fear of heights responses. Two different linear classification algorithms were used in order to perform the classification: LDA and SVM. This research uses the Matlab 2020a implementation of both algorithms, with the standard hyperparameter settings. See Appendix G for an overview of the hyperparameters of the classifiers. It was decided to use the standard hyperparameters because optimization did not have much influence on the classification outcomes and would yield a different set of hyperparameters for every participant's model. The following sub-sections will discuss the feature extraction for the classifiers and how subject-dependent and subject-independent classifiers were trained and tested.

3.2.6.1 Feature Extraction

The gathered fNIRS data consists of Δ[HbO] and Δ[HbR] measurements of 27 channels sampled at 10 Hz, hence there is a measurement for every 0.1 seconds. Similar to the statistical analyses, the time window of 3-15 seconds post-stimulus was used in the feature extraction process.

Based on the within-group statistical analyses, it was determined which fNIRS channels show significant differences between the ground (i.e. fear) condition and the height (i.e. no fear) condition. Only the data of the significant channels was extracted, as the statistical analyses indicated that those channels have the most potential to discriminate between fear and no fear. Due to movement artifacts and hardware malfunctions, data from different amounts of channels were available for the different participants. Therefore, it was decided to average the extracted Δ[HbO] and Δ[HbR] measurements over the significant channels for every participant. The Δ[HbO] and Δ[HbR] averaged over the significant channels will be referred to as $\overline{\Delta[\text{HbO}]}$ and $\overline{\Delta[\text{HbR}]}$, respectively. Then, the averages of $\overline{\Delta[\text{HbO}]}$ and $\overline{\Delta[\text{HbR}]}$ over 1-second windows were computed. This was done by taking the average of every 10 consecutive values of $\overline{\Delta[\text{HbO}]}$ and $\overline{\Delta[\text{HbR}]}$, similar to the procedure of [91]. Short histories of the averaged $\overline{\Delta[\text{HbO}]}$ and $\overline{\Delta[\text{HbR}]}$ signals of every second were also computed, such that the information arising from the changes in the signal over time could be utilized for

classification, similar to [93]. An x amount of seconds was defined to be the short history. For every observation, the current observation and the observations of the x amount of seconds before the current observation were extracted (which made a total of $x + 1$ features for each chromophore (Δ[HbO] and Δ[HbR]) for every observation (second) in the dataset). In order to investigate the effect of the short histories, the classifiers were trained and tested on three different short histories similar to [93]: 1 second, 3 seconds, and 5 seconds.

3.2.6.2 Subject-Dependent Classifiers

Subject-dependent classifiers were trained and tested on the data of every single subject from the experimental group. Hence, a certain subject-dependent classifier is specific to that subject and cannot be used for other subjects. Only the data of the experimental group was used for these classifiers, as there is a clear separation between fearful trials (i.e. height trials) and non-fearful trials (i.e. ground trials) for this group, which can be used to classify between a fear of heights response and no fear of heights response. Therefore, the ground trials were labeled as "no fear response", whereas the height trials were labeled as "fear response", and the goal of the classifiers was to classify the available datapoints into the correct class.

In every participant's dataset, there were a total of 5 ground trials and 5 height trials. The data was divided into 60% training data and 40% testing data. The training data was used to train the classifiers, whereas the testing data was used to test the performance of the classifiers. To this end, the first 6 trials (consisting of 3 ground trials and 3 height trials) were part of the training data, whereas the final 4 trials (consisting of 2 ground trials and 2 height trials) were part of the testing data. The performance of each classifier on the test data was measured by the classifier's accuracy. The accuracy was calculated as the percentage of correctly classified observations in the test dataset. Please note that for every participant, six different subject-dependent classifiers were trained and tested: the LDA classifier and the SVM classifier, both on the 1-second, 3-second, and 5-second history.

3.2.6.3 Subject-Independent Classifiers

The subject-independent classifiers were trained on the data of all the participants of the experimental group minus one, and tested on the data of the final participant. This procedure is similar to k-fold cross validation, where k equals the amount of participants in the experimental group [91]. Again, the data from the ground trials were labeled as "no fear response", whereas the data from the height trials were labeled as "fear response". Both an LDA classifier and SVM classifier were trained for every participant on the 1-second, 3-second, and 5-second history. The performance of every classifier was measured by the classifier's accuracy on the test data again.

The idea behind these classifiers is that the data of the other participants can be used to train a classifier that can be used to classify unseen data from an unknown participant. This could be useful in real-life VRET settings, where training a classifier on the data of every single person would be an exhaustive process. Instead, it would be easier if the data of others could be used by the classifier to make informed decisions about the measurements of a new person [15].

Chapter 4

Results

This chapter contains the results that were generated based on the method described in the previous chapter. First, the selection of participants based on AQ scores, SUDS scores and IPQ scores will be described. Then, the results of the between-group statistical analyses and the within-group statistical analyses of the experimental group will be given. The final section describes the results of the subject-dependent classifiers and the subject-independent classifiers.

4.1 Participant Selection

Three participants did not complete the experiment due to motion sickness caused by the VR application, hence their data was excluded. Based on the post-experiment selection criteria (see Table 3.3) a total of 15 participants ($n_c = 15$) were selected to be part of the control group and 14 participants ($n_e = 14$) were selected to be part of the experimental group. All the results described in this chapter are based on the analyses of the data of these participants. Table 4.1 gives the mean scores and standard deviations of the questionnaire results for both groups. These results show that there are clear distinctions between the two groups in terms of AQ scores (pre-experimental as well as post-experimental, see Appendix J for an overview) and SUDS of the height condition. The two groups are similar in terms of SUDS of the ground condition and the experienced presence in the VEs.

Table 4.1: Mean scores and standard deviations of the questionnaire results for the control group and the experimental group.

Questionnaire	Control group mean (± SD)	Experimental group mean (± SD)
Pre-experiment AQ	10.80 (± 5.66)	56.07 (± 11.20)
Post-experiment AQ	10.73 (± 6.41)	50.36 (± 11.85)
SUDS ground condition	3.00 (± 4.55)	6.43 (± 6.02)
SUDS height condition	11.53 (± 8.08)	69.86 (± 11.55)
IPQ presence	4.20 (± 0.86)	4.21 (± 0.97)

4.2 Statistical Analysis

The statistical analyses that were performed can be divided into between-group analyses and within-group analyses. The latter is only executed on the data of the experimental group. This section is divided into two sub-sections, that each cover one of the analyses types.

4.2.1 Between-Group Analysis

The between-group statistical analyses were performed separately on the Δ[HbO] signals, the Δ[HbR] signals, and the HR extracted from the Δ[HbO] signals. The results of each of these between-group statistical analyses are given below.

4.2.1.1 fNIRS Oxyhemoglobin

The first between-group statistical analysis was performed on the Δ[HbO] data of the control group and the experimental group. Figure 4.1 shows the grand average Δ[HbO] traces of the contrast between the ground condition and the height condition for the two groups for every channel, with the standard error given around every trace. The 3-15 second window over which the mean values were calculated that were used in the permutation tests is indicated by the gray shaded areas in the graphs. The graphs are arranged according to the optode layout that was used during the experiment, as presented in Figure 3.2.

In almost all of the graphs, it can be seen that the grand average contrast trace of the experimental group exceeds that of the control group during the 3-15 second window, hence there seems to be a stronger contrast between ground trials and height trials for the experimental group. However, the outcomes of the FDR-corrected permutation tests indicate that there is only a significant difference between the contrast Δ[HbO] means of both groups in channel 3 ($p = 0.0008$). Appendix F provides a visual representation of the FDR threshold based on the total amount of p-values generated from the statistical tests. From this visual representation, it can be seen which p-values survive the FDR correction.

4.2.1.2 fNIRS Deoxyhemoglobin

The second between-group statistical analysis was performed on the Δ[HbR] data of the control group and the experimental group. Figure 4.2 shows the grand average Δ[HbR] traces of the contrast between the ground condition and the height condition for the two groups for every channel, with the standard error given around every trace. Again, the arrangement of the graphs is according to the optode layout used during the experiments and the 3-15 second window is indicated by the gray shaded areas.

The differences in the grand average contrast Δ[HbR] traces are less apparent in this case than they were for the Δ[HbO] traces. However, there are some channels (i.e. channels 15, 23, 24, and 27) where the grand average trace of the control group behaves somewhat differently than that of the experimental group during the 3-15 second window. However, the outcomes of the FDR-corrected permutation tests show that there are no channels with significant differences between the contrast Δ[HbR] means of the control group and the experimental group.

4.2.1.3 Heart Rate

The third between-group statistical analysis was performed on the HR data (extracted from the Δ[HbO] signals) of the control group and the experimental group. Figure 4.3 shows a box plot of the average contrast HR (of the contrast between the ground condition and the height condition) for both groups separately. As can be seen from this box plot, the median of the contrast HR of the experimental group is positive, while it is negative for the control group. This indicates that, on average, the HR of the participants of the experimental group increased between ground trials and height trials, whereas it decreased for the participants of the control group. However, when taking the range of the box plots into account, it can be seen that the positive and negative changes in average HR between conditions are not the case for every participant in every group. The ranges of the box plots indicate that some participants of the control group experienced increased HR values when going from a ground trial to a height trial, while some participants of the experimental group have experienced decreased HR values in this case.

Based on the permutation test at the predefined significance level $a = 5\%$, the average contrast HRs of the two groups are not significantly different ($p = 0.051$). Although the resulting p-value

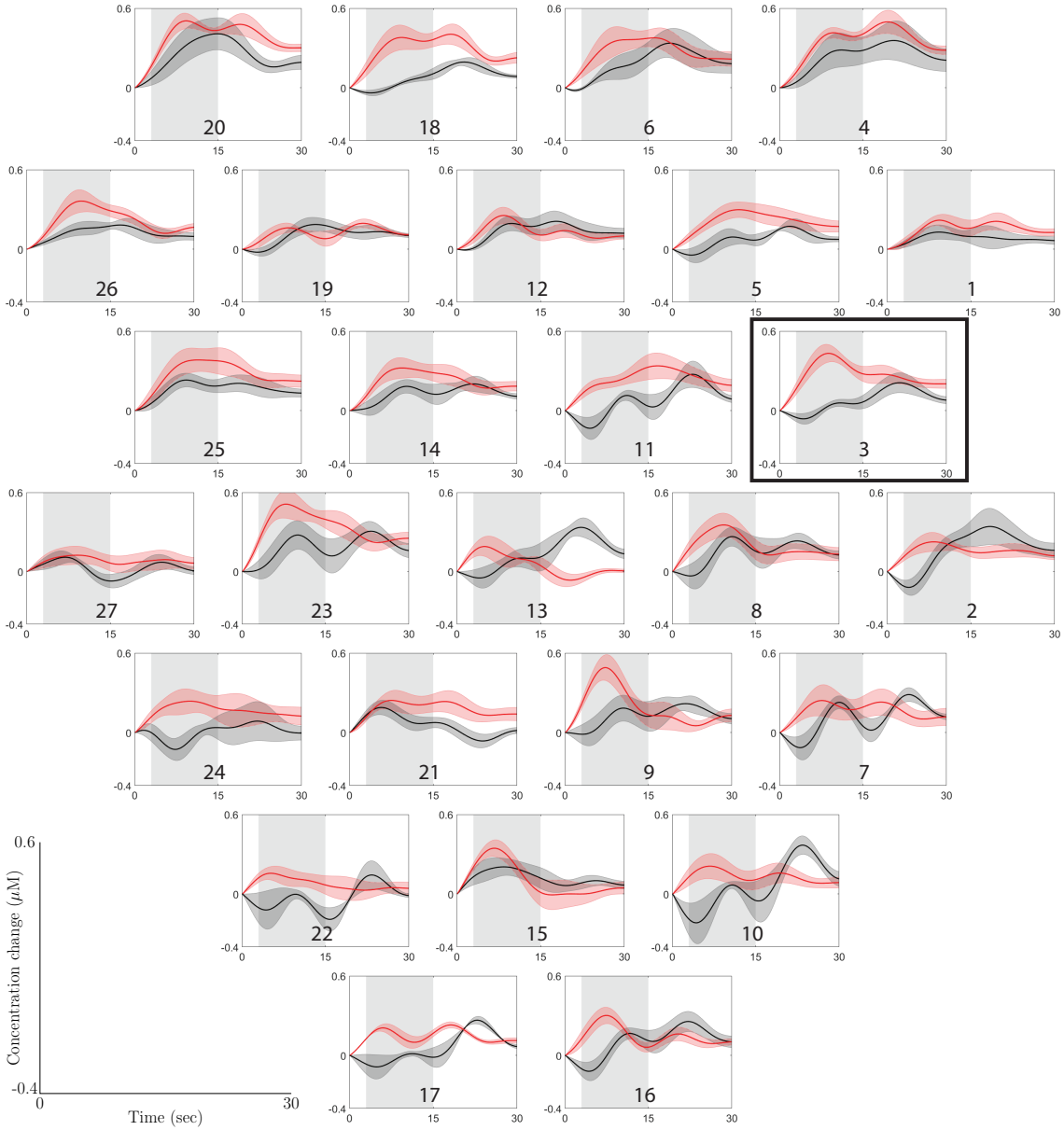


Figure 4.1: Grand average Δ[HbO] traces of the contrast between ground condition and height condition for the two groups: control group (black traces) and experimental group (red traces). Around the traces are the standard errors. The gray shaded area is the window over which the means were taken that were used for the permutation tests. On the x-axis is time (seconds), ranging from 0 to 30. On the y-axis is concentration change (μM), ranging from -0.4 to 0.6. Channel numbers are given in every plot. The plot with the border around it shows the channel where a significant difference was found between the means of the control group and the experimental group.

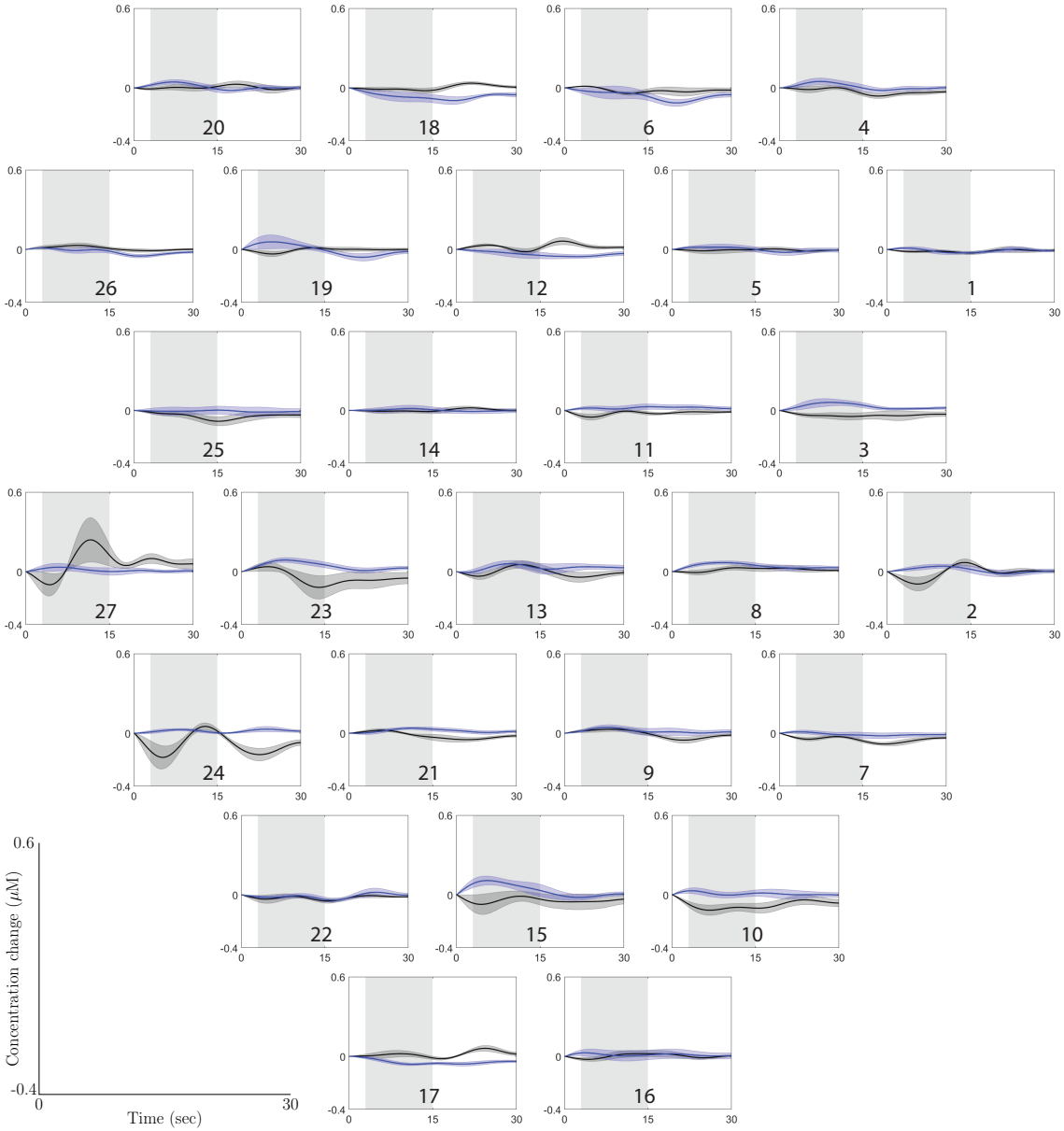


Figure 4.2: Overview of the grand average Δ[HbR] traces of the contrast between ground condition and height condition for the two groups: control group (black traces) and experimental group (blue traces). Around the traces are the standard errors. The gray shaded area is the window over which the means were taken that were used for the permutation tests. On the x-axis is time (seconds), ranging from 0 to 30. On the y-axis is concentration change (μM), ranging from -0.4 to 0.6. Channel numbers are given in every plot.

is very close to a significant result in the case of a single hypothesis test, this result is far from the threshold that is needed to survive the FDR-correction for multiple testing, as can be seen in Appendix F.

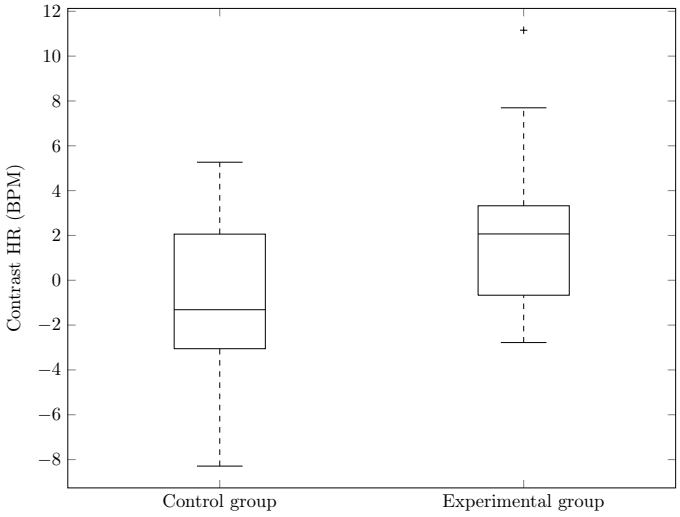


Figure 4.3: Box plot of the average contrast HR in BPM of the control group (left) and the experimental group (right).

4.2.2 Within-Group Analysis

The within-group statistical analyses were performed on the data of the experimental group. Similar to the between-group statistical analyses, the within-group statistical analyses were performed separately on the Δ[HbO] signals, the Δ[HbR] signals, and the HR extracted from the Δ[HbO] signals. The results of each of these within-group statistical analyses are given below.

4.2.2.1 fNIRS Oxyhemoglobin

The first within-group statistical analysis was performed on the Δ[HbO] data of the experimental group. Figure 4.4 shows the grand average Δ[HbO] traces of the ground condition and the height condition for the experimental group, with the standard error given around every trace. Again, the gray shaded area shows the 3-15 second window over which the mean values were calculated that were used for the permutation tests. Also, the graphs are arranged according to the optode layout used during the experiments again.

In all of the graphs, there is a clear visual difference between the grand average traces of the ground condition and of the height condition, where the grand average traces of the height condition exceed those of the ground condition. This effect is especially visible during the 3-15 second window. This indicates that, in general, the participants of the experimental group experienced increased Δ[HbO] values during the height (i.e. fearful) condition. On the contrary, their Δ[HbO] values decreased or remained somewhat constant during the ground (i.e. non-fearful) condition. According to the outcomes of the FDR-corrected permutation tests, there are significant differences between the means of the ground condition and the height condition of the experimental group in the following channels: 1 ($p = 0.0022$), 2 ($p = 0.0022$), 3 ($p = 0.00001$), 4 ($p = 0.0003$), 6 ($p = 0.0022$), 11 ($p = 0.0016$), 12 ($p = 0.0041$), 14 ($p = 0.0016$), 18 ($p = 0.0009$), 20 ($p = 0.00002$), 23 ($p = 0.00002$), 25 ($p = 0.0001$), and 26 ($p = 0.0022$).

