Syracuse University

SURFACE

Dissertations - ALL

SURFACE

December 2018

Machine Learning Methods for functional Near Infrared Spectroscopy

Danushka Sandaruwan Bandara
Syracuse University

Follow this and additional works at: https://surface.syr.edu/etd

Part of the Engineering Commons

Recommended Citation

Bandara, Danushka Sandaruwan, "Machine Learning Methods for functional Near Infrared Spectroscopy" (2018). *Dissertations - ALL*. 953.
https://surface.syr.edu/etd/953

This Dissertation is brought to you for free and open access by the SURFACE at SURFACE. It has been accepted for inclusion in Dissertations - ALL by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

ABSTRACT

Identification of user state is of interest in a wide range of disciplines that fall under the umbrella of human machine interaction. Functional Near Infra-Red Spectroscopy (fNIRS) device is a relatively new device that enables inference of brain activity through non-invasively pulsing infra-red light into the brain. The fNIRS device is particularly useful as it has a better spatial resolution than the Electroencephalograph (EEG) device that is most commonly used in Human Computer Interaction studies under ecologically valid settings. But this key advantage of fNIRS device is underutilized in current literature in the fNIRS domain.

We propose machine learning methods that capture this spatial nature of the human brain activity using a novel preprocessing method that uses 'Region of Interest' based feature extraction. Experiments show that this method out performs the F1 score achieved previously in classifying 'low' vs 'high' valence state of a user.

We further our analysis by applying a Convolutional Neural Network (CNN) to the fNIRS data, thus preserving the spatial structure of the data and treating the data similar to a series of images to be classified. Going further, we use a combination of CNN and Long Short-Term Memory (LSTM) to capture the spatial and temporal behavior of the fNIRS data, thus treating it similar to a video classification problem. We show that this method improves upon the accuracy previously obtained by valence classification methods using EEG or fNIRS devices. Finally, we apply the above model to a problem in classifying combined task-load and performance in an across-subject, across-task scenario of a Human Machine Teaming environment in order to achieve optimal productivity of the system.

MACHINE LEARNING METHODS FOR FUNCTIONAL NEAR INFRARED SPECTROSCOPY

By

Mallika Arachchilage Danushka Sandaruwan Bandara
B.S. (Hons.), University of Moratuwa, Sri Lanka, 2009
M.S., Syracuse University, USA, 2013

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Syracuse University
December 2018

Copyright © 2018 Mallika Arachchilage Danushka Sandaruwan Bandara

All Rights Reserved

ACKNOWLEDGMENTS

The path to this dissertation has not been straightforward. It was riddled with many challenges and obstacles on the way. I am lucky to have had the support of my co advisers, Dr. Senem Velipasalar and Dr. Leanne Hirshfield who guided me patiently through those trying times. I want to thank Dr. Velipasalar for taking on this interdisciplinary project with an open mind and scientific curiosity. Also for putting in a lot of her time proof reading my manuscripts towards the night. I also want to thank Dr. Hirshfield for introducing the fNIRS device to me and encouraging me to take on this daunting task of scientific exploration. It was a pleasure to work with her as her student and I want to thank her for her continuous support for my degree progress. I am also thankful to my committee members, Dr. Pramod K. Varshney and Dr. Reza Zafarani for their invaluable feedback and constructive criticism on my research ideas and during this dissertation project. Moreover, I am delighted to have Dr. Jianshun Zhang serve as the oral examination chair for my defense and want to thank Dr. M Cenk Gursoy for being part of my defense committee. The feedback from my dissertation committee is highly valued and appreciated.

I want to acknowledge Dr. Chilukuri Mohan for having faith in my potential and giving me the opportunity to start a Ph.D. at Syracuse University. I also want to thank Dr. Jae Oh for his encouragement as my Ph.D. life comes to an end. Their belief in me made me more committed and determined.

I am thankful to all my colleagues that worked in the research groups that I was lucky to be a part of, Trevor Grant, Natalie Sommer, Sarah Bratt, Mark Costa, Nupur Kulka-

rni, Yu Zheng, Burak Kakillioglu, Yantao Lu, Maria Cornacchia, Tianlin Liu, Krishna Ravi Shankar, Abdul Serwadda and Sujit Poudel. We collaborated on many interesting problems over the years and I learnt a lot from them. Also I appreciate their feedback on my research ideas which helped guide me towards completion of my projects. I also want to mention my friends Maneesha Jayasuriya, Chulaka Gunasekara, Pivithuru Wijegunawardana, Saleh Raheem, Krishna Shivaram, Felix Strassburg, who took their valuable time to take part in some of my human subject experiments.