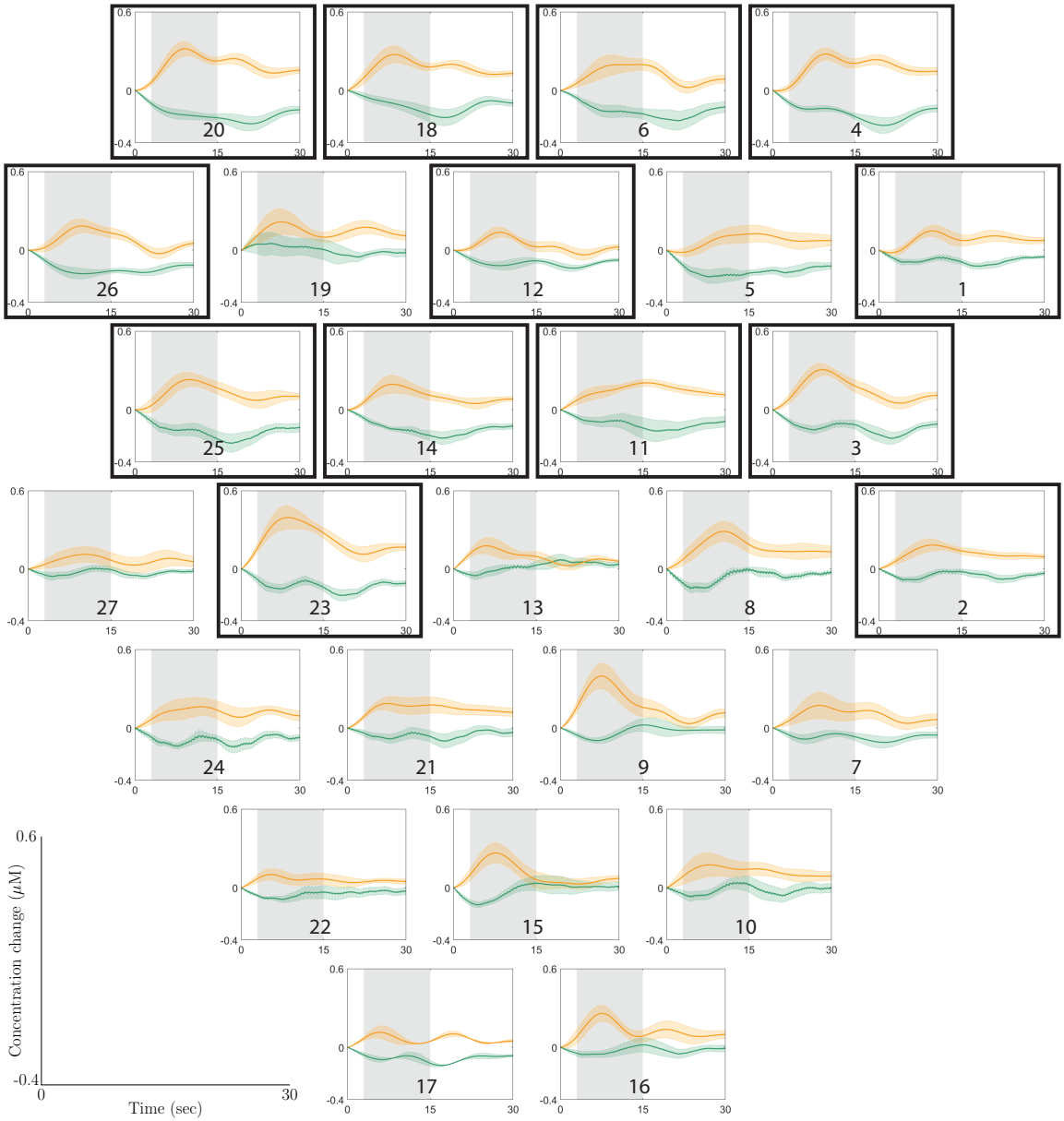


Figure 4.4: Overview of the grand average Δ[HbO] traces of the ground condition (green traces) and height condition (orange traces) of the experimental group. Around the traces are the standard errors. The gray shaded area is the window over which the means were taken that were used for the permutation tests. On the x-axis is time (seconds), ranging from 0 to 30. On the y-axis is concentration change (μM), ranging from -0.4 to 0.6. Channel numbers are given in every plot. The plots with the border around it show the channels where a significant difference were found between the means of the ground condition and the height condition for the experimental group.

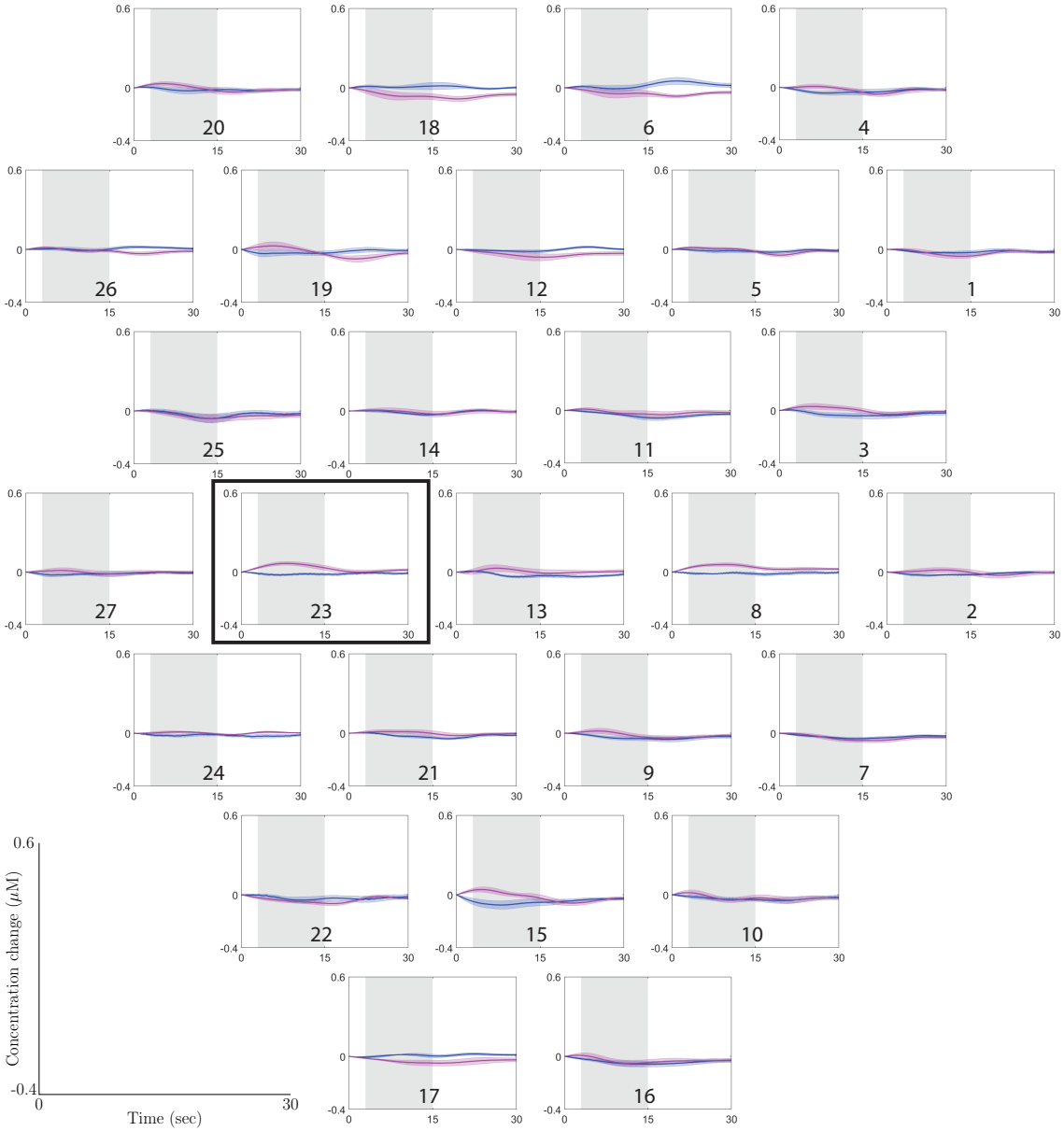


Figure 4.5: Overview of the grand average Δ[HbR] traces of the ground condition (blue traces) and height condition (purple traces) of the experimental group. Around the traces are the standard errors. The gray shaded area is the window over which the means were taken that were used for the permutation tests. On the x-axis is time (seconds), ranging from 0 to 30. On the y-axis is concentration change (μM), ranging from -0.4 to 0.6. Channel numbers are given in every plot. The plot with the border around it shows the channel where a significant difference was found between the means of the ground condition and the height condition for the experimental group.

4.2.2.2 fNIRS Deoxyhemoglobin

The second within-group statistical analysis was performed on the Δ[HbR] data of the experimental group. Figure 4.5 shows the grand average Δ[HbR] traces of the ground condition and the height condition for the experimental group, with the standard error given around every trace. Similar to the previous figures, the graphs are arranged according to the optode layout and the gray shaded areas indicate the 3-15 second window over which the mean values that were used in the permutation tests were calculated.

The grand average Δ[HbR] traces shown in the graphs seem rather flat for both the ground and the height condition, hence no activation and no significant differences between the conditions are expected based on visual inspection. However, the outcomes of the FDR-corrected permutation tests show that there is a significant difference in channel 23 ($p = 0.0017$) between the means of the grand average Δ[HbR] signal of the ground condition and that of the height condition over the 3-15 second window.

4.2.2.3 Heart Rate

The third within-group statistical analysis was performed on the average baseline-corrected HR over ground trials and the average baseline-corrected HR over height trials of the experimental group. See figure 4.6 for a box plot. It can be seen that the ranges of the box plots of the different conditions are somewhat similar, although the height condition box plot contains values that are more negative than those of the ground condition box plot. This indicates that, compared to the baseline, the average HR during the height condition dropped more severely for some participants than it did during the ground condition. However, both box plots contain positive as well as negative values, which means that for both conditions the average HRs dropped and increased with respect to the baseline. Therefore, no distinctive patterns can be observed between the baseline-corrected HRs of the two conditions, which suggests that the change in HR with respect to the baseline is rather subject-dependent. Based on the outcome of the permutation test, the average baseline-corrected HRs of the ground condition and the height condition within the experimental group are not significantly different ($p = 0.381$).

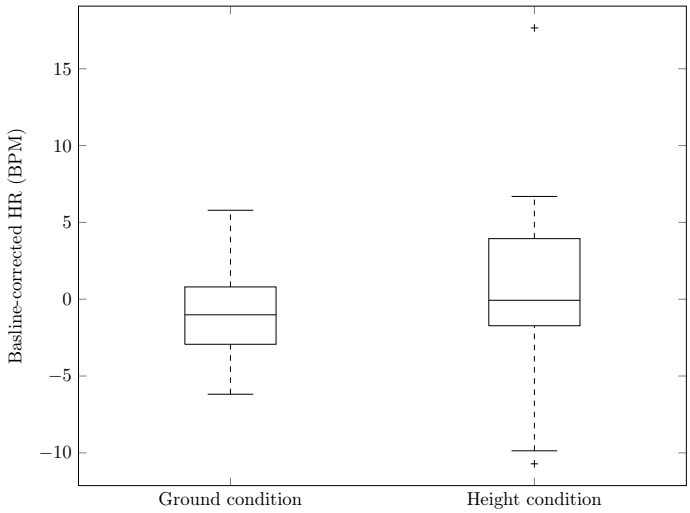


Figure 4.6: Box plot of the average baseline-corrected HR in BPM during the ground condition (left) and during the height condition (right) for the experimental group.

4.3 Classification

Multiple classifiers were trained and tested on the data of the experimental group, with the data from the ground trials labeled as "no fear response" and the data from the height trials labeled as "fear response". The within-group statistical analyses of the experimental group show that the Δ[HbO] signals of channels 1, 2, 3, 4, 6, 11, 12, 14, 18, 20, 23, 25, 26 and the Δ[HbR] signals of channel 23 differ significantly between the ground and height condition. Furthermore, the average HRs of the different conditions do not differ significantly. Therefore, only the data of the significant Δ[HbO] channels and the significant Δ[HbR] channel were used in the classifiers.

Two different types of classifiers were trained and tested: subject-dependent classifiers and subject-independent classifiers. Due to motion artifacts and hardware malfunctions, some channels were excluded from the analyses. Every subject-dependent classifier therefore used the data of the significant channels that were available for their respective participant. The subject-independent classifiers only used the data of the significant Δ[HbO] channels, as the significant Δ[HbR] channel was excluded for some participants, which made it impossible to train and test classifiers on this data. The classification results of both types of classifiers will be given in the following sub-sections.

4.3.1 Subject-Dependent Classifiers

The accuracies of the subject-dependent classifiers on the test dataset are given in Table 4.2. See Appendix H.1 for the confusion matrices of the classifiers. For every subject, six different subject-dependent classifiers were trained and tested: LDA and SVM, both on 1-second history, 3-second history, and 5-second history.

Table 4.2: Accuracies of the subject-dependent classifiers.

	1s history		3s history		5s history	
Participant	LDA	SVM	LDA	SVM	LDA	SVM
1	45.83	66.67	47.92	68.75	52.08	60.42
2	20.83	18.75	14.58	14.58	10.42	14.58
3	60.42	62.50	75.00	64.58	68.75	64.58
4	85.42	83.33	77.08	85.42	70.83	81.25
5	93.75	91.67	79.17	89.58	79.17	77.08
6	60.42	60.42	77.08	64.58	72.92	68.75
7	95.83	100.00	95.83	100.00	100.00	100.00
8	79.17	81.25	77.08	85.42	75.00	91.67
9	54.17	56.25	54.17	47.92	54.17	52.08
10	91.67	91.67	95.83	91.67	95.83	91.67
11	89.58	89.58	77.08	87.50	58.33	85.42
12	77.78	75.00	80.56	72.22	83.33	72.22
13	62.50	62.50	68.75	62.50	56.25	62.50
14	77.08	75.00	64.58	75.00	66.67	72.92
Mean	71.03	72.47	70.34	72.12	67.41	71.08
(± SD)	(± 21.47)	(± 20.61)	(± 20.74)	(± 21.80)	(± 21.89)	(± 21.16)

The performances of the six different subject-dependent classifiers can be determined based on their mean accuracies over participants. In this regard, the SVM on the 1-second history performs best, with a mean accuracy of 72.47% (SD 20.61) over participants. However, the mean accuracies of the other classifiers are close to that of the 1-second history SVM, with a maximum difference of roughly 5%. Therefore, the amount of history taken into the classifier seems to have minimal effect on this metric. Similarly, the choice between LDA or SVM classifier also seems to have a minimal effect, with a maximum of less than 4% difference between the accuracies of the different classifiers.

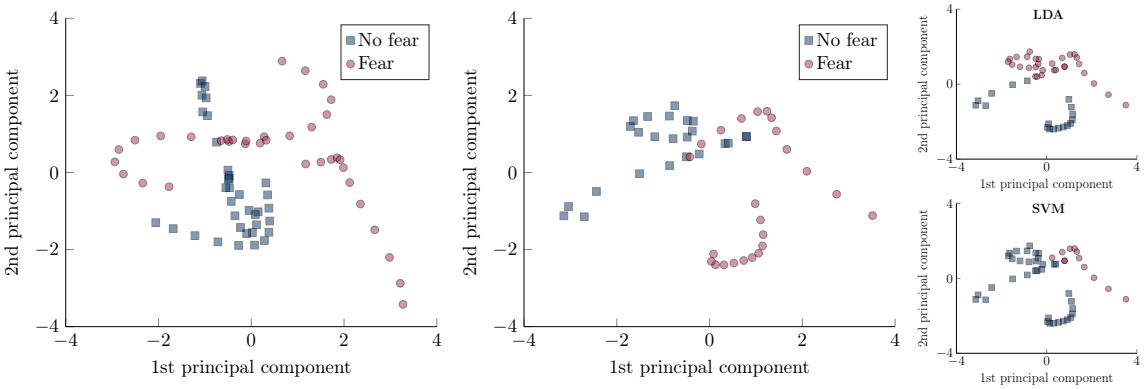
The performances of the different classifiers vary considerably within participants. For example, the accuracies of the different classifiers of participant 1 range from 45.83% (1-second history LDA)

to 68.75% (3-second history SVM), which is a difference of almost 30.00% accuracy between the worst and best performing classifier. Furthermore, the mean accuracies over participants do not always reflect the performances of the different classifiers on the data of a single subject. For example, the 5-second history LDA is the worst performing classifier in terms of mean accuracies over participants, while it is one of the best performing classifiers for participants 7 (accuracy of 100.00%), 10 (accuracy of 95.83%), and 12 (accuracy of 83.33%).

Besides the within-participant variation in accuracies of different classifiers, it is worth noting that the accuracies vary considerably among the different participants. This is also reflected by the rather high standard deviation of the accuracies of the different classifiers. The accuracies among participants range from as low as 10.42% (LDA classifier on 5-second history, participant 2) to even 100.00% (multiple classifiers of participant 7). In general, there are some participants for whom (almost) all of the classifiers perform around or even below the level of chance. Those participants are participant 1, participant 2, and participant 9. In the cases of these participants, the classifiers have no added value. However, there are also some participants for whom (almost) all of the classifiers perform very good, with accuracies above 90.00%. This is the case for participant 7 and participant 10. Therefore, the performances of these classifiers seems to be largely dependent on the subject.

4.3.1.1 Error Analysis

The differences in accuracies among participants are further analyzed in this section. To this end, the train and test data of participant 1, 2, and 9, for whom the classifier accuracies are around or below the level of chance, were inspected. For comparison, the train and test data of participant 7, for whom the subject-dependent classifiers perform the best, was also reviewed. Figures 4.7, 4.8, 4.9 and 4.10 contain scatter plots of the train data, test data, and the decisions of the subject-dependent classifiers on 1-second history of participants 1, 2, 7, and 9, respectively. PCA was used to find the first and second principal components of the data on which the classifiers are trained, which serve as the axes for the scatter plots. See Appendix I for the PCA procedure. The same plots can be made for the 3-second history and 5-second history data of those participants. However, the patterns of these plots are very similar to those of the 1-second history data. Therefore, the plots of the 3-second history and 5-second history data of these participants are given in Appendix I, while this section focuses on the 1-second history data only.

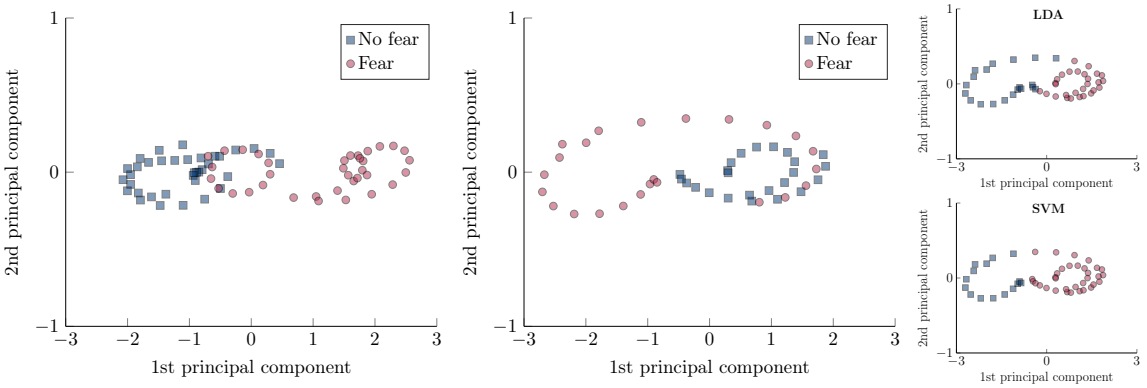


(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.7: Train and test data of the 1-second subject-dependent classifiers of participant 1.

Figure 4.7 contains scatter plots of the data labeled as "fear" (red circles) and "no fear" (blue squares) along the first and second principal component for participant 1. The data on which the classifiers are trained is plotted in Figure 4.7a, while the data on which the classifiers' performances are tested is plotted in Figure 4.7b. Figure 4.7c shows the decisions made by the LDA and SVM classifiers, plotted along the same axes as the train and test data. The test data depicted in Figure

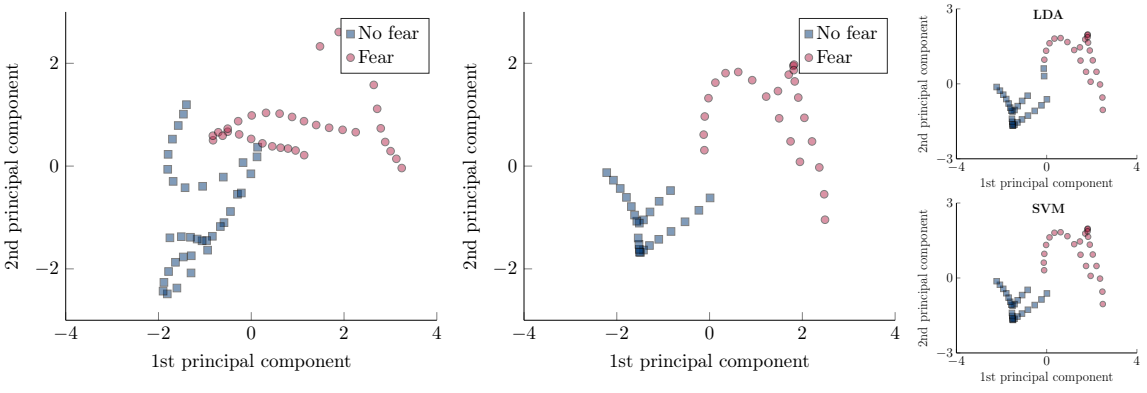
4.7b shows good linear separability along the first principal component. However, the train data shown in Figure 4.7a consists of a different pattern and is not linearly separable along the first principal component. Therefore, the linear classifiers trained on the data of Figure 4.7a are unlikely to generalize well to the test data of Figure 4.7b. From Figure 4.7c it becomes clear that the LDA and SVM try to separate the datapoints along another direction instead, which explains why the accuracies of the classifiers of participant 1 are mediocre.



(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.8: Train and test data of the 1-second subject-dependent classifiers of participant 2.

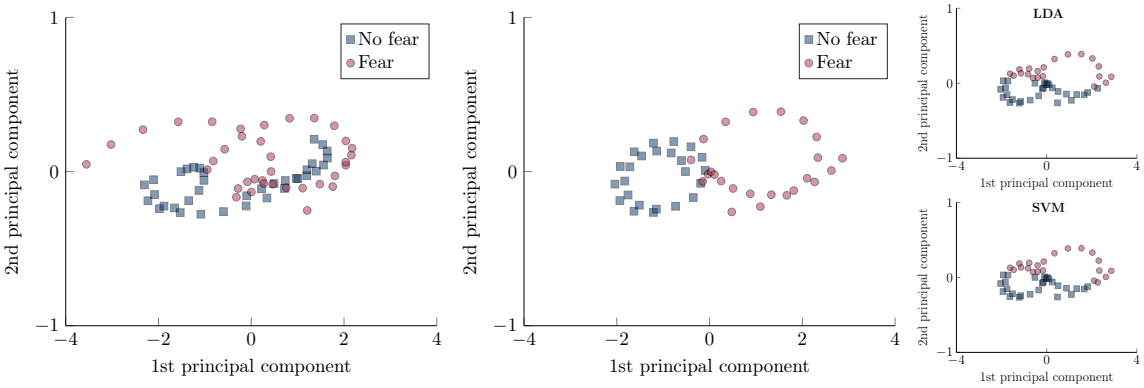
Similar scatter plots are given for participant 2 in Figure 4.8. From these figures, it can clearly be seen that the train data does not give a good representation of the test data. The train data of Figure 4.8a consists of mostly negative values along the first principal component for the data labeled as "no fear", and mostly positive values along the first principal component for the data labeled as "fear". However, in the test set the opposite is true, see Figure 4.8b. Therefore, it is impossible to train a classifier on the provided train data that generalizes well on the given test data, which explains why the classifiers perform at exceptionally low accuracies in the case of participant 2. From Figure 4.8c, it can be seen that the classification decisions are almost the opposite of the true test data.



(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.9: Train and test data of the 1-second subject-dependent classifiers of participant 7.

The scatter plots for participant 7 are given in Figure 4.9. The distribution of the test data shown in Figure 4.9b is rather similar to the distribution of the train data in Figure 4.9a. Additionally, the test data demonstrates very clear linear separability. This explains why the classifiers trained on the train data of participant 7 generalize very well to this participant's test data, as can be seen in Figure 4.9c. Therefore, high accuracies are achieved for participant 7.



(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.10: Train and test data of the 1-second subject-dependent classifiers of participant 9.

Lastly, Figure 4.10 contains the scatter plots of participant 9. Similar patterns can be observed as for participant 1. The test data plotted in Figure 4.10b shows good linear separability along the first principal component, with positive values for the data labeled as "fear" and negative values for the data labeled as "no fear". However, this pattern is absent in the train data in Figure 4.10a. In fact, the train data can better be separated along the second principal component, although perfect separation is impossible. The classifier decisions depicted in Figure 4.10c show that the test data is indeed separated along the second principal component during classification. Therefore, the classifiers of participant 9 yield accuracies around the level of chance.

4.3.2 Subject-Independent Classifiers

The accuracies of the subject-independent classifiers on the test dataset are given in Table 4.3. See Appendix H.2 for the confusion matrices of the classifiers. Again, six different classifiers were trained and tested for every participant: the LDA and SVM on 1-second history, 3-second history, and 5-second history.

Table 4.3: Accuracies of the subject-independent classifiers.

	1s history		3s history		5s history	
Participant	LDA	SVM	LDA	SVM	LDA	SVM
1	72.50	74.17	75.83	72.50	75.00	75.00
2	60.83	60.83	57.50	57.50	57.50	57.50
3	80.83	80.83	82.50	82.50	82.50	82.50
4	83.33	83.33	85.83	85.83	85.00	85.00
5	85.83	87.50	88.33	89.17	89.17	89.17
6	74.17	75.83	86.67	87.50	85.00	85.83
7	84.17	83.33	86.67	86.67	89.17	86.67
8	58.33	60.83	60.83	58.33	60.83	60.83
9	69.17	70.00	70.00	72.50	67.50	70.83
10	89.17	88.33	93.33	94.17	92.50	93.33
11	81.67	80.83	83.33	83.33	82.50	80.83
12	70.37	69.44	65.74	68.52	60.19	71.30
13	72.50	70.00	70.83	70.00	70.83	70.00
14	70.00	67.50	71.67	71.67	73.33	73.33
Mean	75.21	75.20	77.08	77.16	76.50	77.29
(± SD)	(± 9.29)	(± 9.12)	(± 11.15)	(± 11.48)	(± 11.74)	(± 10.64)

Based on the mean accuracies over participants, it can be concluded that the SVM on the 5-second history performs best on average, with a mean accuracy of 77.29% (SD 10.64) over participants. On the contrary, the SVM on the 1-second history performs the worst on average, with a mean accuracy of 75.20% (SD 9.12). Again, the difference between the accuracies of the classifiers that perform best and worst on average is only a few percent, which indicates that the amount of history and the type of classifier (LDA or SVM) have only a small influence on the classification performance.

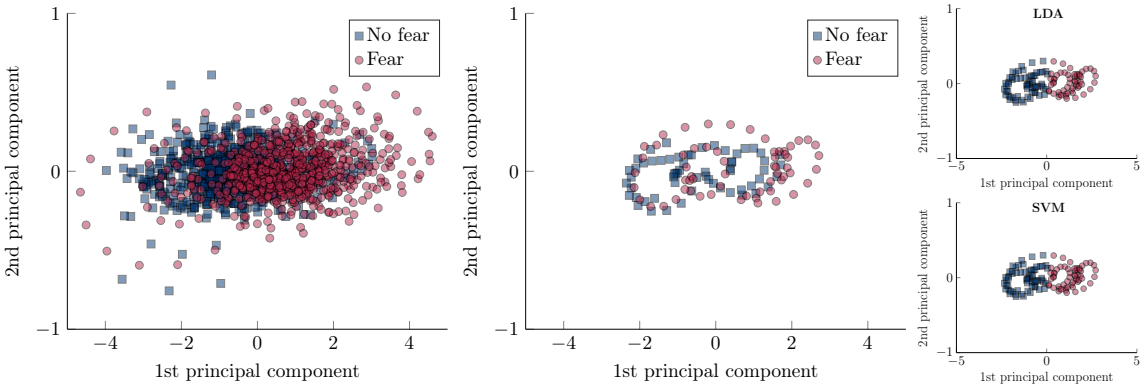
The mean accuracies within participants vary less between the different classifiers than in the case of the subject-dependent classifiers. However, still the mean accuracies of the classifiers do not always reflect the performance in the case of every individual participant. For example, the 1-second history SVM has the lowest mean accuracy over participants, while it is one of the best performing classifiers in the cases of participant 2 and participant 8 (both with an accuracy of 60.83%).

Furthermore, the accuracies vary considerably among participants again, although also this effect is less severe than in the case of the subject-dependent classifiers. The accuracies of the subject-independent classifiers range from 57.50% (multiple classifiers for participant 2) to 94.17% (3-second history SVM for participant 10). The overall performance is worst for participant 2, with accuracies ranging from 57.50% (3-second and 5-second history classifiers) to 60.83% (1-second history classifiers). However, this is still above the level of chance. The classifiers perform best in the case of participant 10, which accuracies ranging from 88.33% (1-second history SVM) to 94.17% (3-second history SVM).

Overall, the subject-independent classifiers seem to be more stable than the subject-dependent classifiers. This is mainly because of the smaller variation in accuracies within and between participants, and the higher mean accuracies overall.

4.3.2.1 Error Analysis

This section provides an analysis of the differences in the accuracies of the subject-independent classifiers of the participants for whom the performance was worst (participant 2) and best (participant 10). Figures 4.11 and 4.12 contain scatter plots of the 1-second history train data, test data, and classifier decisions for participant 2 and 10, respectively. Again, the datapoints are plotted along the first and second principal components of the train data. The plots of the 3-second history and 5-second history data show patterns that are very similar to those of the 1-second history data. Therefore, this section focuses on the analysis of the 1-second history data only, while the scatter plots of the 3-second history and 5-second history data of these participants can be found in Appendix I.

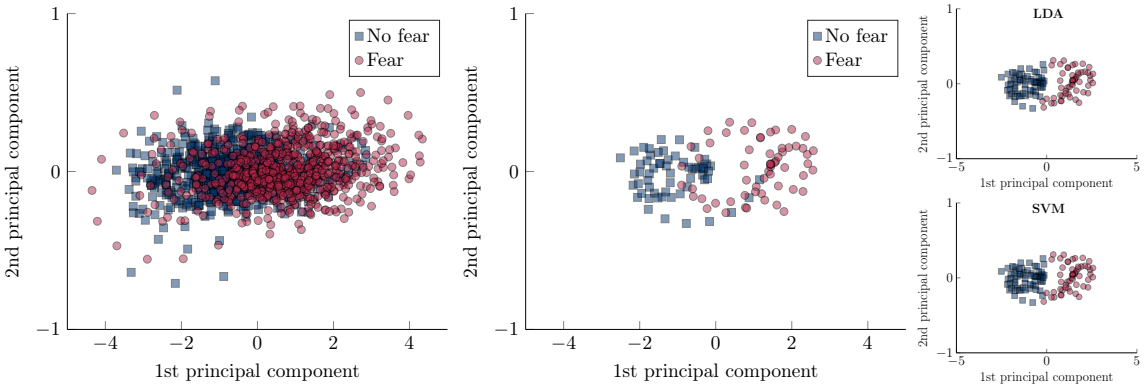


(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.11: Train and test data of the 1-second subject-independent classifiers of participant 2.

Figure 4.11 contains the scatter plots of the 1-second history train data, test data, and classifier decisions of the subject-independent classifier of participant 2. The train data depicted in Figure 4.11a shows that the data labeled as "no fear" is mostly centered around the negative values of the

first principal component, while the opposite is true for the data labeled as "fear". However, the test data of Figure 4.11b does not follow the same pattern. Instead, the data of the different labels are much more distributed over the positive and negative values on the first principal component, which makes it impossible to separate the test data of the different labels by a linear decision boundary. The differences in the distributions of the data shown in Figures 4.11a and 4.11b explain why the classifiers trained on the train data of Figure 4.11a perform mediocre on the test data of Figure 4.11b. Figure 4.11c shows that the decision boundary separates the test data along the first principal component, where it is actually not linearly separable.



(a) 1s history train data. **(b)** 1s history test data. **(c)** Classifier decision

Figure 4.12: Train and test data of the 1-second subject-independent classifiers of participant 10.

Similar scatter plots are given for participant 10 in Figure 4.12. Again, the train data depicted in Figure 4.12a shows that the data labeled as "no fear" is centered around the negative values of the first principal component, while the data labeled as "fear" is centered around the positive values of the first principal component. Figure 4.12b shows that the test data of participant 10 is distributed in a similar manner, albeit more densely. A linear classifier trained on the data of Figure 4.12a can generalize well to the test data of Figure 4.12b, as can be seen in Figure 4.12c.

Chapter 5

Discussion

This chapter provides the discussion of the results that were generated by this research. The results of the statistical analyses and the offline classification are discussed and compared with the literature. Furthermore, the contributions and limitations of this research, along with recommendations for future work on this topic are provided.

5.1 Statistical Analyses

5.1.1 Between-Group fNIRS Analysis

The results of the between-group statistical analysis of the fNIRS signals show that the grand average contrast Δ[HbO] signals of the control group and the experimental group are significantly different in channel 3. No significant differences were found between the grand average contrast Δ[HbR] signals of the two groups. The fact that only one out of 27 channels shows a significant difference between the two groups for only one chromophore suggests that the fNIRS signals of people with fear of heights responses and people without fear of heights responses are not very different from each other.

These results cannot be compared to the literature, as no previous work was found that compares fNIRS signals of an experimental group (people with a certain fear) and a control group (people without this fear) to each other. Instead, only literature was found that compares the Δ[HbO] and sometimes the Δ[HbR] signals of the same group of participants during fearful conditions and control conditions to each other [74–81]. The grand average Δ[HbO] plots of the contrast given in Figure 4.1 show that, in most cases, the Δ[HbO] signal of the experimental group peaks higher than that of the control group. This is somewhat in line with the findings from literature, where it is reported that increased Δ[HbO] is measured in (some areas of) the PFC during fearful conditions [74, 76–81]. Although higher Δ[HbO] values are measured for the experimental group than for the control group, their Δ[HbR] values seem more or less equal, with a few exceptions. See Figure 4.2. This is contrary to what was found in some other works that reported decreased Δ[HbR] values during fearful conditions [75, 76, 78], although the majority do not report this decrease in Δ[HbR] [74, 77, 79, 80].

It is important to note that previous works also measured increased Δ[HbO] values over the PFC while participants were experiencing other mental states, such as mental workload [53–56, 58, 59, 61–64], mental stress [67, 69–73], affective responses [82–84], attention [87, 88, 90, 91], deception [93–97], preference [52, 98–100], anticipation [101–103], suspicion [104, 105], and frustration [105–107]. This indicates that increased Δ[HbO] values are not only an indication of fear responses, but also of many more mental states. Therefore, it is possible that in this research the Δ[HbO] measurements of the control group might have been influenced by other mental states, which could have affected the statistical differences between the two groups. This effect is less likely for the experimental group, as they indicated that they were feeling afraid during the height exposure, which makes it improbable that they also experienced other mental states.

5.1.2 Between-Group HR Analysis

The contrast HR values of the experimental group and the control group are not significantly different. However, a rather low p-value was obtained ($p = 0.051$), which suggests that the difference in contrast HR values of both groups is close to significance. One other work was found that analyzed the HR between groups (phobics versus control group), which also did not find a significant difference [125]. Other works that assessed the HR values of an experimental group and a control group during virtual exposure reported that the HR values of both groups increased from virtual ground conditions to virtual height conditions [123, 124]. The box plot given in Figure 4.3 shows that this was also the case for some participants of the control group of this research, which could have affected the statistical significance of the differences in HR between groups. Based on these observations, it can be said that the results of the between-group HR analysis are similar to those found in the literature.

5.1.3 Within-Group fNIRS Analysis

The results of the within-group statistical analysis of the fNIRS signals show that the grand average Δ[HbO] values are significantly higher during the height (i.e. "fear") condition than during the ground (i.e. "no fear") condition. This significant difference was observed in a total of 13 channels, which are all located towards the frontal part of the PFC. These results indicate that during fear responses, the Δ[HbO] values increase significantly as compared to no fear responses, which is in accordance with the vast majority of literature on fNIRS measurements taken during fear responses [74, 76–81].

Additionally, the results of the within-group analysis show that the grand average Δ[HbR] values of the height (i.e. "fear") condition and the ground (i.e. "no fear") condition are significantly different in channel 23. Surprisingly, the grand average Δ[HbR] signal of the height condition is higher than that of the ground condition in this channel. This contradicts with some findings from the literature, where decreased Δ[HbR] values are reported for fearful conditions [75, 76, 78]. It remains unclear why this result differs from the literature.

The significant differences that were found in the Δ[HbO] and Δ[HbR] signals of the height (i.e. "fear") condition and the ground (i.e. "no fear") condition for the experimental group suggest that fear responses can be detected based on fNIRS data. Therefore, the results of this within-group statistical analysis provide the foundation for the use of classifiers to detect fear responses based on fNIRS data.

5.1.4 Within-Group HR Analysis

The within-group statistical analysis of the HR values of the experimental group did not show significant differences between the baseline-corrected HR values of the ground condition and the height condition. This contradicts with other findings described in the literature, which found significant increases in HR values between virtual ground and height conditions [122–124]. A possible explanation for the discrepancies between the result of this thesis research and the results found in the literature might be the experimental procedure. The experimental procedures of [122–124] presented their participants once to each condition for longer periods of time (up to 10 minutes), whereas the experiment of this thesis research presented participants multiple times to every condition for much shorter periods of time (30 seconds). The frequent change between conditions and the relatively short time intervals could have had an influence on the HR of the participants.

5.2 Classification

5.2.1 Subject-Dependent Classifiers

The subject-dependent classifier that achieved the highest average accuracy over participants is the 1-second history SVM, with an average accuracy of 72.47% (SD 20.61). This average accuracy

is only a few percent higher than the average accuracies of the other subject-dependent classifiers, which operated on 3-second history, 5-second history, and/or used the LDA algorithm. These results suggest that the amount of history and the choice between the LDA or SVM algorithm have minimal influence on the subject-dependent classification.

It is striking that the subject-dependent classifiers do not perform well for participants 1, 2, and 9, while they are all trained on the data of their respective subject. The error analysis showed that in the case of participant 2, the test data has a very different pattern than the train data, which makes it impossible to train a model that performs well on the classification of the test data of this participant. Furthermore, in the cases of participant 1 and 9, the test data is linearly separable along one dimension, however the classifiers learned a decision boundary that separates the data along another dimension. It can be said that the train data of participants 1, 2 and 9 does not give an accurate representation of their test data, which caused the subject-dependent classifiers to perform at accuracies around or below the level of chance. It remains unclear what caused the observed differences between the train and test data of these participants. One possible explanation could be that the fear responses and accompanying fNIRS measurements of these participants were not stable over time. Recall that the subject-dependent classifiers are trained on the data of the first 6 trials, while their accuracies are tested on the data of the final 4 trials. It could be the case that the fear responses of these subjects changed over time, which influenced their fNIRS measurements, while the labels in the train and test datasets remained the same.

5.2.2 Subject-Independent Classifiers

The subject-independent classifiers perform better than the subject-dependent classifiers, with an average accuracy of 77.29% (SD 10.64) over participants for the 5-second history SVM. Again, the other classifiers that operated on 1-second history, 3-second history, and/or used the LDA algorithm yielded very similar average accuracies. Therefore, also in the case of the subject-independent classifier, the choice between LDA and SVM and the choice for the amount of history to take into account do not have much influence on the performance of the classifier.

There is one participant for whom the subject-independent classifiers perform only slightly better than chance: participant 2. The test data of this participant shows that, in fact, no clear separation between the "fear" and "no fear" fNIRS measurements exists, see Figure 4.11b. This is surprising, as this participant had high AQ, SUDS, and IPQ scores (pre-experiment AQ = 59, post-experiment AQ = 64, SUDS height trials = 80, SUDS ground trials = 0, average IPQ = 4.4). Based on these scores, it can be assumed that this participant has a strong fear of heights, felt very anxious during the height trials, felt relaxed during the ground trials, and felt present in the VEs. Therefore, it would be expected that the "fear" and "no fear" fNIRS data of this participant would be more distinctive from each other than they actually are. The error analysis of the train and test data of the subject-dependent classifier of participant 2 revealed that the fNIRS measurements of this participant are not stable over time. The train data, taken from the first 6 trials, shows almost the opposite pattern as the test data, taken from the final 4 trials. See Figure 4.8. This explains why the test data of participant 2 for the subject-independent classifier is distributed over the data space without a clearly recognizable pattern. However, it remains unknown what caused the instability observed in the data over time.