Special thanks goes out to Dean D'Amore and Derrick D'Amore of GreenView Energy Management Systems. It was a pleasure to work with them during my time as a student to support my Ph.D. financially. I also really enjoyed my time working with them as a team member and friend.

I also want to thank my Syracuse 'Family' for their support and motivation. Including my friends from Syracuse University Outing Club and the Sri Lankan community who kept me sane through this process. Although your numbers are too numerous to mention here, I appreciate all of you for adding color to my days in Syracuse.

I want to give my heartfelt gratitude to my Parents, Thilakarathna Bandara and Mangalika Thilakarathne along with my loving sister, Thamali Thilakarathna and my brother in law Nirman Dias. The emotional support and encouragement from my family has been critical in my Ph.D. career. Finally, I want to thank my loving girlfriend, Ximeng Chen for always being there, pushing me to reach higher in my Ph.D. and my career. Thank you for putting up with me and cheering me on!

TABLE OF CONTENTS

List of Figures **ix**

1 Introduction **1**

1.1 Objectives 2

1.2 Research Impact 3

1.3 Organization of the Dissertation 4

2 Building Predictive Models of Emotion with Functional Near-Infrared Spectroscopy **5**

2.1 Introduction 5

2.2 Background and Literature Review 7

2.2.1 Review of Brain Measurement Techniques 7

2.2.2 The Construct of Emotion 8

2.2.3 Objective Measures of Emotion 10

2.2.4 Challenges in Using Machine Learning on Cognitive Data 12

2.3 Experiment 13

2.3.1 Selection of Stimulus Material 14

2.3.2 Equipment Setup 15

2.3.3 Protocol 15

2.4 Data Analysis and Results 17

2.4.1 Survey Data Analysis and Results 17

2.4.2 Label Selection 19

2.4.3 fNIRS Data Analysis and Results 20

2.4.4 Region of Interest Analysis and Feature Generation 20
2.4.5 Correlation Tests on the fNIRS Features vs the SAM Survey Labels 22
2.4.6 Machine Learning 24
2.5 Discussion 27
2.6 Conclusion and Future Work 30

3 Classification of Affect Using Deep Learning on Brain Blood Flow Data 32
3.1 Introduction 32
3.2 Background and Related Work 34
3.2.1 Machine Learning challenges for Brain Data 34
3.2.2 Objective Measurement of Emotion 35
3.2.3 Functional Near-Infrared Spectroscopy (fNIRS) 36
3.2.4 Related Results 41
3.3 Experiment Design 42
3.3.1 Description of Dataset Collection 42
3.4 Data Analysis and Results 45
3.4.1 Preprocessing 47
3.4.2 Classification Model Development 48
3.4.3 Spatial Dependency in Input Data 50
3.4.4 Combining HbO and Hb Data 51
3.4.5 Predictiveness of fNIRS Channels 53
3.4.6 Effect of the Number of Timesteps on Model Performance 53
3.4.7 Label-wise Model Performance 55
3.5 Analysis of Results 56
3.6 Conclusion 57

4 Identification of Potential Task Shedding Events Using Brain Activity Data 58
4.1 Introduction 58

4.2 Background 60
    4.2.1 Human Machine Teaming and Task Shedding 60
    4.2.2 Using Psychophysiological Sensors to Measure Workload 61
    4.2.3 Machine Learning Classifiers on fNIRS data 62
4.3 Methods 62
    4.3.1 Experimental Protocol 62
    4.3.2 Stimulus Materials: Training Data 64
    4.3.3 Stimulus Materials: Test Data 66
4.4 Data Analysis 69
    4.4.1 fNIRS Data 69
    4.4.2 Label Calculation 70
    4.4.3 Discretization of ground truth Labels 73
    4.4.4 Classification Model 74
    4.4.5 Model Evaluation 75
4.5 Results 76
4.6 Conclusion 84

5 Conclusions and Future Work **86**
5.1 Conclusions 86
5.2 Future Work 87

References **89**

LIST OF FIGURES

2.1 fNIRS background 8
2.2 The circumplex model of affect. 9
2.3 fNIRS probe positions 15
2.4 Music video stimuli 16
2.5 The rating distribution of participants' responses 18
2.6 Post task survey label distribution 19
2.7 Brodmann regions covered by the fNIRS 21
2.8 Frequency mapping of predictive brodman regions 26