It is remarkable that for most participants, the subject-independent classifiers outperform the subject-dependent classifiers. Instead, it was expected that the subject-dependent classifiers would perform better, as they are trained on the data that contains the characteristics that are specific to the subject, whereas the subject-independent classifiers do not. It was already discovered that the subject-dependent classifiers of participants 1, 2, and 9 do not perform well due to the fact that the train data of the first 6 trials deviates from the test data of the last 4 trials. However, from the error analysis of the subject-dependent classifiers it was observed that the data from the final 4 trials of participants 1 and 9 was very close to perfect linear separability, with positive values along the first principal component for the data labeled as "fear" and negative values along the first principal component for the data labeled as "no fear". Similar patterns can be observed for the train data of the subject-independent classifiers, shown in Figures 4.11a and 4.12a. It could be the case that the

data of the participants for whom the subject-dependent classifiers performed worse, converges over time to patterns similar to those of the other participants. This would explain the relatively good performance of the subject-independent classifiers.

The fact that the subject-independent classifiers perform at least as good as the subject-dependent classifiers (based on accuracy), indicates the potential of these classifiers for online detection scenarios. In such a scenario, the subject-independent classifier can be trained on the data of a given number of people, based on which it can classify the fear responses of an unknown person. The results of this research suggest that training the subject-independent classifier on the fNIRS data of 13 people is already enough to classify the fear responses of an unknown person at an average accuracy above 75%.

5.2.3 Overall Classification Performance

The overall average classification performances of the subject-dependent classifiers are around an accuracy of 70%, whereas the subject-independent classifiers have average accuracies around 75% to 77%. These accuracies are significantly higher than the level of chance, which is 50%. However, an accuracy around 70% to 77% is not sufficient for real-life applications, where this technology could for example be used during VRET settings. At an accuracy of 70%, the classifier makes the wrong decision 30% of the time. Imagine that 30% of an exposure therapy session consists of the wrong exposure scenario based on these decisions. This would not only decrease the effectiveness of the exposure therapy, it could also cause very inconvenient situations, such as patients having panic attacks because they are exposed to the wrong scenario.

The average accuracies obtained by the classifiers of this research are mediocre compared to those of other classifiers that use fNIRS data to detect other mental states, see Table 2.1. Other works reported that they were able to detect mental workload [56], attentional state [88, 90–92], deception [93], and frustration [106] at average accuracies ranging from 74.8% to 90.7%. However, lower average accuracies ranging from 63% to 72.9% were also reported in the literature for the detection of mental stress [71], affective state [84], and preference [98]. It must be noted that it is difficult to compare the performance of the classifiers of this research to those found in the literature, as none of the other works focused on the detection of a fear response.

The average accuracies at which the classifiers of this research perform are also mediocre to inferior compared to those of other studies that focused on the detection of fear elicited by VR exposure, using different physiological signals. Accuracies ranging from 76% to 89.5% were reported for the classification of fear versus no fear based on the BVP data of 7 participants [128] and the combination of GSR, HR, and EEG data of 8 participants [15]. However, the smaller number of participants, and thus smaller amount of data used in those works, makes their results less reliable than the results generated by this thesis research.

5.3 Contributions

There are several contributions made by this thesis research. First of all, this research is the first to expose both a control group and an experimental group to virtual heights while their fNIRS measurements were taken and to show that their cortical hemodynamic responses are different to some extent. Furthermore, this research has shown the feasibility of the combination of immersive VR (presented through an HMD) and fNIRS measurements, to elicit and detect fear responses. Finally, this research demonstrated that fNIRS measurements have the potential to be used to detect fear responses to some extent, using linear classifiers.

5.4 Limitations

The research described in this thesis comes with several limitations. These limitations will be discussed in this section.

First of all, the amount of participants is a limitation. Eventually, only 15 participants were part of the experimental group and 14 participants were part of the control group. The limited amount of participants in both groups makes the statistical significance of the outcomes of this research debatable, as larger groups of participants could potentially reveal different results.

Secondly, the use of the AQ to select participants with a fear of heights and participants without a fear of heights was less reliable than expected. Although several other works about fear of heights responses used the AQ to assess the severity of their participants' fear of heights [27–30, 75], it was not always a reliable indication in this research. Based on the post-experiment AQ scores, discrepancies between people's fear of heights according to the AQ scores were discovered, see Appendix J. This made the distribution of participants over the control and experimental group complex and potentially even inaccurate, which could have had an impact on the statistical results and the classification performances.

Moreover, the statistical analyses of the fNIRS data and the classification performances are based on the mean Δ[HbO] and mean Δ[HbR] values only. However, from the literature on mental state detection it is know that signal features like amplitude, slope, standard deviation, kurtosis, skewness, and signal peaks were used in other classifiers for mental state detection [56, 71, 90, 92], see Table 2.1. This thesis research did not investigate to what extent the statistical significance and classification performances could be improved when using different features of the fNIRS signals.

Furthermore, the amount of data that was captured from every participant was limited. Participants were only asked to undergo 10 trials of virtual exposure during which their fNIRS measurements were taken. The limited amount of data makes it more difficult to perform reliable statistical analyses and to train and test robust classifiers. The amount of data also limits the possibilities of the classifier. Now, only binary classifiers, that classify between "fear response" and "no fear response", could be trained and tested. Unfortunately, the limited amount of data does not allow the training of multi-class classifiers that could detect the level of the fear response.

Besides the small amount of trials that were recorded, the trial duration was also limited to only 30 seconds. The short trial duration impacted the analysis of cardiac signals derived from the fNIRS data. Although the HR could successfully be extracted from the fNIRS signals, the HRV data of the participants could not be assessed, as conventional time- and frequency-domain measurements of HRV require at least 2 to 5 minute data epochs [162]. HRV is suggested to be a useful feature to detect fear [121, 122]. However, due to the short trial duration it remains unknown to what extent the HRV could have improved the statistical results or classification performances of this research.

Additionally, the hardware components were not optimal for simultaneous usage. The fNIRS cap (i.e. the Artinis Brite 24) and the VR HMD (i.e. the Oculus Rift S) caused an uncomfortable feeling for most participants when they were worn simultaneously. The Oculus Rift S had to be tightened with a headband around the participant's head. This put some extra pressure on the optodes of the fNIRS cap, which felt unpleasant for the participant. This might have influenced the experience of the participants. Since it is difficult to quantify the effect of the uncomfortable feeling caused by the hardware components, it is unknown to what extent it affected the participants and the results generated by this research.

Finally, the fNIRS technology proved to be less resistant to motion artifacts than was expected from the literature [19, 42, 43]. Therefore, participants were instructed to look around very slowly in the VEs and to limit their bodily movements. This is contrary to the research of Landowska et al. [75], who instructed their participants to walk over a wooden plank during the exposure. The limited movement that was allowed during the experiment could have made the experience of the VEs less realistic for the participants, which could have influenced the data that was collected. Furthermore, even looking around slowly caused motion artifacts in the fNIRS signals, due to which some channels could not be used in the data analyses.

5.5 Recommendations for Future Work

Based on the results and the limitations of this research, several recommendations for future work on the topic of fear detection based on fNIRS data can be made. These recommendations will be

discussed below.

A first recommendation would be to verify the results of this thesis research by conducting more experiments with more participants, both belonging to the control group and the experimental group. Data from more participants will make the statistical analyses, and thus the results, more reliable than they are now.

A second recommendation would be to recruit more participants who suffer from varying degrees of fear of heights. Data from, for example, people with little, moderate, and severe fear of heights can be used to train multi-class classifiers. This way, it can be investigated whether a classifier can be used to distinguish between different degrees of fear using fNIRS data.

Apart from collecting more fNIRS data by recruiting more participants, future research could also focus on the collection of more fNIRS data from every individual participant. Longer trial durations allow to extract the HRV from the fNIRS data, which could be used to investigate whether the fNIRS derived HRV data between or within groups are significantly different and if it could benefit the detection of fear responses. Additionally, data from longer trial durations could be used to study the effects of fearful stimuli on fNIRS data within a given time period.

Furthermore, it is advised to find better metrics to test before the experiment if someone has a fear of heights, in order to establish a clearer distinction between the control group and the experimental group. Besides the AQ, other questionnaires like the VHI can be used to assess a person's fear of heights, similar to [15, 127]. Another suggestion would be to let people experience various situations involving heights before the actual experiment, and to let them fill out the pre-experiment AQ after that. This might make it easier for people to self-report on their actual fear of heights.

Moreover, it should be investigated if different signal features could improve the statistical significance of channels and the performances of the classifiers. The grand average Δ[HbO] traces presented in Chapter 4 revealed that the traces of the experimental group generally rise to a peak value, whereas this pattern is less apparent for the grand average traces of the control group, see Figure 4.1. A similar observation was made for the grand average Δ[HbO] traces of the height condition and ground condition of the experimental group, see Figure 4.4. Based on these observations, it is expected that the additional usage of features such as the maximum signal value, the time to peak, and the signal slope have the potential to improve the results.

Finally, when this research would be executed on a larger scale or when the technology would be implemented in real VRET settings, it would be advisable to develop a single hardware component that contains both the fNIRS cap and the VR HMD. This hardware device should be developed in such a way that the optodes and the VR HMD do not interfere with each other anymore, such that they do not cause unpleasant pressing feelings on the user's head. This can be achieved by integrating the fNIRS optodes into the headbands of the HMD.

Chapter 6

Conclusion

This chapter provides the conclusions that can be drawn based on the research described in this thesis. To this end, the research questions that were posed in the introduction will be answered.

This thesis started with the motivation to detect fear responses during VR exposure, using non-invasive fNIRS measurements. Based on the findings from the literature research, an experiment was designed similar to that of Landowska et al. [75]. The goal of the experiment was to create non-fearful situations (i.e. the ground condition) and fear eliciting situations (i.e. the height condition) in the VEs, while collecting fNIRS data from the participants. A total of 41 participants were invited to participate in the experiment. Both people with fear of heights and people without fear of heights were recruited, based on an assessment of the severity of their fear of heights using the AQ. The fNIRS data that was captured from the participants consisted of Δ[HbO] signals, from which HR was extracted, and Δ[HbR] signals. Post-experimental AQ scores, SUDS scores, and IPQ scores were used to decide which participants were experiencing fear responses during the experiment (experimental group) and which participants were not (control group). A total of 15 participants were selected to be part of the control group for the analysis, while 14 participants were selected to be part of the experimental group for the analysis.

> **1** *To what extent do the fNIRS signals captured from people with a fear of heights response and people without a fear of heights response differ?*

The grand averages of the Δ[HbO] signals of the contrast between the ground (i.e. "no fear") condition and the height (i.e. "fear") condition show that in most channels the signals of the experimental group exceed those of the control group between the 3-15 second post-stimulus window. However, this effect is only significant at channel 3. Therefore, it can only be concluded that at channel 3 there exists a significant difference between the grand average Δ[HbO] contrast signals of people with a fear of heights response and people without a fear of heights response, where the grand average Δ[HbO] contrast signal of the people with the fear of heights response shows a clear peak in the 3-15s window, whereas this peak is non-existent for the people without a fear of heights response.

On the contrary, no significant differences were found between the Δ[HbR] contrast signals and the contrast HR (extracted from the Δ[HbO] signals) of the people with fear of heights responses and the people without fear of heights responses. Therefore, it is concluded that the grand average Δ[HbR] contrast signals and the contrast HRs of people with fear of heights and people without fear of heights do not differ significantly.

> **2** *To what extent can a person's fear of heights response to a virtual reality environment be detected using fNIRS data?*

The within-group statistical analysis of the experimental group showed that there are significant differences in the grand average Δ[HbO] values during fear responses and during no-fear responses, where the Δ[HbO] values of the fear responses were significantly higher than those of the no-fear responses in channels 1, 2, 3, 4, 6, 11, 12, 14, 18, 20, 23, 25, and 26. Another significant channel was found for the grand average Δ[HbR] signals, which is channel 23.

Subject-dependent as well as subject-independent classifiers can be used to detect a person's fear of heights response to a VR environment based on the fNIRS data of the significant channels of the experimental group. The subject-dependent SVM classifier on 1-second history of the fNIRS signals can detect a fear of heights response with an average accuracy of 72.47% (SD 20.61) over participants. However, the accuracies of this classifier range from 10.42% (far below the level of chance) to 100.00%. The subject-independent SVM classifier on 5-second history of the fNIRS signals can detect a fear of heights response with an average accuracy of 77.29% (SD 10.64). The accuracies of this classifier range from 57.50% to 94.17%. Based on the variation in accuracies, it is concluded that it is very dependent on the person to what extent his/her fear of heights response to a VE can be detected using fNIRS data. The subject-independent classifiers show potential for usage in online detection situations, as they can be trained beforehand on existing fNIRS data and can classify the unseen data of a new person at average accuracies above 75.00%.

Bibliography

[1] G. Riva, "Virtual reality in psychotherapy: Review," *CyberPsychology & Behavior*, vol. 8, no. 3, pp. 220–231, Jul. 2005. DOI: 10.1089/cpb.2005.8.220.

[2] G. Riva and B. Wiederhold, "The new dawn of virtual reality in health care: Medical simulation and experiential interface," *Annual Review of CyberTherapy and Telemedicine*, vol. 13, pp. 3–6, Dec. 2015. DOI: 10.3233/978-1-61499-595-1-3.

[3] G. Tieri, G. Morone, S. Paolucci, and M. Iosa, "Virtual reality in cognitive and motor rehabilitation: Facts, fiction and fallacies," *Expert Review of Medical Devices*, vol. 15, no. 2, pp. 107–117, 2018. DOI: 10.1080/17434440.2018.1425613.

[4] M. Tarr and W. Warren, "Virtual reality in behavioral neuroscience and beyond," *Nature neuroscience*, vol. 5, pp. 1089–1092, Dec. 2002. DOI: 10.1038/nn948.

[5] W. P. Teo, M. Muthalib, S. Yamin, A. M. Hendy, K. Bramstedt, E. Kotsopoulos, S. Perrey, and H. Ayaz, "Does a combination of virtual reality, neuromodulation and neuroimaging provide a comprehensive platform for neurorehabilitation? – a narrative review of the literature," *Frontiers in Human Neuroscience*, vol. 10, 2016. DOI: 10.3389/fnhum.2016.00284.

[6] V. Eswaran, M. Veezhinathan, G. Balasubramanian, and A. Taneja, "Virtual reality therapy for mental stress reduction," *Journal of Clinical and Diagnostic Research*, vol. 12, no. 10, JC11–JC16, Oct. 2018. DOI: 10.7860/JCDR/2018/36055.12109.

[7] S. B. Moro, S. Bisconti, M. Muthalib, M. Spezialetti, S. Cutini, M. Ferrari, G. Placidi, and V. Quaresima, "A semi-immersive virtual reality incremental swing balance task activates prefrontal cortex: A functional near-infrared spectroscopy study," *NeuroImage*, vol. 85, no. 1, pp. 451–460, 2014. DOI: https://doi.org/10.1016/j.neuroimage.2013.05.031.

[8] M. A. Martens, A. Antley, D. Freeman, M. Slater, P. J. Harrison, and E. M. Tunbridge, "It feels real: Physiological responses to a stressful virtual reality environment and its impact on working memory," *Journal of Psychopharmacology*, vol. 33, no. 10, pp. 1264–1273, 2019. DOI: 10.1177/0269881119860156.

[9] D. Gromala, X. Tong, A. Choo, M. Karamnejad, and C. D. Shaw, "The virtual meditative walk: Virtual reality therapy for chronic pain management," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2015, pp. 521–524, ISBN: 9781450331456. DOI: 10.1145/2702123.2702344.

[10] A. Li, Z. Montaño, V. J. Chen, and J. I. Gold, "Virtual reality and pain management: Current trends and future directions," *Pain Management*, vol. 1, no. 2, pp. 147–157, 2011. DOI: 10.2217/pmt.10.15.

[11] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: The link between presence and emotions," *CyberPsychology & Behavior*, vol. 10, no. 1, pp. 45–56, 2007. DOI: 10.1089/cpb.2006.9993.

[12] W.-P. Brinkman, G. Sandino, and C. Mast, "Field observations of therapists conducting virtual reality exposure treatment for the fear of flying," Jan. 2009.

[13] S. G. Hofmann, "Cognitive processes during fear acquisition and extinction in animals and humans: Implications for exposure therapy of anxiety disorders," *Clinical Psychology Review*, vol. 28, no. 2, pp. 199–210, 2008. DOI: 10.1016/j.cpr.2007.04.009.

[14] D. Boeldt, E. McMahon, M. McFaul, and W. Greenleaf, "Using virtual reality exposure therapy to enhance treatment of anxiety disorders: Identifying areas of clinical adoption and potential obstacles," *Frontiers in Psychiatry*, vol. 10, no. 773, 2019. DOI: 10.3389/fpsyt.2019.00773.

[15] O. Balan, G. Moise, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy," *Sensors*, vol. 20, no. 2, p. 496, Jan. 2020. DOI: 10.3390/s20020496.

[16] A. Rizzo, B. John, B. Newman, J. Williams, A. Hartholt, C. Lethin, and J. G. Buckwalter, "Virtual reality as a tool for delivering ptsd exposure therapy and stress resilience training," *Military Behavioral Health*, vol. 1, no. 1, pp. 52–58, 2013. DOI: 10.1080/21635781.2012.721064.

[17] A. P. Hill and C. J. Bohil, "Applications of optical neuroimaging in usability research," *Ergonomics in Design*, vol. 24, no. 2, pp. 4–9, 2016. DOI: 10.1177/1064804616629309.

[18] A. Rodríguez, B. Rey, M. Clemente, M. Wrzesien, and M. Alcañiz, "Assessing brain activations associated with emotional regulation during virtual reality mood induction procedures," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1699–1709, 2015. DOI: https://doi.org/10.1016/j.eswa.2014.10.006.

[19] N. Noman and H. Keum-Shik, "fNIRS-based brain-computer interfaces: a review.," *Frontiers in Human Neuroscience*, vol. 9, p. 3, 2015. DOI: 10.3389/fnhum.2015.00003.

[20] Artinis Medical Systems. (2020). Brite - a wearable multi-channel fnirs device for brain oxygenation measurement., [Online]. Available: https://www.artinis.com/brite. (accessed: 30.03.2020).

[21] A. Garcia-Palacios, H. Hoffman, A. Carlin, T. Furness, and C. Botella, "Virtual reality in the treatment of spider phobia: A controlled study," *Behaviour Research and Therapy*, vol. 40, no. 9, pp. 983–993, 2002. DOI: 10.1016/S0005-7967(01)00068-7.

[22] A. Miloff, P. Lindner, P. Dafgård, S. Deak, M. Garke, W. Hamilton, J. Heinsoo, G. Kristofersson, J. Rafi, K. Sindemark, J. Sjölund, M. Zenger, L. Reuterskiöld, G. Andersson, and P. Carlbring, "Automated virtual reality exposure therapy for spider phobia vs. in-vivo one-session treatment: A randomized non-inferiority trial," *Behaviour Research and Therapy*, vol. 118, pp. 130–140, 2019. DOI: 10.1016/j.brat.2019.04.004.

[23] P. Lindner, A. Miloff, C. Bergman, G. Andersson, W. Hamilton, and P. Carlbring, "Gamified, automated virtual reality exposure therapy for fear of spiders: A single-subject trial under simulated real-world conditions," *Frontiers in Psychiatry*, vol. 11, p. 116, 2020. DOI: 10.3389/fpsyt.2020.00116.

[24] B. O. Rothbaum, L. Hodges, S. Smith, J. H. Lee, and L. Price, "A controlled study of virtual reality exposure therapy for the fear of flying.," *Journal of Consulting and Clinical Psychology*, vol. 68, no. 6, pp. 1020–1026, 2000. DOI: 10.1037/0022-006X.68.6.1020.

[25] B. O. Rothbaum, P. Anderson, E. Zimand, L. Hodges, D. Lang, and J. Wilson, "Virtual reality exposure therapy and standard (in vivo) exposure therapy in the treatment of fear of flying," *Behavior Therapy*, vol. 37, no. 1, pp. 80–90, 2006. DOI: 10.1016/j.beth.2005.04.004.

[26] N. Maltby, I. Kirsch, M. Mayers, and G. J. Allen, "Virtual reality exposure therapy for the treatment of fear of flying: A controlled investigation," *Journal of Consulting and Clinical Psychology*, vol. 70, no. 5, pp. 1112–1118, 2002. DOI: 10.1037//0022-006x.70.5.1112.

[27] P. M. G. Emmelkamp, M. Bruynzeel, L. Drost, and C. A. P. G. van der Mast, "Virtual reality treatment in acrophobia: A comparison with exposure in vivo," *CyberPsychology & Behavior*, vol. 4, no. 3, pp. 335–339, 2001. DOI: 10.1089/109493101300210222.

[28] T. Donker, S. V. Esveld, N. Fischer, and A. V. Straten, "0phobia – towards a virtual cure for acrophobia: Study protocol for a randomized controlled trial," *Trials*, vol. 19, no. 433, 2018. DOI: https://doi.org/10.1186/s13063-018-2704-6.