3.1 The circumplex model of affect 32
3.2 Theoretical background of fNIRS 37
3.3 Diagram of LSTM cell 40
3.4 Music video stimuli 43
3.5 Valence survey item that was administered after each video session. 43
3.6 Comparison between the survey labels and the stimulus labels in all the video sessions. 45
3.7 Frequency distribution of survey labels. 46
3.8 Frequency distribution of survey labels after merging the labels. 46
3.9 Preprocessing of fNIRS data 48
3.10 Structure of the CNN + LSTM Model 49
3.11 Detailed structure of the CNN + LSTM model 50
3.12 Channel predictivity heatmap 54

3.13 Variation of Classification Performance with Number of LSTM timesteps (T) . . . 55

3.14 Aggregated Confusion Matrix for HbO + Hb with T=10 56

4.1 The 3 x 11 fNIRS probe configuration. 63

4.2 Adaptive Words Task Presentation. 64

4.3 Visual search task presentation. 65

4.4 The go no-go task presentation. 65

4.5 N-back task presentation. 66

4.6 Triage source intrusion. 67

4.7 The Multi-Attribute Task Battery. 69

4.8 Data ranges for Task Load, Performance and Task Shedding Index. 71

4.9 Apportioning tasks to the time segments. 72

4.10 CNN+LSTM classification model. 75

4.11 Confusion matrix for Ensemble model tested on Triage task. 78

4.12 Confusion matrix for Ensemble model tested on MATB task. 78

4.13 An HMT system swapping tasks from overloaded (High TSX) users to underloaded (low TSX) users. 82

CHAPTER 1

INTRODUCTION

The human brain is one of the most sophisticated structures in the known universe. Up until the invention of precise measurement and imaging devices, the functional details of the human brain had remained a mystery. But with recent advances in measurement technology and analysis methods, we are provided with a window into the workings of the human brain. Treating the brain as a sensor, we can obtain the most objective view possible; into the state of the human. Functional Near Infrared Spectroscopy (fNIRS) is one such non-invasive brain activity measurement device. It has been introduced in the mid 1990's and has several advantages over existing brain measurement technologies such as Electroencephalography (EEG) and Functional Magnetic Resonance Imaging (fMRI). fNIRS devices measure the blood flow in the brain through pulsing near infra-red light into the brain tissue. The reflected light intensity can be used to infer the concentration of oxygenated (HbO) and deoxygenated (Hb) blood in that particular area. The fNIRS has become popular in human computer interaction field due to the ease of setup, calibration and portability. One of the key advantages of fNIRS over the EEG devices is the high spatial resolution of its data. However, most of the existing body of literature on fNIRS ignores the spatial structure of the fNIRS dataset, treating it instead as a tabular dataset. In our research, we attempt to fill this gap in existing fNIRS research by including the spatial information in the analysis of fNIRS data. We also introduce a method to capture the temporal dynamics of the fNIRS data. Additionally, we demonstrate the use of fNIRS in diverse applications in the Human Computer Interaction domain. In the first two chapters of this work, we use an fNIRS dataset labeled with emotion labels based on [1], and show how our method improves on the state of the art brain activity classification. We look at valence, or the positivity or negativity of emotion, which has been historically difficult to classify with high accuracy using physiological data. In the third chapter, we extend this classifier model to the Human Machine Teaming domain where we predict the potential for human performance degradations due to high task load. Our method situates the fNIRS as a useful device in measuring the mental state of a human.

1.1 Objectives

The main focus of this dissertation is to capture the spatial and temporal nature of fNIRS data in novel machine learning methods as well as demonstrate several basic research areas that can benefit from the method. These basic methods such as valence classification and performance degradation classification can be applied in a wide range of applied settings such as,

- Wearable devices, especially head mounted systems such as Virtual Reality or Augmented Reality systems can benefit from information about the wearer's mental state so that the device can optimize the user interface to the user state.

- Assistive technologies, Users with disabilities such as muscular dys-trophy have difficulties expressing emotion using facial expressions. An emotion detection system can use their brain activity to communicate their emotion to other humans that interact with them.

- Robotic Systems, As robots become more ubiquitous, the need for them to understand and respond to human state changes becomes critical. Our research is a first step in creating Artificial Intelligence (AI) that is emotion aware. Which will in turn make the vision of domesticated AI a reality.

- Interactive media, As the multitude of entertainment content providers saturate the market with content such as movies, video games and TV shows, the demand for content that can adapt to the user is also growing. This can be seen in open ended video games and other content that has recently become popular. Such content can benefit from the research presented in this work where the content can adapt to user state, be it emotional state or the cognitive overload of the user.