[29] D. Freeman, P. Haselton, J. Freeman, B. Spanlang, S. Kishore, E. Albery, M. Denne, P. Brown, M. Slater, and A. Nickless, "Automated psychological therapy using immersive virtual reality for treatment of fear of heights: A single-blind, parallel-group, randomised controlled trial," *The Lancet Psychiatry*, vol. 5, no. 8, pp. 625–632, 2018. DOI: 10.1016/S2215-0366(18)30226-8.

[30] D. Gromer, O. Madeira, P. Gast, M. Nehfischer, M. Jost, M. Müller, A. Mühlberger, and P. Pauli, "Height simulation in a virtual reality cave system: Validity of fear responses and effects of an immersion manipulation," *Frontiers in Human Neuroscience*, vol. 12, 2018. DOI: 10.3389/fnhum.2018.00372.

[31] J. Wald and S. Taylor, "Efficacy of virtual reality exposure therapy to treat driving phobia: A case report," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 31, no. 3, pp. 249–257, 2000. DOI: 10.1016/S0005-7916(01)00009-X.

[32] B. O. Rothbaum, L. F. Hodges, D. Ready, K. Graap, and R. D. Alarcon, "Virtual reality exposure therapy for vietnam veterans with posttraumatic stress disorder," *The Journal of Clinical Psychiatry*, vol. 62, no. 8, pp. 617–622, 2001. DOI: 10.4088/JCP.v62n0808.

[33] J. Difede and H. G. Hoffman, "Virtual reality exposure therapy for world trade center posttraumatic stress disorder: A case report," *CyberPsychology & Behaviour*, vol. 5, no. 6, pp. 529–535, 2002. DOI: 10.1089/109493102321018169.

[34] M. Gerardi, B. O. Rothbaum, K. Ressler, M. Heekin, and A. Rizzo, "Virtual reality exposure therapy using a virtual iraq: Case report," *Journal of Traumatic Stress*, vol. 21, no. 2, pp. 209–213, 2008. DOI: 10.1002/jts.20331.

[35] B. O. Rothbaum, M. Price, T. Jovanovic, S. D. Norrholm, M. Gerardi, B. Dunlop, M. Davis, B. Bradley, E. J. Duncan, A. Rizzo, and K. J. Ressler, "A randomized, double-blind evaluation of d-cycloserine or alprazolam combined with virtual reality exposure therapy for posttraumatic stress disorder in iraq and afghanistan war veterans," *American Journal of Psychiatry*, vol. 171, no. 6, pp. 640–648, 2014. DOI: 10.1176/appi.ajp.2014.13121625.

[36] Airbus. (2019). Airbus brings cockpit to you with new virtual reality flight trainer, [Online]. Available: https://services.airbus.com/en/newsroom/stories/2019/12/airbus-brings-cockpit-to-you-with-new-virtual-reality-flight-trainer.html. (accessed: 09.04.2020).

[37] H. G. Hoffman, T. Richards, B. Coda, A. Richards, and S. R. Sharar, "The illusion of presence in immersive virtual reality during an fmri brain scan," *CyberPsychology & Behavior*, vol. 6, no. 2, pp. 127–131, 2003. DOI: 10.1089/109493103321640310.

[38] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application," *NeuroImage*, vol. 63, no. 2, pp. 921–935, Mar. 2012. DOI: 10.1016/j.neuroimage.2012.03.049.

[39] V. Quaresima, S. Bisconti, and M. Ferrari, "A brief review on the use of functional near-infrared spectroscopy (fnirs) for language imaging studies in human newborns and adults," *Brain and language*, vol. 121, no. 2, pp. 79–89, Apr. 2011. DOI: 10.1016/j.bandl.2011.03.009.

[40] P. Pinti, F. Scholkmann, A. Hamilton, P. Burgess, and I. Tachtsidis, "Current status and issues regarding pre-processing of fnirs neuroimaging data: An investigation of diverse signal filtering methods within a general linear model framework," *Frontiers in Human Neuroscience*, vol. 12, p. 505, 2019. DOI: 10.3389/fnhum.2018.00505.

[41] D. A. Boas, C. E. Elwell, M. Ferrari, and G. Taga, "Twenty years of functional near-infrared spectroscopy: Introduction for the special issue," *NeuroImage*, vol. 85, no. 1, pp. 1–5, 2014. DOI: https://doi.org/10.1016/j.neuroimage.2013.11.033.

[42] T. Wilcox and M. Biondi, "Fnirs in the developmental sciences," *WIREs Cognitive Science*, vol. 6, no. 3, pp. 263–283, 2015. DOI: 10.1002/wcs.1343.

[43] P. Pinti, I. Tachtsidis, A. Hamilton, J. Hirsch, C. Aichelburg, S. Gilbert, and P. W. Burgess, "The present and future use of functional near-infrared spectroscopy (fnirs) for cognitive neuroscience," *Annals of the New York Academy of Sciences*, vol. 1464, no. 1, pp. 5–29, 2020. DOI: 10.1111/nyas.13948.

[44] F. Scholkmann, S. Kleiser, A. Metz, R. Zimmermann, J. Pavia, U. Wolf, and M. Wolf, "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *NeuroImage*, vol. 85, no. 1, May 2013. DOI: 10.1016/j.neuroimage.2013.05.004.

[45] S. Prahl. (1998). Tabulated molar extinction coefficient for hemoglobin in water, [Online]. Available: https://omlc.org/spectra/hemoglobin/summary.html. (accessed: 09.04.2020).

[46] L.-C. Chen, "Cortical plasticity in cochlear implant users," PhD thesis, Jan. 2016. DOI: 10.13140/RG.2.2.33251.76324.

[47] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," *Physics in Medicine and Biology*, vol. 33, no. 12, pp. 1433–1442, Dec. 1988. DOI: 10.1088/0031-9155/33/12/008.

[48] FieldTrip. (2018). Preprocessing and averaging of single-channel nirs data, [Online]. Available: http://www.fieldtriptoolbox.org/tutorial/nirs_singlechannel/. (accessed: 25-03-2020).

[49] L. M. Hocke, I. K. Oni, C. C. Duszynski, A. V. Corrigan, B. deB. Frederick, and J. F. Dunn, "Automated processing of fnirs data—a visual guide to the pitfalls and consequences," *Algorithms*, vol. 11, no. 5, pp. 67–92, 2018. DOI: 10.3390/a11050067.

[50] Artinis Medical Systems. (2017). The do's and dont's of baselines, [Online]. Available: https://www.artinis.com/blogpost-all/2017/10/27/the-dos-and-donts-of-baselines. (accessed: 24.03.2020).

[51] A. Chaddad, "Brain function diagnosis enhanced using denoised fnirs raw signals," *Journal of biomedical science and engineering*, vol. 7, no. 4, pp. 218–227, Nov. 2013. DOI: 10.4236/jbise.2014.74025.

[52] E. Peck, D. Afergan, and R. Jacob, "Investigation of fnirs brain sensing as input to information filtering systems," Mar. 2013, pp. 142–149. DOI: 10.1145/2459236.2459261.

[53] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," *NeuroImage*, vol. 59, no. 1, pp. 36–47, 2012, Neuroergonomics: The human brain in action and at work. DOI: 10.1016/j.neuroimage.2011.06.023.

[54] R. McKendrick, R. Parasuraman, R. Murtza, A. Formwalt, W. Baccus, M. Paczynski, and H. Ayaz, "Into the Wild: Neuroergonomic Differentiation of Hand-Held and Augmented Reality Wearable Displays during Outdoor Navigation with Functional Near Infrared Spectroscopy.," *Frontiers in Human Neuroscience*, vol. 10, no. 216, 2016. DOI: 10.3389/fnhum.2016.00216.

[55] A. Unni, K. Ihme, H. Surm, L. Weber, A. Lüdtke, D. Nicklas, M. Jipp, and J. W. Rieger, "Brain activity measured with fNIRS for the prediction of cognitive workload.," *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2015. DOI: 10.1109/CogInfoCom.2015.7390617.

[56] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring Mental Workload with EEG+fNIRS.," *Frontiers in Human Neuroscience*, vol. 11, no. 359, 2017. DOI: 10.3389/fnhum.2017.00359.

[57] E. M. Peck, D. Afergan, B. F. Yuksel, F. Lalooses, and R. J. K. Jacob, "Using fnirs to measure mental workload in the real world," in *Advances in Physiological Computing*, S. H. Fairclough and K. Gilleade, Eds. London: Springer London, 2014, pp. 117–139, ISBN: 978-1-4471-6392-3. DOI: 10.1007/978-1-4471-6392-3_6.

[58] G. Durantin, J.-F. Gagnon, S. Tremblay, and F. Dehais, "Using near infrared spectroscopy and heart rate variability to detect mental overload.," *Behavioural Brain Research*, vol. 259, pp. 16–23, 2014. DOI: 10.1016/j.bbr.2013.10.042.

[59] S. W. Hincks, D. Afergan, and R. J. K. Jacob, "Using fnirs for real-time cognitive workload assessment," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, D. D. Schmorrow and C. M. Fidopiastis, Eds., Springer International Publishing, 2016, pp. 198–208, ISBN: 978-3-319-39955-3.

[60] M. Pike, H. Maior, M. Wilson, and S. Sharples, "Continuous detection of workload overload: An fnirs approach," Apr. 2014. DOI: 10.1201/b16742-79.

[61] F. Fishburn, M. Norr, A. Medvedev, and C. Vaidya, "Sensitivity of fNIRS to cognitive state and load.," *Frontiers in Human Neuroscience*, vol. 8, no. 76, 2014. DOI: 10.3389/fnhum.2014.00076.

[62] S. Bunce, K. Izzetoglu, H. Ayaz, P. Shewokis, M. Izzetoglu, K. Pourrezaei, and B. Onaral, "Implementation of fnirs for monitoring levels of expertise and mental workload," Jul. 2011, pp. 13–22. DOI: 10.1007/978-3-642-21852-1_2.

[63] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS.," *Frontiers in Human Neuroscience*, vol. 7, no. 935, 2014. DOI: 10.3389/fnhum.2013.00935.

[64] T. Gateau, H. Ayaz, and F. Dehais, "In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI.," *Frontiers in Human Neuroscience*, vol. 12, no. 187, 2018. DOI: 10.3389/fnhum.2018.00187.

[65] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang, "Using fnirs brain sensing to evaluate information visualization interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13, Paris, France: Association for Computing Machinery, 2013, pp. 473–482, ISBN: 9781450318990. DOI: 10.1145/2470654.2470723.

[66] F. Al-Shargie, M. Kiguchi, N. Badruddin, S. Dass, A. F. Mohd Hani, and T. B. Tang, "Mental stress assessment using simultaneous measurement of eeg and fnirs," *Biomedical Optics Express*, vol. 7, pp. 3882–3898, Sep. 2016. DOI: 10.1364/BOE.7.003882.

[67] R. Arefi, K. Setarehdan, and A. Motie Nasrabadi, "Classification of mental stress levels by analyzing fnirs signal using linear and non-linear features," *International Clinical Neuroscience Journal*, vol. 5, no. 2, pp. 55–61, Jun. 2018. DOI: 10.15171/icnj.2018.11.

[68] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Mental stress grading based on fnirs signals," vol. 2016, Aug. 2016, pp. 5140–5143. DOI: 10.1109/EMBC.2016.7591884.

[69] M. Tanida, M. Katsuyama, and K. Sakatani, "Relation between mental stress-induced prefrontal cortex activity and skin conditions: A near-infrared spectroscopy study," *Brain Research*, vol. 1184, pp. 210–216, 2007. DOI: https://doi.org/10.1016/j.brainres.2007.09.058.

[70] D. Rosenbaum, P. Hilsendegen, M. Thomas, F. B. Haeussinger, F. G. Metzger, H.-C. Nuerk, A. J. Fallgatter, V. Nieratschker, and A.-C. Ehlis, "Cortical hemodynamic changes during the trier social stress test: An fnirs study," *NeuroImage*, vol. 171, pp. 107–115, 2018. DOI: https://doi.org/10.1016/j.neuroimage.2017.12.061.

[71] M. Parent, V. Peysakhovich, K. Mandrick, S. Tremblay, and M. Causse, "The diagnosticity of psychophysiological signatures: Can we disentangle mental workload from acute stress with ecg and fnirs?" *International Journal of Psychophysiology*, vol. 146, pp. 139–147, 2019. DOI: https://doi.org/10.1016/j.ijpsycho.2019.09.005.

[72] D. Rosenbaum, M. Thomas, P. Hilsendegen, F. G. Metzger, F. B. Haeussinger, H.-C. Nuerk, A. J. Fallgatter, V. Nieratschker, and A.-C. Ehlis, "Stress-related dysfunction of the right inferior frontal cortex in high ruminators: An fnirs study," *NeuroImage: Clinical*, vol. 18, pp. 510–517, 2018. DOI: https://doi.org/10.1016/j.nicl.2018.02.022.

[73] H. Modi, H. Singh, G.-Z. Yang, A. Darzi, and D. Leff, "Neural correlates of stress resilience in the operating room," *Journal of the American College of Surgeons*, vol. 227, no. 4, Oct. 2018. DOI: 10.1016/j.jamcollsurg.2018.08.563.

[74] D. Rosenbaum, E. J. Leehr, J. Rubel, M. J. Maier, V. Pagliaro, K. Deutsch, J. Hudak, F. G. Metzger, A. J. Fallgatter, and A.-C. Ehlis, "Cortical oxygenation during exposure therapy – in situ fnirs measurements in arachnophobia," *NeuroImage: Clinical*, vol. 26, 2020. DOI: https://doi.org/10.1016/j.nicl.2020.102219.

[75] A. Landowska, D. Roberts, P. Eachus, and A. Barrett, "Within- and between-session prefrontal cortex response to virtual reality exposure therapy for acrophobia," *Frontiers in Human Neuroscience*, vol. 12, no. 362, 2018. DOI: 10.3389/fnhum.2018.00362.

[76] E. Glotzbach, A. Mühlberger, K. Gschwendtner, A. J. Fallgatter, P. Pauli, and M. J. Herrmann, "Prefrontal brain activation during emotional processing: A functional near infrared spectroscopy study (fnirs)," *The Open Neuroimaging Journal*, vol. 5, pp. 33–39, 2011. DOI: 10.2174/1874440001105010033.

[77] D. Zhang, Y. Zhou, X. Hou, Y. Cui, and C. Zhou, "Discrimination of emotional prosodies in human neonates: A pilot fnirs study," *Neuroscience Letters*, vol. 658, pp. 62–66, 2017. DOI: https://doi.org/10.1016/j.neulet.2017.08.047.

[78] A. Köchel, F. Schöngassner, and A. Schienle, "Cortical activation during auditory elicitation of fear and disgust: A near-infrared spectroscopy (nirs) study," *Neuroscience Letters*, vol. 549, pp. 197–200, 2013. DOI: https://doi.org/10.1016/j.neulet.2013.06.062.

[79] A. Roos, F. Robertson, C. Lochner, B. Vythilingum, and D. J. Stein, "Altered prefrontal cortical function during processing of fear-relevant stimuli in pregnancy," *Behavioural Brain Research*, vol. 222, no. 1, pp. 200–205, 2011. DOI: https://doi.org/10.1016/j.bbr.2011.03.055.

[80] Q. Ma, Y. Huang, and L. Wang, "Left prefrontal activity reflects the ability of vicarious fear learning: A functional near-infrared spectroscopy study," *TheScientificWorldJournal*, vol. 2013, p. 652 542, Nov. 2013. DOI: 10.1155/2013/652542.

[81] A. Landowska, "Measuring Prefrontal Cortex Response to Virtual Reality Exposure Therapy in Freely Moving Participants," PhD thesis, University of Salford, 2018.

[82] S. V. Tupak, T. Dresler, A. Guhn, A.-C. Ehlis, A. J. Fallgatter, P. Pauli, and M. J. Herrmann, "Implicit emotion regulation in the presence of threat: Neural and autonomic correlates," *NeuroImage*, vol. 85, no. 1, pp. 372–379, 2014, Celebrating 20 Years of Functional Near Infrared Spectroscopy (fNIRS). DOI: https://doi.org/10.1016/j.neuroimage.2013.09.066.

[83] M. Strait and M. Scheutz, "Using near infrared spectroscopy to index temporal changes in affect in realistic human-robot interactions," 2014, pp. 385–392. DOI: 10.5220/0004902203850392.

[84] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," Sep. 2013, pp. 832–837. DOI: 10.1109/ACII.2013.156.

[85] R. Aoki, H. Sato, T. Katura, K. Utsugi, H. Koizumi, R. Matsuda, and A. Maki, "Relationship of negative mood with prefrontal cortex activity during working memory tasks: An optical topography study," *Neuroscience Research*, vol. 70, no. 2, pp. 189–196, 2011. DOI: https://doi.org/10.1016/j.neures.2011.02.011.

[86] R. Aoki, H. Sato, T. Katura, R. Matsuda, and H. Koizumi, "Correlation between prefrontal cortex activity during working memory tasks and natural mood independent of personality effects: An optical topography study," *Psychiatry Research: Neuroimaging*, vol. 212, no. 1, pp. 79–87, 2013. DOI: https://doi.org/10.1016/j.pscychresns.2012.10.009.

[87] T. Numata, M. Kiguchi, and H. Sato, "Multiple-time-scale analysis of attention as revealed by eeg, nirs, and pupil diameter signals during a free recall task: A multimodal measurement approach," *Frontiers in Neuroscience*, vol. 13, Dec. 2019. DOI: 10.3389/fnins.2019.01307.

[88] A. Harrivel, D. Weissman, D. Noll, T. Huppert, and S. Peltier, "Dynamic filtering improves attentional state prediction with fnirs," *Biomedical Optics Express*, vol. 7, no. 3, p. 979, Mar. 2016. DOI: 10.1364/BOE.7.000979.

[89] C. Bogler, J. Mehnert, J. Steinbrink, and J.-D. Haynes, "Decoding vigilance with nirs," *PLOS ONE*, vol. 9, no. 7, Jul. 2014. DOI: 10.1371/journal.pone.0101729.

[90] Z. Zhang, X. Jiao, J. Jiang, J. Pan, Y. Cao, H. Yang, and F. Xu, "Passive bci based on sustained attention detection: An fnirs study," in *Advances in Brain Inspired Cognitive Systems*, C.-L. Liu, A. Hussain, B. Luo, K. C. Tan, Y. Zeng, and Z. Zhang, Eds., Springer International Publishing, 2016, pp. 220–227, ISBN: 978-3-319-49685-6.

[91] G. Derosière, S. Dalhoumi, S. Perrey, G. Dray, and T. Ward, "Towards a near infrared spectroscopy-based estimation of operator attentional state," *PLOS ONE*, vol. 9, no. 3, Mar. 2014. DOI: 10.1371/journal.pone.0092045.

[92] M. Khan and K.-S. Hong, "Passive bci based on drowsiness detection: An fnirs study," *Biomedical Optics Express*, vol. 6, no. 10, pp. 4063–78, Oct. 2015. DOI: 10.1364/BOE.6.004063.

[93] X.-S. Hu, K.-S. Hong, and S. Ge, "Fnirs-based online deception decoding," *Journal of Neural Engineering*, vol. 9, no. 2, Feb. 2012. DOI: 10.1088/1741-2560/9/2/026012.

[94] F. Tian, V. Sharma, F. Kozel, and H. Liu, "Functional near-infrared spectroscopy to investigate hemodynamic responses to deception in the prefrontal cortex," *Brain research*, vol. 1303, Sep. 2009. DOI: 10.1016/j.brainres.2009.09.085.

[95] X. P. Ding, X. Gao, G. Fu, and K. Lee, "Neural correlates of spontaneous deception: A functional near-infrared spectroscopy (fnirs)study," *Neuropsychologia*, vol. 51, no. 4, Jan. 2013. DOI: 10.1016/j.neuropsychologia.2012.12.018.

[96] X. P. Ding, L. Sai, G. Fu, J. Liu, and K. Lee, "Neural correlates of second-order verbal deception: A functional near-infrared spectroscopy (fnirs) study," *NeuroImage*, vol. 87, pp. 505–514, Feb. 2014. DOI: 10.1016/j.neuroimage.2013.10.023.

[97] L. Sai, X. Zhou, X. P. Ding, G. Fu, and B. Sang, "Detecting concealed information using functional near-infrared spectroscopy," *Brain Topography*, vol. 27, no. 5, Feb. 2014. DOI: 10.1007/s10548-014-0352-z.

[98] H. Hosseini, Y. Mano, M. Rostami, M. Takahashi, M. Sugiura, and R. Kawashima, "Decoding what one likes or dislikes from single-trial fnirs measurements," *Neuroreport*, vol. 22, no. 6, pp. 269–273, Mar. 2011. DOI: 10.1097/WNR.0b013e3283451f8f.

[99] K. Laghari, R. Gupta, S. Arndt, J.-N. Voigt-Antons, S. Mollery, and T. Falk, "Characterization of human emotions and preferences for text-to-speech systems using multimodal neuroimaging methods," May 2014. DOI: 10.1109/CCECE.2014.6901142.

[100] U. Kreplin and S. Fairclough, "Activation of the rostromedial prefrontal cortex during the experience of positive emotion in the context of esthetic experience. an fnirs study," *Frontiers in Human Neuroscience*, vol. 7, no. 879, 2013. DOI: 10.3389/fnhum.2013.00879.

[101] E. Vassena, R. Gerrits, J. Demanet, T. Verguts, and R. Siugzdaite, "Anticipation of a mentally effortful task recruits dorsolateral prefrontal cortex: An fnirs validation study," *Neuropsychologia*, vol. 123, pp. 106–115, 2019. DOI: https://doi.org/10.1016/j.neuropsychologia.2018.04.033.

[102] M.-Y. Wang, F.-M. Lu, Z. Hu, J. Zhang, and Z. Yuan, "Optical mapping of prefrontal brain connectivity and activation during emotion anticipation," *Behavioural Brain Research*, vol. 350, pp. 122–128, 2018.

[103] M. Suzuki, I. Miyai, T. Ono, and K. Kubota, "Activities in the frontal cortex and gait performance are modulated by preparation. an fnirs study," *NeuroImage*, vol. 39, no. 2, pp. 600–607, 2008. DOI: https://doi.org/10.1016/j.neuroimage.2007.08.044.

[104] L. Hirshfield, P. Bobko, A. Barelka, N. Sommer, and S. Velipasalar, "Toward interfaces that help users identify misinformation online: Using fnirs to measure suspicion," *Augmented Human Research*, vol. 4, no. 1, Apr. 2019. DOI: 10.1007/s41133-019-0011-8.

[105] L. Hirshfield, P. Bobko, A. Barelka, S. Hirshfield, M. Farrington, S. Gulbronson, and D. Paverman, "Using noninvasive brain measurement to explore the psychological effects of computer malfunctions on users during human-computer interactions," *Advances in Human-Computer Interaction*, vol. 2014, no. 2, pp. 1–13, 2014. DOI: 10.1155/2014/101038.

[106] K. Ihme, A. Unni, M. Zhang, J. W. Rieger, and M. Jipp, "Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy," *Frontiers in Human Neuroscience*, vol. 12, p. 327, 2018. DOI: 10.3389/fnhum.2018.00327.

[107] K. Ihme, A. Unni, J. Rieger, and M. Jipp, "Assessing driver frustration using functional near infrared spectroscopy (fnirs)," in *1st International Conference on Neuroergonomics*, Oct. 2016.

[108] K. Lange, L. M. Williams, A. W. Young, E. T. Bullmore, M. J. Brammer, S. C. Williams, J. A. Gray, and M. L. Phillips, "Task instructions modulate neural responses to fearful facial expressions," *Biological Psychiatry*, vol. 53, no. 3, pp. 226–232, 2003. DOI: https://doi.org/10.1016/S0006-3223(02)01455-5.

[109] M. Nomura, H. Ohira, K. Haneda, T. Iidaka, N. Sadato, T. Okada, and Y. Yonekura, "Functional association of the amygdala and ventral prefrontal cortex during cognitive evaluation of facial expressions primed by masked angry faces: An event-related fmri study," *NeuroImage*, vol. 21, no. 1, pp. 352–363, 2004. DOI: https://doi.org/10.1016/j.neuroimage.2003.09.021.

[110] A. Etkin and T. D. Wager, "Functional neuroimaging of anxiety: A meta-analysis of emotional processing in ptsd, social anxiety disorder, and specific phobia," *American Journal of Psychiatry*, vol. 164, no. 10, pp. 1476–1488, 2007. DOI: 10.1176/appi.ajp.2007.07030504.

[111] L. M. Shin and I. Liberzon, "The neurocircuitry of fear, stress, and anxiety disorders," *Neuropsychopharmacol*, vol. 35, pp. 169–191, 2010. DOI: https://doi.org/10.1038/npp.2009.83.

[112] R. B. Price, D. A. Eldreth, and J. Mohlman, "Deficient prefrontal attentional control in late-life generalized anxiety disorder: An fmri investigation," *Translational Psychiatry*, vol. 1, no. 10, 2011. DOI: 10.1038/tp.2011.46.

[113] A. L. Krain, A. M. Wilson, R. Arbuckle, F. X. Castellanos, and M. P. Milham, "Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making," *NeuroImage*, vol. 32, no. 1, pp. 477–484, 2006. DOI: https://doi.org/10.1016/j.neuroimage.2006.02.047.

[114] A. Dimoka, "What does the brain tell us about trust and distrust? evidence from a functional neuroimaging study," *MIS Quarterly*, vol. 34, no. 2, pp. 373–396, Jun. 2010. DOI: 10.2307/20721433.

[115] B. Seraglia, L. Gamberini, K. Priftis, P. Scatturin, M. Martinelli, and S. Cutini, "An exploratory fnirs study with immersive virtual reality: A new method for technical implementation," *Frontiers in Human Neuroscience*, vol. 5, 2011. DOI: 10.3389/fnhum.2011.00176.

[116] D. Dong, L. K. Wong, and Z. Luo, "Assessment of prospective memory using fnirs in immersive virtual reality environment," *Journal of Behavioral and Brain Science*, vol. 7, pp. 247–258, 2017. DOI: 10.4236/jbbs.2017.76018.

[117] D. Dong, L. K. Wong, and Z. Luo, "Assess ba10 activity in slide-based and immersive virtual reality prospective memory task using functional near-infrared spectroscopy (fnirs)," *Applied Neuropsychology: Adult*, vol. 26, no. 5, pp. 465–471, 2018. DOI: 10.1080/23279095.2018.1443104.

[118] G. Kim, N. Buntain, L. Hirshfield, M. Costa, and T. Chock, "Processing racial stereotypes in virtual reality: An exploratory study using functional near-infrared spectroscopy (fnirs)," in *Augmented Cognition. HCII 2019. Lectures Notes in Computer Science*, D. Schmorrow and C. Fidopiastis, Eds. Springer Cham, 2019, pp. 407–417, ISBN: 978-3-030-22418-9. DOI: 10.1007/978-3-030-22419-6_29.

[119] J. Hudak, F. Blume, T. Dresler, F. B. Haeussinger, T. J. Renner, A. J. Fallgatter, C. Gawrilow, and A.-C. Ehlis, "Near-infrared spectroscopy-based frontal lobe neurofeedback integrated in virtual reality modulates brain and behavior in highly impulsive adults," *Frontiers in Human Neuroscience*, vol. 11, 2017. DOI: 10.3389/fnhum.2017.00425.

[120] E. Aksoy, K. Izzetoglu, E. Baysoy, A. Agrali, D. Kitapcioglu, and B. Onaral, "Performance monitoring via functional near infrared spectroscopy for virtual reality based basic life support training," *Frontiers in Human Neuroscience*, vol. 13, 2019. DOI: 10.3389/fnins.2019.01336.

[121] B. K. Wiederhold, D. P. Jang, S. I. Kim, and M. D. Wiederhold, "Physiological monitoring as an objective tool in virtual reality therapy," *CyberPsychology & Behavior*, vol. 5, no. 1, pp. 77–82, 2004. DOI: 10.1089/109493102753685908.

[122] S. M. Peterson, E. Furuichi, and D. P. Ferris, "Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading," *PLOS ONE*, vol. 13, no. 7, 2018. DOI: 10.1371/journal.pone.0200306.

[123] P. Maron, V. Powell, and W. Powell, "The differential effect of neutral and fear-stimulus virtual reality exposure on physiological indicators of anxiety in acrophobia," in *International Conference on Disability, Virtual Reality and Associated Technologies*, Sep. 2016.

[124] J. Diemer, N. Lohkamp, A. Mühlberger, and P. Zwanzger, "Fear and physiological arousal during a virtual height challenge–effects in patients with acrophobia and healthy controls," *Journal of Anxiety Disorders*, vol. 37, no. 9, pp. 30–39, 2016. DOI: 10.1016/j.janxdis.2015.10.007.

[125] M. Lee, G. Bruder, and G. Welch, "The virtual pole: Exploring human responses to fear of heights in immersive virtual environments," *Journal of Virtual Reality and Broadcasting*, vol. 14, no. 6, 2019. DOI: 10.20385/1860-2037/14.2017.6.

[126] J. Šalkevicius, R. Damaševičius, R. Maskeliunas, and I. Laukienė, "Anxiety level recognition for virtual reality therapy system using physiological signals," *Electronics*, vol. 8, no. 9, 2019. DOI: 10.3390/electronics8091039.

[127] F. Hu, H. Wang, J. Chen, and J. Gong, "Research on the characteristics of acrophobia in virtual altitude environment," in *2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*, 2018, pp. 238–243.

[128] W. Handouzi, C. Maaoui, A. Pruski, A. Moussaoui, and Y. Bendiouis, "Short-term anxiety recognition induced by virtual reality exposure for phobic people," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 3145–3150.

[129] D. S. Collingridge, "A primer on quantitized data analysis and permutation testing," *Journal of Mixed Methods Research*, vol. 7, no. 1, pp. 81–97, 2013. DOI: 10.1177/1558689812454457.

[130] P. Legendre and L. Legendre, "Statistical testing by permutation," in *Numerical Ecology*. Amsterdam: Elsevier Science BV, 1998, pp. 17–26, ISBN: 9780080523170.

[131] M. Marozzi, "Some remarks about the number of permutations one should consider to perform a permutation test," *Statistica*, vol. 64, no. 1, pp. 193–201, 2004. DOI: 10.6092/issn.1973-2201/32.

[132] P. Cohen and D. Jensen, "Overfitting explained," *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 115–122, Apr. 1997.

[133] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, no. 4, pp. 870–878, 2002. DOI: https://doi.org/10.1006/nimg.2001.1037.

[134] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, G. Casella, S. Fienberg, and I. Olkin, Eds. Springer, 2014, ISBN: 9781461471370.

[135] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, 169-190. DOI: 10.3233/AIC-170729.

[136] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006, ISBN: 0387310738.

[137] A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, G. Casella, S. Fienberg, and I. Olkin, Eds. Springer, 2008, ISBN: 9780387781891.

[138] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, 2008, ISBN: 9780387848587.

[139] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, vol. 5, 1992, pp. 144–152.

[140] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, Second Edition. Pearson, 2018, ISBN: 9780133128901.

[141] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. DOI: https://doi.org/10.1007/BF00994018.

[142] D. C. Cohen, "Comparison of self-report and overt-behavioral procedures for assessing acrophobia," *Behavior Therapy*, vol. 8, no. 1, pp. 17–23, 1977. DOI: https://doi.org/10.1016/S0005-7894(77)80116-0.

[143] M. M. Antony, "Measures for specific phobia," in *Practitioner's Guide to Empirically Based Measures of Anxiety*, M. M. Antony, S. M. Orsillo, and L. Roemer, Eds. Boston, MA: Springer US, 2001, pp. 133–158, ISBN: 978-0-306-47628-0. DOI: 10.1007/0-306-47628-2_12.

[144] Facebook Technologies, LLC. (2020). Oculus rift s, [Online]. Available: https://www.oculus.com/rift-s/. (accessed: 01.09.2020).

[145] Janet Orega. (2018). Virtual reality - how does it work? [Online]. Available: https://medium.com/@georginaoregx/virtual-reality-how-does-it-work-7af29f650b23. (accessed: 01.05.2020).

[146] Unity Technologies. (2020). Unity real-time development platform., [Online]. Available: https://unity.com/. (accessed: 02.04.2020).

[147] J. Wolpe, *The Practice of Behavior Therapy*. New York: Pergamon Press, 1969, ISBN: 0080065635.

[148] C. L. Benjamin, K. A. O'Neil, S. A. Crawley, and R. S. Beidas, "Patterns and predictors of subjective units of distress in anxious youth," *Behavioural and Cognitive Psychotherapy*, vol. 38, no. 4, pp. 497–504, 2010. DOI: 10.1017/S1352465810000287.

[149] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 266–281, 2001. DOI: 10.1162/105474601300343603.

[150] H. T. Regenbrecht, T. W. Schubert, and F. Friedmann, "Measuring the sense of presence and its relations to fear of heights in virtual environments," *International Journal of Human–Computer Interaction*, vol. 10, no. 3, pp. 233–249, 1998. DOI: 10.1207/s15327590ijhc1003_2.

[151] H. M. Peperkorn, J. Diemer, and A. Mühlberger, "Temporal dynamics in the relation between presence and fear in virtual reality," *Computers in Human Behavior*, vol. 48, pp. 542–547, 2015. DOI: https://doi.org/10.1016/j.chb.2015.02.028.

[152] IGroup. (2016). Igroup presence questionnaire (ipq) overview, [Online]. Available: http://www.igroup.org/pq/ipq/index.php. (accessed: 21.04.2020).

[153] F. A. Fishburn, R. S. Ludlum, C. J. Vaidya, and A. V. Medvedev, "Temporal derivative distribution repair (tddr): A motion correction method for fnirs," *NeuroImage*, vol. 184, pp. 171–179, 2019. DOI: https://doi.org/10.1016/j.neuroimage.2018.09.025.

[154] X. Cui, S. Bray, and A. L. Reiss, "Functional near infrared spectroscopy (nirs) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics," *NeuroImage*, vol. 49, no. 4, pp. 3039–3046, 2010. DOI: https://doi.org/10.1016/j.neuroimage.2009.11.050.

[155] K. L. Perdue, A. Westerlund, S. A. McCormick, and I. Charles A. Nelson, "Extraction of heart rate from functional near-infrared spectroscopy in infants," *Journal of Biomedical Optics*, vol. 19, no. 6, p. 67 010, 2014. DOI: 10.1117/1.JBO.19.6.067010.

[156] L. Holper, E. Seifritz, and F. Scholkmann, "Short-term pulse rate variability is better characterized by functional near-infrared spectroscopy than by photoplethysmography," *Journal of Biomedical Optics*, vol. 21, no. 9, p. 091 308, 2016. DOI: 10.1117/1.JBO.21.9.091308.

[157] S. K. S. Naser Hakimi, "Stress assessment by means of heart rate derived from functional near-infrared spectroscopy," *Journal of Biomedical Optics*, vol. 23, no. 11, p. 115 001, 2018. DOI: 10.1117/1.JBO.23.11.115001.

[158] M. Mirbagheri, N. Hakimi, E. Ebrahimzadeh, and S. Setarehdan, "Quality analysis of heart rate derived from functional near-infrared spectroscopy in stress assessment," *Informatics in Medicine Unlocked*, vol. 18, p. 100 286, 2020. DOI: https://doi.org/10.1016/j.imu.2019.100286.

[159] P. Lachert, D. Janusek, P. Pulawski, A. Liebert, D. Milej, and K. J. Blinowska, "Coupling of oxy- and deoxyhemoglobin concentrations with eeg rhythms during motor task," *Scientific Reports*, vol. 7, p. 15 414, 2017. DOI: https://doi.org/10.1038/s41598-017-15770-2.

[160] M. N. A. Khan, M. R. Bhutta, and K. Hong, "Task-specific stimulation duration for fnirs brain-computer interface," *IEEE Access*, vol. 8, pp. 89 093–89 105, 2020.

[161] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl, "Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, bmi, and time of year: Retrospective, longitudinal cohort study of 92,457 adults," *PLoS ONE*, vol. 15, no. 2, e0227709, 2020. DOI: https://doi.org/10.1371/journal.pone.0227709.

[162] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, no. 258, 2017. DOI: 10.3389/fpubh.2017.00258.

Appendices

Appendix A

Deriving Equations for Δ[HbO] and Δ[HbR]

According to the MBLL, the OD can be found with the following equation:

$$OD(t,\lambda) = -\log_{10}\Big(\frac{I(t,\lambda)}{I_0(t,\lambda)}\Big) = \sum_i \varepsilon_i(\lambda)\cdot\Delta c_i\cdot l\cdot DPF(\lambda) \tag{A.1}$$

Therefore, the change in OD (ΔOD) over time can be found by inserting Δt into the equation and solving for two different points in time:

$$\Delta OD(\Delta t,\lambda) = -\log 10\Big(\frac{I(\lambda,t_1)}{I_0(\lambda,t_1)}\Big) - \Big(-\log 10\frac{I(\lambda,t_0)}{I_0(\lambda,t_0)}\Big) \tag{A.2}$$

with $\Delta t = t_0 - t_1$. The emitted light intensity I_0 remains the same over time, hence $I_0(t_1) = I_0(t_0) = I_0$. Therefore:

$$\begin{aligned}\Delta OD(\Delta t,\lambda) &= \log_{10}\Big(\frac{I(\lambda,t_0)}{I_0}\Big) - \log_{10}\Big(\frac{I(\lambda,t_1)}{I_0}\Big)\\ &= \log_{10}\Big(\frac{I(\lambda,t_0)}{I_0}\cdot\frac{I_0}{I(\lambda,t_1)}\Big)\\ &= \log_{10}\Big(\frac{I(\lambda,t_0)}{I(\lambda,t_1)}\Big) = -\log_{10}\Big(\frac{I(t_1,\lambda)}{I(t_0,\lambda)}\Big)\end{aligned}$$

$$\Delta OD(\Delta t,\lambda) = -\log_{10}\Big(\frac{I(t_1,\lambda)}{I(t_0,\lambda)}\Big) = \sum_i \varepsilon_i(\lambda)\cdot\Delta c_i\cdot l\cdot DPF(\lambda) \tag{A.3}$$

Based on equation A.3, the change in OD for wavelength λ_1 can be computed as follows:

$$\begin{aligned}\Delta OD(\Delta t,\lambda_1) &= \varepsilon_{HbR}(\lambda_1)\cdot\Delta[HbR]\cdot DPF(\lambda_1)\cdot l + \varepsilon_{HbO}(\lambda_1)\cdot\Delta[HbO]\cdot DPF(\lambda_1)\cdot l\\ &= l\cdot DPF(\lambda_1)\cdot(\varepsilon_{HbR}(\lambda_1)\cdot\Delta[HbR] + \varepsilon_{HbO}(\lambda_1)\cdot\Delta[HbO])\end{aligned} \tag{A.4}$$

with $\Delta[HbO]$ and $\Delta[HbR]$ denoting Δc_{HbO} and Δc_{HbR}, respectively. Likewise, the change in OD for wavelength λ_2 can be found as follows:

$$\Delta OD(\Delta t, \lambda_2) = l \cdot DPF(\lambda_2) \cdot (\varepsilon_{HbR}(\lambda_2) \cdot \Delta[HbR] + \varepsilon_{HbO}(\lambda_2) \cdot \Delta[HbO]) \tag{A.5}$$

Equation A.5 can be solved for $\Delta[HbR]$:

$$\begin{aligned}
\frac{\Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2)} &= \varepsilon_{HbR}(\lambda_2) \cdot \Delta[HbR] + \varepsilon_{HbO}(\lambda_2) \cdot \Delta[HbO] \\
\frac{\Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2)} - \varepsilon_{HbO}(\lambda_2) \cdot \Delta[HbO] &= \varepsilon_{HbR}(\lambda_2) \cdot \Delta[HbR] \\
\Delta[HbR] &= \frac{\Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)} - \frac{\varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)} \cdot \Delta[HbO]
\end{aligned} \tag{A.6}$$

Now equation A.6 can be substituted into equation A.4, which can be solved to find the expression for $\Delta[HbO]$:

$$\Delta OD(\Delta t, \lambda_1) = l \cdot DPF(\lambda_1) \cdot (\varepsilon_{HbR}(\lambda_1) \cdot \Big(\frac{\Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)} - \frac{\varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)} \cdot \Delta[HbO]\Big)) + \varepsilon_{HbO}(\lambda_1) \cdot \Delta[HbO]$$

$$\frac{\Delta OD(\Delta t, \lambda_1)}{l \cdot DPF(\lambda_1)} = \frac{\varepsilon_{HbR}(\lambda_1) \cdot \Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)} - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)} \cdot \Delta[HbO] + \varepsilon_{HbO}(\lambda_1) \cdot \Delta[HbO]$$

$$\frac{\Delta OD(\Delta t, \lambda_1)}{l \cdot DPF(\lambda_1)} - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)} = \Delta[HbO] \cdot \Big(\varepsilon_{HbO}(\lambda_1) - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)}\Big)$$

$$\Big[\frac{\Delta OD(\Delta t, \lambda_1)}{l \cdot DPF(\lambda_1)} - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)}\Big] \cdot \Big[\varepsilon_{HbO}(\lambda_1) - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)}\Big]^{-1} = \Delta[HbO]$$

$$\begin{aligned}
&\Big[\varepsilon_{HbO}(\lambda_1) \cdot \frac{\varepsilon_{HbR}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)} - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)}\Big]^{-1} = \\
&\Big[\frac{\varepsilon_{HbO}(\lambda_1) \cdot \varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}{\varepsilon_{HbR}(\lambda_2)}\Big]^{-1} = \\
&\frac{\varepsilon_{HbR}(\lambda_2)}{\varepsilon_{HbO}(\lambda_1) \cdot \varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbR}(\lambda_1) \cdot \varepsilon_{HbO}(\lambda_2)}
\end{aligned}$$

$$\Delta[HbO] = \frac{\varepsilon_{HbR}(\lambda_2)}{\varepsilon_{HbO}(\lambda_1) \cdot \varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbO}(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_1)} \cdot \Big(\frac{\Delta OD(\Delta t, \lambda_1)}{l \cdot DPF(\lambda_1)} - \frac{\varepsilon_{HbR}(\lambda_1) \cdot \Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_2)}\Big)$$

$$\Delta[HbO] = \frac{\varepsilon_{HbR}(\lambda_2) \cdot \frac{\Delta OD(\Delta t, \lambda_1)}{l \cdot DPF(\lambda_1)} - \varepsilon_{HbR}(\lambda_1) \cdot \frac{\Delta OD(\Delta t, \lambda_2)}{l \cdot DPF(\lambda_2)}}{\varepsilon_{HbO}(\lambda_1) \cdot \varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbO}(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_1)} \tag{A.7}$$

Since equation A.1 is symmetric for $\Delta[HbO]$ and $\Delta[HbR]$, substituting $\Delta[HbR]$ for $\Delta[HbO]$ and $\Delta[HbO]$ for $\Delta[HbR]$ yields the following expression for $\Delta[HbR]$:

$$\Delta[HbR] = \frac{\varepsilon_{HbO}(\lambda_1) \cdot \frac{\Delta OD(\Delta t,\lambda_2)}{l \cdot DPF(\lambda_2)} - \varepsilon_{HbO}(\lambda_2) \cdot \frac{\Delta OD(\Delta t,\lambda_1)}{l \cdot DPF(\lambda_1)}}{\varepsilon_{HbO}(\lambda_1) \cdot \varepsilon_{HbR}(\lambda_2) - \varepsilon_{HbO}(\lambda_2) \cdot \varepsilon_{HbR}(\lambda_1)} \tag{A.8}$$

Appendix B

Mental States Measured with fNIRS

Table B.1: Overview of the mental states that can be measured with fNIRS. For every mental state, the characteristics of the measured signal and the brain areas with significant activations are given.