- Human Machine Teaming, as the line between human and machine become blurred due to the conglomeration of wearables and implants that are designed to augment the user's abilities, there is a need to optimize this human machine symbiosis. The last chapter of this dissertation touches on this topic in detail.

1.2 Research Impact

The research presented in this dissertation has resulted in multiple publications.

- Bandara, D., Velipasalar, S., Bratt, S., & Hirshfield, L. (2018). Building predictive models of emotion with functional near-infrared spectroscopy. International Journal of Human-Computer Studies, 110, 75-85.

- Bandara, D., Hirshfield, L., & Velipasalar S. (Under Review) Classification of affect using deep learning on brain blood flow data. Journal of Near Infrared Spectroscopy.

- Bandara, D., Hirshfield, L., & Velipasalar S. (Under Review) Identification of Potential Task Shedding Events Using Brain Activity Data. ACM Transactions on Computer Human Interaction.

- Bandara, D., Song, S., Hirshfield, L., & Velipasalar, S. (2016, July). A more complete picture of emotion using electrocardiogram and electrodermal activity to complement cognitive data. In International Conference on Augmented Cognition (pp. 287-298). Springer, Cham.

- Bandara, D., Song, S., Hirshfield, L., & Velipasalar, S. (2016, July). A more complete picture of emotion using electrocardiogram and electrodermal activity to complement cognitive data. In International Conference on Augmented Cognition (pp. 287-298). Springer, Cham.
- Serwadda, A., Phoha, V. V., Poudel, S., Hirshfield, L. M., Bandara, D., Bratt, S. E., & Costa, M. R. (2015, September). fnirs: A new modality for brain activity-based biometric authentication. In Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on (pp. 1-7). IEEE.
- Hirshfield, L., Costa, M., Bandara, D., & Bratt, S. (2015, August). Measuring situational awareness aptitude using functional near-infrared spectroscopy. In International Conference on Augmented Cognition (pp. 244-255). Springer, Cham.
- Bandara, D., Hirshfield, L., & Velipasalar, S. (2014, June). Insights into User Personality and Learning Styles through Cross Subject fNIRS Classification. In International Conference on Augmented Cognition (pp. 181-189). Springer, Cham.

1.3 Organization of the Dissertation

The remainder of this dissertation is organized as follows, in Chapter 2 we show how preprocessing can be used to augment a traditional machine learning classifier such as Support Vector Machines for application to fNIRS data. Chapter 3 considers the use of Convolutional Neural Networks (CNN) and LSTM as a method to incorporate the spatial and Temporal nature of fNIRS data. Chapter 4 demonstrates an extension of the method to the area of performance degradation dues to high task load.

CHAPTER 2

BUILDING PREDICTIVE MODELS OF EMOTION WITH FUNCTIONAL NEAR-INFRARED SPECTROSCOPY

2.1 Introduction

Accurately assessing human emotion has long been a goal of researchers and practitioners in human-computer interaction (HCI), as emotion is essential for understanding users' experiences with new technologies and for designing affect-based adaptive systems [2][1] [3]. Emotion is a complex phenomenon often difficult to recognize for humans, never mind machines [4]. While emotions are frequently measured with self- report surveys, many HCI researchers recognize the shortcomings associated with self-report methods, such as the tendency to inaccurately assess personal emotions. Furthermore, these self-report techniques are administered after a task completion which interrupts the user experience and fails to capture real-time information about the user's changing emotional states during the task. For this reason, researchers have attempted to measure and predict changing emotional states using a variety of objective physiological sensors such as functional magnetic resonance imaging (fMRI), Electroencephalography (EEG), Galvanic Skin