Mental state	Signal characteristics	Brain areas	Reference
Mental workload	Increased Δ[HbO]	PFC	[55, 56, 61, 63]
		dlPFC	[54, 62]
		Left dlPFC	[58]
		Left PFC	[53, 59]
		Left anterior PFC	[64]
	Decreased Δ[HbR]	PFC	[60, 63, 65]
		Right PFC	[57]
Mental stress	Increased Δ[HbO]	Right PFC	[67, 69]
		Right dlPFC	[70, 72]
		Left vlPFC	[72]
		vlPFC	[73]
		Sensory association cortex	[72]
	Decreased Δ[HbO]	Right PFC	[66, 68]
Fear response	Increased Δ[HbO]	PFC	[76, 79]
		Left PFC	[80]
		dlPFC	[81]
		Anterior PFC	[81]
		Left dlPFC	[74]
		Left vlPFC	[74]
		Right supramarginal gyrus	[77]
	Decreased Δ[HbR]	Supramarginal gyrus	[78]
		Right superior temporal gyrus	[78]
		Right dlPFC	[75]
		PFC	[76]
	Decreased Δ[HbO]	dlPFC	[75]
		Anterior PFC	[75]
Affective response	Increased Δ[HbO]	PFC	[83, 84]
		vlPFC	[82]
	Decreased Δ[HbR]	PFC	[84]
		vlPFC	[82]
	Decreased Δ[HbO]	Left dlPFC	[85, 86]
Attention	Increased Δ[HbO]	Right PFC	[87, 91]
		dlPFC	[88]

	Decreased Δ[HbO]	Right PFC	[92]
Attention loss	Late Δ[HbO] signal peak	PFC	[89]
Deception	Increased Δ[HbO]	Left PFC	[93, 94]
		Left SFG	[95]
		Right anterior PFC	[93]
		Right SFG	[94, 96]
		dlPFC	[97]
Preference	Increased Δ[HbO]	OFC	[98, 99]
		Anterior PFC	[100]
		Right PFC	[52]
	Decreased Δ[HbR]	OFC	[99]
	Increased Δ[HbR]	OFC	[98]
Anticipation	Increased Δ[HbO]	PFC	[103]
		dlPFC	[101]
		Left dlPFC	[102]
Suspicion	Increased Δ[HbO]	ACC	[105]
		OFC	[104]
		TPJ	[104]
Frustration	Increased Δ[HbO]	dlPFC	[105]
		vlPFC	[107]
		Occipitotemporal area	[106, 107]

Appendix C

Experiment Questionnaires

Table C.1: AQ items, 0 = not at all anxious, 6 = extremely anxious.

Situations	0	1	2	3	4	5	6
Diving off the low board at a swimming pool.	O	O	O	O	O	O	O
Stepping over rocks crossing a stream.	O	O	O	O	O	O	O
Looking down a circular stairway from several flights up.	O	O	O	O	O	O	O
Standing on a ladder leaning against a house, second story.	O	O	O	O	O	O	O
Sitting in front of an upper balcony of a theater.	O	O	O	O	O	O	O
Riding a ferris wheel.	O	O	O	O	O	O	O
Walking up a steep incline in country hiking.	O	O	O	O	O	O	O
Airplane trip to San Francisco.	O	O	O	O	O	O	O
Standing next to an open window on the third floor.	O	O	O	O	O	O	O
Walking on a footbridge over a highway.	O	O	O	O	O	O	O
Driving over a large bridge (Golden Gate).	O	O	O	O	O	O	O
Being away from the window in an office on the 15th floor of a building.	O	O	O	O	O	O	O
Seeing window washers 10 flights up on a scaffold.	O	O	O	O	O	O	O
Walking over a sidewalk grating.	O	O	O	O	O	O	O
Standing on the edge of a subway platform.	O	O	O	O	O	O	O
Climbing a fire escape to the 3rd floor landing.	O	O	O	O	O	O	O
On the roof of a 10 story apartment building.	O	O	O	O	O	O	O
Riding an elevator to the 50th floor.	O	O	O	O	O	O	O
Standing on a chair to get something off a shelf.	O	O	O	O	O	O	O
Walking up the gangplank of an ocean liner.	O	O	O	O	O	O	O

Table C.2: SUDS items.

Feeling of distress	Score
No distress; totally relaxed.	0
Alert and awake; concentrating well.	10
Minimal anxiety/distress.	20
Mild anxiety/distress; no interference with functioning.	30
Mild-to-moderate anxiety/distress.	40
Moderate anxiety/distress; uncomfortable, but can continue to function.	50
Moderate-to-strong anxiety/distress.	60
Quite anxious/distressed; interfering with functioning. Physiological signs	70

may be present.	
Very anxious/distressed; can't concentrate. Physiological signs present.	80
Extreme anxiety/distress.	90
Highest anxiety/distress that you have every felt.	100

Table C.3: IPQ items.

Statements/questions	0	1	2	3	4	5	6
	Not at all						Very much
In the computer generated world I had a sense of "being there".	O	O	O	O	O	O	O
	Fully disagree						Fully agree
Somehow I felt that the virtual world surrounded me.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I felt like I was just perceiving pictures.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I did not feel present in the virtual space.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I had a sense of acting in the virtual space, rather than operating something from outside.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I felt present in the virtual space.	O	O	O	O	O	O	O
	Extremely aware						Not aware at all
How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)?	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I was not aware of my real environment.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I still paid attention to the real environment.	O	O	O	O	O	O	O
	Fully disagree						Fully agree
I was completely captivated by the virtual world.	O	O	O	O	O	O	O
	Completely real						Not real at all
How real did the virtual world seem to you?	O	O	O	O	O	O	O
	Not consistent						Very consistent
How much did your experience in the virtual environment seem consistent with your real	O	O	O	O	O	O	O

world experience?

	About as real as an imagined world						Indistinguishable from the real world
How real did the virtual world seem to you?	O	O	O	O	O	O	O
	Fully disagree						Fully agree
The virtual world seemed more realistic than the real world.	O	O	O	O	O	O	O

Appendix D

Interview Experiment

Below are the interview questions that were asked to the participants after the experiment:

1. How did you experience the virtual environments?
2. Were there moments during the virtual exposure when you felt afraid? If yes, can you describe these moments?
3. Were there moments during the virtual exposure when you felt other emotions? If yes, can you describe these emotions and moments?

Appendix E

TDDR Motion Correction

This section provides an explanation of the TDDR motion correction algorithm proposed by Fishburn et al. [153]. TDDR assumes that the fluctuations in the fNIRS signals that are not due to motion are normally distributed, that the majority of signal fluctuations are not related to motion, and that the fluctuations in the signal due to motion artifacts have greater magnitude than the ones that are not related to motion. The procedure of the TDDR algorithm is given below, based on the steps described in [153]:

1. Let's take a signal x, which is a function of time t, to be the raw, unfiltered input signal. Calculate the temporal derivative x, which represents the fluctuations in the input signal:

$$y_t = x_t - x_{t-1} \tag{E.1}$$

2. Initialize the observation weights as a vector consisting of ones:

$$w_t = 1 \tag{E.2}$$

3. Estimate robust observation weights:

   (a) Use the weights to estimate the weighted mean of the fluctuations in the signal:

$$\mu = \frac{1}{\sum(w)} \sum(w_t y_t) \tag{E.3}$$

   (b) Subtract the weighted mean from the signal fluctuations to obtain the absolute residuals:

$$r_t = |y_t - \mu| \tag{E.4}$$

   (c) Calculate a robust estimate of the standard deviation of the absolute residuals:

$$\sigma = \text{median}(r) \cdot 1.4826 \tag{E.5}$$

   Where 1.4826 is the scaling factor of the median absolute deviation of the normal distribution.

   (d) Use σ to scale the deviations of each observation, using the tuning constant 4.685:

$$d_t = \frac{r_t}{4.685\sigma} \tag{E.6}$$

   (e) Use Tukey's biweight function to calculate new weights:

$$w_t = \begin{cases} (1 - d_t^2)^2, & d_t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{E.7}$$

   Fluctuations that are far from the mean get lower weights assigned. In case of an extreme fluctuation, the weight is set to 0.

(f) Iterate until μ converges.

4. Center the fluctuations in the signal by subtracting their weighted mean, and scale with the robust weights to obtain the corrected signal fluctuations:

$$y'_t = w_t(y_t - \mu_t) \tag{E.8}$$

5. Compute the corrected signal by integrating the corrected signal fluctuations:

$$x'_t = \sum_{t=1}^{N} y'_t \tag{E.9}$$

See Figure E.1 for an example of motion correction with the TDDR algorithm.

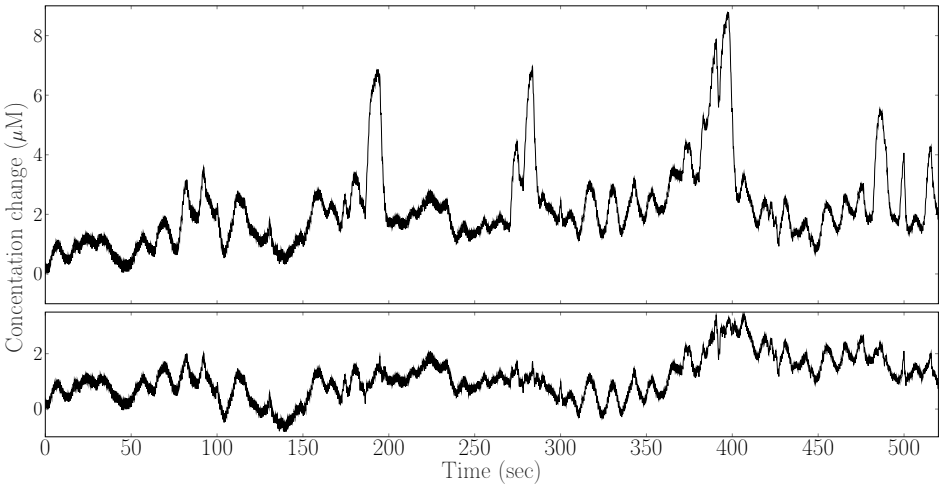


Figure E.1: Example of motion correction with the TDDR algorithm on a Δ[HbO] trace of one of the participants. The top part of this figure contains the uncorrected Δ[HbO] signal, the bottom part contains the TDDR-corrected Δ[HbO]. Large spikes and baseline shifts that exist in the uncorrected signal are mostly removed by the TDDR algorithm.

Appendix F

FDR Correction Threshold

Figure F.1 shows a bar graph containing the 27 lowest p-values (arranged in ascending order) that were generated by the 110 statistical tests that were performed. Recall that a p-value survives the FDR correction if the following holds:

$$p_i \leq \frac{i}{T} \cdot q \tag{F.1}$$

Where i is the index of the p-value under examination when all p-values are arranged in ascending order, T is the total amount of hypothesis tests that were performed, and q is the FDR-rate, which equals 0.05 in this research. In Figure F.1 the FDR correction threshold calculated based on equation F.1 is visualized by the red line. It can be seen that the 15 smallest p-values survive the FDR correction at $q = 0.05$, as their values are still below the FDR correction threshold.

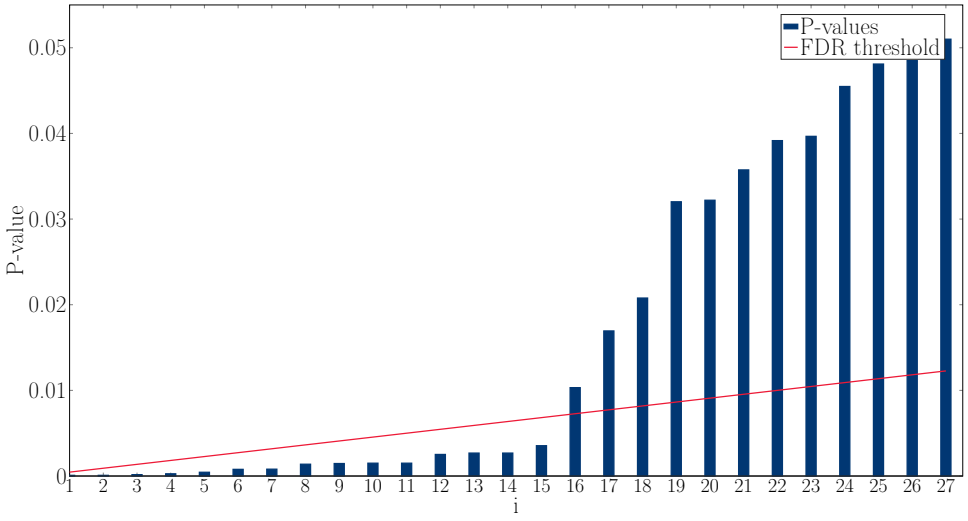


Figure F.1: Graph containing the 27 smallest p-values generated by the 110 statistical tests performed in this research and the FDR correction threshold. P-values below the threshold survive the FDR correction.

Appendix G

Classifier Hyperparameters

Table G.1: The hyperparameters of the LDA.

Parameter	Value	Description
Discriminant type	Linear	Specifies the choice between linear and quadratic discriminant analysis
Gamma	0	Regularization parameter for the covariance matrix
Delta	0	Linear coefficient threshold, to reduce dimensionality of the data space

Table G.2: The hyperparameters of the SVM.

Parameter	Value	Description
Box constraint	1	Parameter C in equation 2.19, to regulate the trade-off between the width of the margin and the number of datapoints that are misclassified during training
Kernel scale	1	Scaling of the data features
Standardize data	Yes	Standardization of the data
Kernel function	Linear	Specifies the choice between different types of kernel functions: linear, quadratic, cubic, gaussian

Appendix H

Confusion Matrices

H.1 Subject-Dependent Classifiers

Table H.1: Confusion matrix of the subject-dependent LDA over 1-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	251	85	74.70
	Fear	107	217	66.98
	Precision	70.11	71.85	70.91

Table H.2: Confusion matrix of the subject-dependent SVM over 1-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	258	78	76.79
	Fear	104	220	67.90
	Precision	71.27	73.83	72.42

Table H.3: Confusion matrix of the subject-dependent LDA over 3-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	244	92	72.62
	Fear	105	219	67.59
	Precision	69.91	70.42	70.15

Table H.4: Confusion matrix of the subject-dependent SVM over 3-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	256	80	76.19
	Fear	104	220	67.90
	Precision	71.11	73.33	72.12

Table H.5: Confusion matrix of the subject-dependent LDA over 5-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	229	107	68.15
	Fear	110	214	66.05
	Precision	67.55	66.67	67.12

Table H.6: Confusion matrix of the subject-dependent SVM over 5-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	248	88	73.81
	Fear	103	221	68.21
	Precision	70.66	71.52	71.06

H.2 Subject-Independent Classifiers

Table H.7: Confusion matrix of the subject-independent LDA over 1-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	665	175	79.17
	Fear	238	590	71.26
	Precision	73.64	77.12	75.24

Table H.8: Confusion matrix of the subject-independent SVM over 1-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	678	162	80.71
	Fear	251	577	69.69
	Precision	72.98	78.08	75.24

Table H.9: Confusion matrix of the subject-independent LDA over 3-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	680	160	80.95
	Fear	221	607	73.31
	Precision	75.47	79.14	77.16

Table H.10: Confusion matrix of the subject-independent SVM over 3-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	696	144	82.86
	Fear	236	592	71.50
	Precision	74.68	80.43	77.22

Table H.11: Confusion matrix of the subject-independent LDA over 5-second history.

		Predicted		
		No fear	Fear	Recall
Actual	No fear	679	161	80.83
	Fear	229	599	72.34
	Precision	74.78	78.82	76.62

Table H.12: Confusion matrix of the subject-independent SVM over 5-second history.

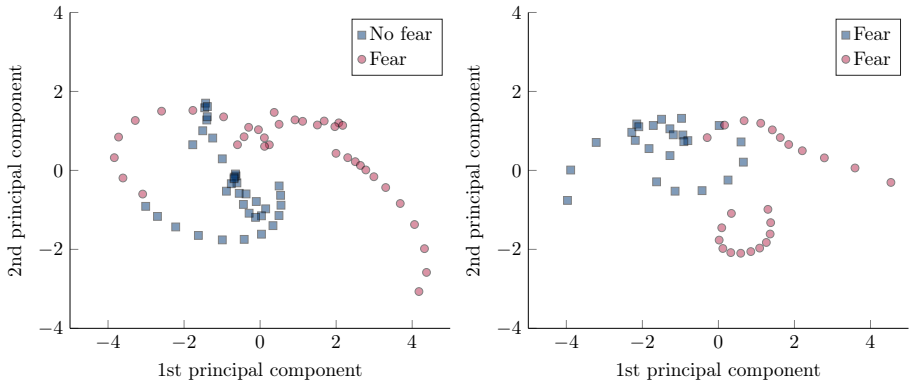
		Predicted		
		No fear	Fear	Recall
Actual	No fear	696	144	82.86
	Fear	234	594	71.74
	Precision	74.84	80.49	77.34

Appendix I

Scatter Plots Error Analysis

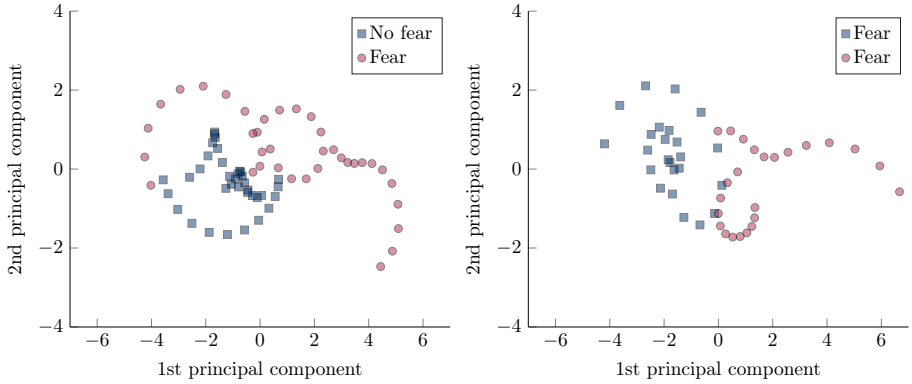
I.1 Subject-Dependent Classifiers

Scatter plots of the train and test data of the subject-dependent classifiers on the 3-second history and 5-second history data of participants 1, 2, 7, and 9 are given in Figures I.1, I.2, I.3, and I.4, respectively. PCA was used to compute the first two principal components of the train data of the 3-second history and the 5-second history separately, see section I.3. For both types of history, the train and test data are plotted against the first two principal components of the train data.



(a) 3s history train data.

(b) 3s history test data.

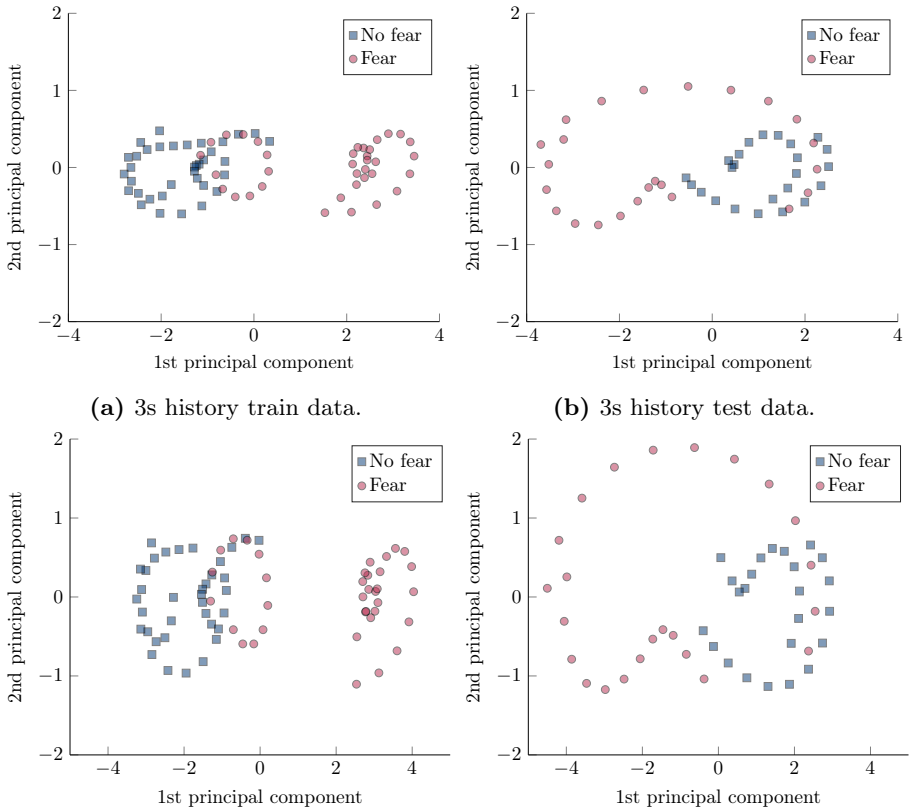


(c) 5s history train data.

(d) 5s history test data.

Figure I.1: Train and test data of the subject-dependent classifiers on 3-second history and 5-second history of participant 1.

The scatter plots given in Figure I.1 show that the test data of participant 1 is almost perfectly linearly separable along the first principal component for both the 3-second history and the 5-second history. However, the corresponding train data shows a pattern that is slightly different, where the data can be separated more accurately along the second principal component instead. Applying the separation along the second principal component to the test data yields accuracies around the level of chance.



(a) 3s history train data.

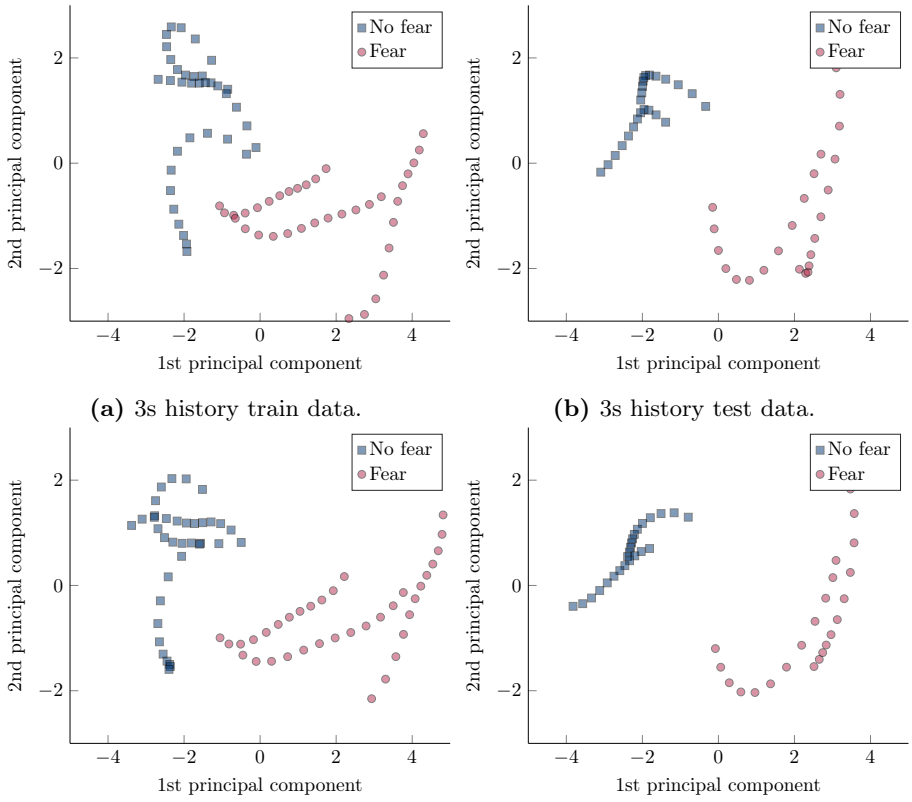
(b) 3s history test data.

(c) 5s history train data.

(d) 5s history test data.

Figure I.2: Train and test data of the subject-dependent classifiers on 3-second history and 5-second history of participant 2.

From Figure I.2 it can clearly be seen that the train data of participant 2 shows a pattern that is very contradictory to that of its test data. Therefore, the trained classifiers will not generalize well on the available test data of this participant, leading to very poor accuracies.



(a) 3s history train data.

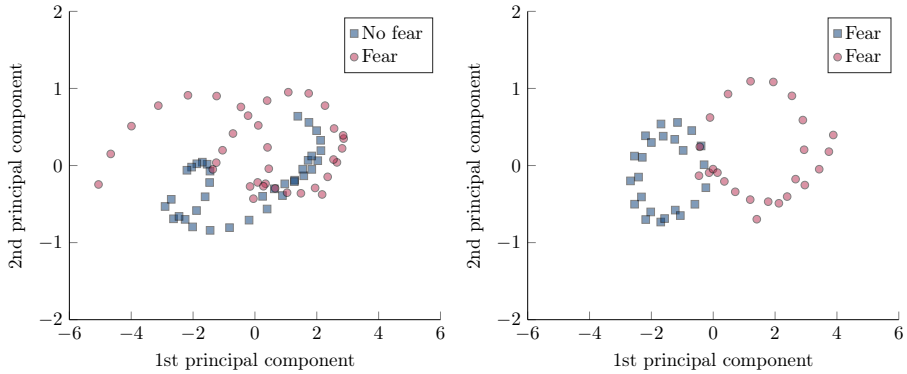
(b) 3s history test data.

(c) 5s history train data.

(d) 5s history test data.

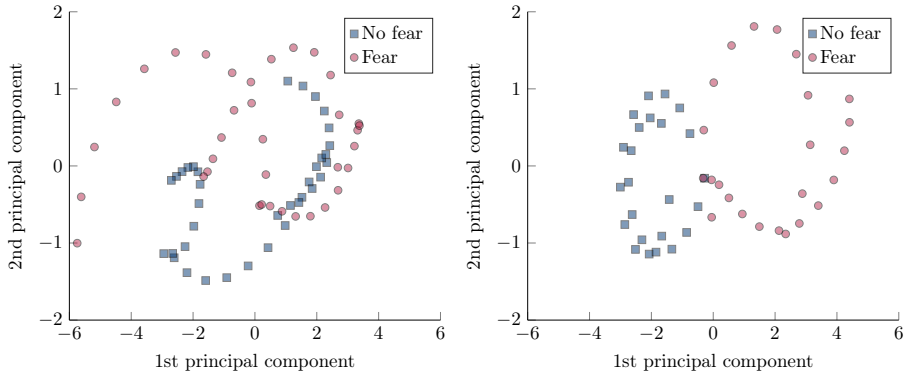
Figure I.3: Train and test data of the subject-dependent classifiers on 3-second history and 5-second history of participant 7.

The scatter plots given in Figure I.3 show that the train and test data of participant 7 follow very similar patterns. Furthermore, a clear distinction is visible between the test data labeled as "Fear" and the test data labeled as "No fear", which makes it easy to separate the data using a linear classifier. This explains why the subject-dependent classifiers of participant 7 reach very high accuracies.



(a) 3s history train data.

(b) 3s history test data.



(c) 5s history train data.

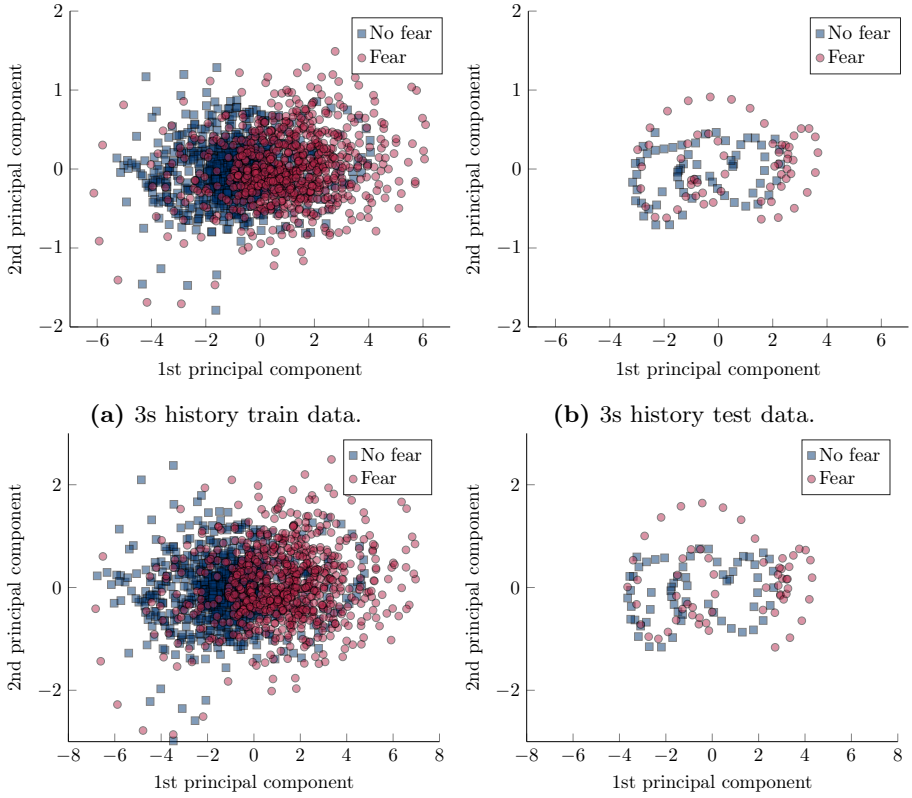
(d) 5s history test data.

Figure I.4: Train and test data of the subject-dependent classifiers on 3-second history and 5-second history of participant 9.

Finally, the scatter plots of the data of participant 9 in Figure I.4 show a pattern that is similar to that of participant 1. The test data is linearly separable along the first principal component. However, this is not the case for the train data, which can better be separated along the second principal component. Therefore, the models trained on this participant's train data are unlikely to generalize well to the test data.

I.2 Subject-Independent Classifiers

Scatter plots of the train and test data of the subject-independent classifiers on the 3-second history and 5-second history data of participants 2 and 10 are given in Figures I.5 and I.6, respectively. Again, PCA was used to compute the first two principal component of the train data of the different types of history, see section I.3, which serve as the axes along which both the train and test data are plotted.



(a) 3s history train data.

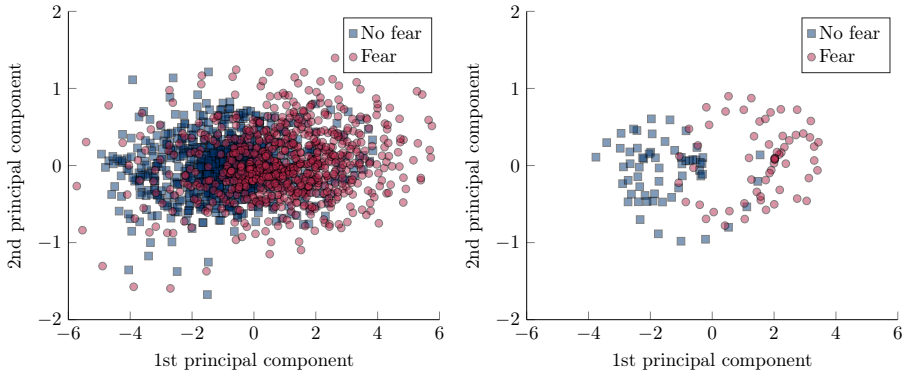
(b) 3s history test data.

(c) 5s history train data.

(d) 5s history test data.

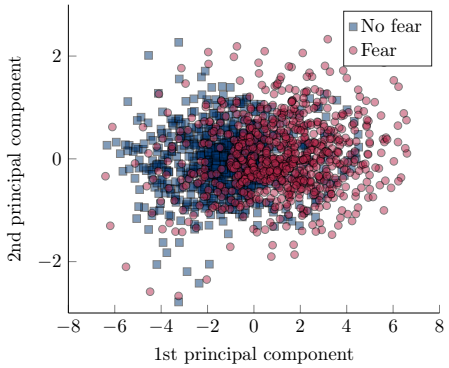
Figure I.5: Train and test data of the subject-independent classifiers on 3-second history and 5-second history of participant 2.

Figure I.5 shows that the train data of participant 2 consists of mostly negative values along the first principal component for the "No fear" data, whereas the data of the "Fear" class has mostly positive values along the first principal component. However, this pattern is not reflected in the test data of participant 2. In fact, the test data of participant 2 is not linearly separable at all. This explains why the classifiers trained on the train data of participant 2 do not perform well on the test data.

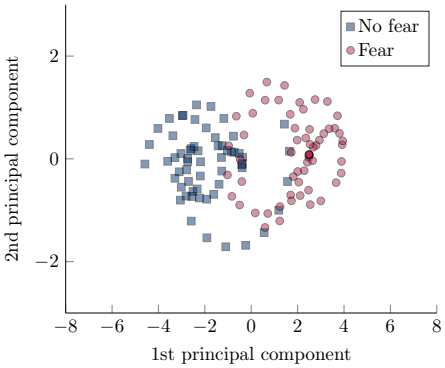


(a) 3s history train data.

(b) 3s history test data.



(c) 5s history train data.



(d) 5s history test data.

Figure I.6: Train and test data of the subject-independent classifiers on 3-second history and 5-second history of participant 10.

The train data of participant 10 also shows that the data labeled as "No fear" consists of mostly negative values along the first principal component, whereas the data labeled as "Fear" consists of mostly positive values along the first principal component. See Figure I.6. Similar patterns can be observed from this participant's test data. The test data of participant 10 is not perfectly linearly separable. However, the similarities in the patterns of the train and test data of this participant explain why the classifiers perform at rather high accuracies in this case.

I.3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality-reduction method that aims to find a low-dimensional representation of a given dataset that contains as much of the variation of the original dataset as possible [134]. It does so by finding an orthogonal projection of the original data onto a lower-dimensional subspace where the variance in the projected data is maximized [136, 137]. Let's consider a dataset $\{\mathbf{x}_n\}$ where $n = 1, ..., N$ and every datapoints has f features (dimensions). PCA can be used to reduce the dimensionality of the data from f to k, where $f \leq k$. The step-to-step approach to this dimensionality reduction using PCA is as follows:

1. PCA assumes that the data of every dimension has zero mean. Therefore, the data has to be standardized first.

2. Calculate the covariance matrix of the standardized data:

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T \tag{I.1}$$

Where $\overline{\mathbf{x}}$ represents the sample mean of the original dataset.

3. Calculate the eigenvalues and eigenvectors of the covariance matrix Σ using the following relation:

$$\Sigma \mathbf{v} = \lambda \mathbf{v} \tag{I.2}$$

Where λ is the eigenvalue of the eigenvector $\mathbf{v}$ associated with Σ.

4. The variance in the data is largest when it is projected onto the eigenvector with the largest eigenvalue. Therefore, project the original data onto the k-dimensional subspace, using the k eigenvectors with the largest eigenvalues:

$$\mathbf{z}_n = \begin{bmatrix} \mathbf{v}_1^T \mathbf{x}_n \\ \mathbf{v}_2^T \mathbf{x}_n \\ \dots \\ \mathbf{v}_k^T \mathbf{x}_n \end{bmatrix} \tag{I.3}$$

Where $\mathbf{z}_n$ denotes the projected data.

Appendix J

Pre-Experiment and Post-Experiment AQ Scores

Figures J.1 and J.2 give the pre-experiment and post-experiment AQ scores of the control group and the experimental group, respectively. The figures show that there are some inconsistencies in the pre-experiment and post-experiment AQ scores for both groups. In Figure J.1 it can be seen that two participants of the control group exceed the threshold AQ score of 20 with their post-experiment AQ scores. Figure J.2 shows that four participants of the experimental group score below the threshold AQ score of 35 with their post-experiment AQ scores.

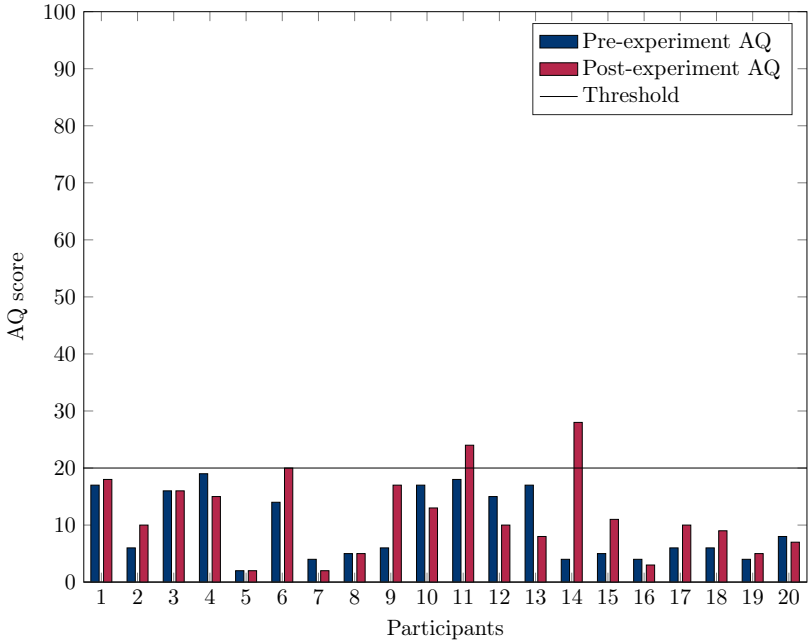


Figure J.1: Overview of the pre-experiment and post-experiment AQ scores of the control group. Participants were selected to be part of the control group if both their pre-experiment and post-experiment AQ scores were below the threshold value of 20.

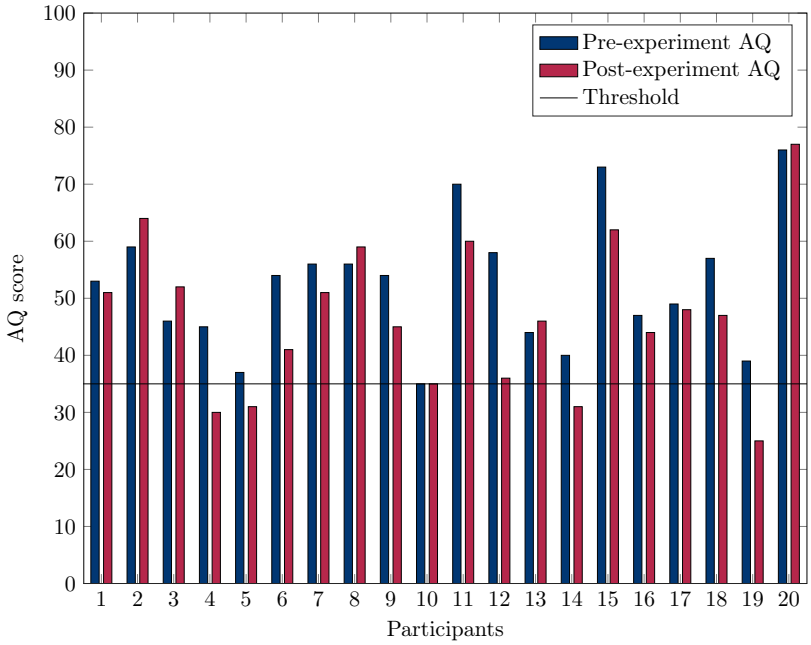


Figure J.2: Overview of the pre-experiment and post-experiment AQ scores of the experimental group. Participants were selected to be part of the experimental group if both their pre-experiment and post-experiment AQ scores were equal to or higher than the threshold value of 35.