Response (GSR), and Heart Rate Variability (HRV), as detailed in the next section. While much progress has been made in objectively measuring and predicting user emotions, further interdisciplinary research is needed to develop robust models for accurately predicting real-time changes in emotional state. As biologists and neuroscientists continue to analyze the physiology of emotion, biotechnology experts are developing new non-invasive sensors that are practical, robust to noise, and highly accurate [5] [6]. Meanwhile, computer scientists continue developing machine learning and data mining models capable of making real-time predictions from this wide array of multi-modal physiological sensor data [7] [8] [9] [1]. The focus of our research involves the use of functional near-infrared spectroscopy (fNIRS), a relatively new, non-invasive brain measurement technique that is resilient to noise, portable, and allows for naturalistic participant movement (as compared to fMRI). Further, fNIRS has higher spatial resolution than EEG and enables the localization of specific brain regions of activation while taking measurements under normal working conditions [10] [11] [12] [13]. Our goal is to leverage the high spatial resolution of fNIRS to develop machine learning classifiers capable of predicting valence and arousal in participants with a high degree of accuracy. In the experiment described in this chapter, participants' brain function was measured with fNIRS while they viewed a variety of clips extracted from music videos. These videos have been shown to elicit various levels of valence and arousal. After each video clip they filled out the Self-Assessment Manikin (SAM) [14] to indicate their valence and arousal. The self-report values from the SAM were used as labels during subsequent supervised machine learning classification. This research makes two primary contributions in the realm of HCI: First, we demonstrate the capability of classifying and distinguishing between affective states on the valence and arousal dimensions using fNIRS, a practical non-invasive device. Our fNIRS results show that specific functional brain regions are recruited during changes in valence and arousal and these regions are consistent with those identified by fMRI research on emotion. Second, we develop models to classify and predict emotional states across subjects, creating the capacity to generalize the model to new participants rather than training each model per individual. The F1-scores achieved by our classifiers suggest that fNIRS is particularly useful at distinguishing between levels of valence, which has proven to be difficult to measure with physiological sensors. The remainder of this chapter proceeds as follows. We first provide related background information and a review of the relevant literature. Next, we describe our experimental set-up and protocol. We then report the analysis procedures and results and discuss findings in the context of our research goals. Last, we describe study limitations and possible avenues for future work stemming from this research.

2.2 Background and Literature Review

This section describes the fNIRS device and how it compares to other popular brain measurement techniques. We then describe conceptualizations of emotion, the measurement of emotion using subjective and objective brain measurement techniques, and describe challenges faced in conducting research using machine learning on cognitive data for emotional state predictions.

2.2.1 Review of Brain Measurement Techniques

The measurement of brain activity has significant potential for evaluating the physiological correlates of emotion. Sensor technologies such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) provide valuable insight into the functions and structures of the brain. However, they constrain subject movement and in the case of fMRI, require that subjects remain completely still. Further, they can expose subjects to hazardous materials (PET) or to loud noises (fMRI) [3] and are not ideal for assessing the neural activity of participants under normal working conditions. The use of the electroencephalograph (EEG) has attracted researchers interested in non-invasively measuring users' brain activity [15] [8]. The EEG has gained popularity for research use be- cause of its cost-effectiveness, ease of use, and granular temporal resolution. In the 1990s fNIRS was introduced, a tool which can augment and overcome some of the limitations of EEG and other brain-imaging devices [16]. The fNIRS device pulses near-infrared light in the wavelength range (690-900 nm) into the brain (Fig. 2.1).

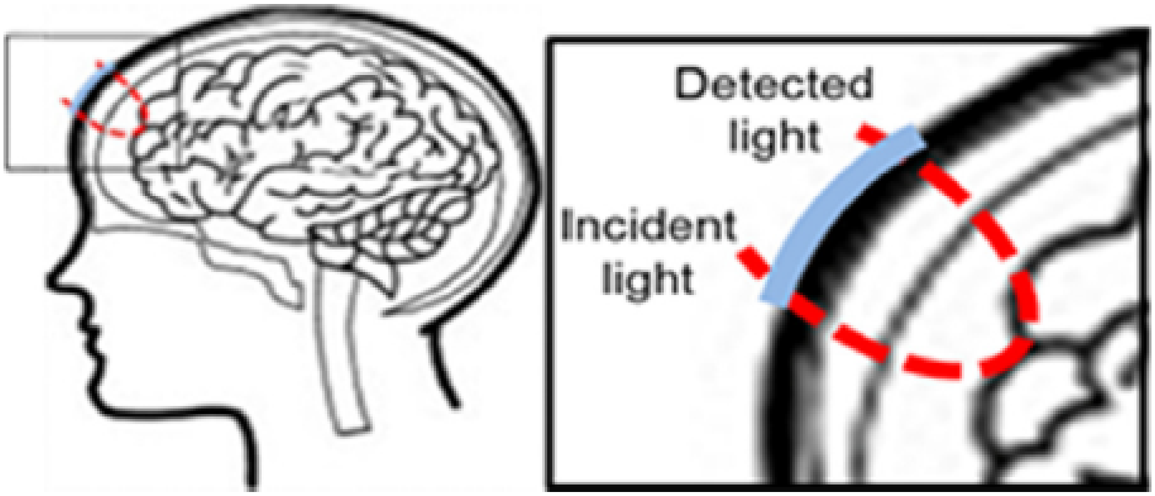


Fig. 2.1: Near-infrared light is emitted from a diode into the cortex, and detectors measure the light reflected out of the cortex.

The primary absorbers of near-infrared light are deoxygenated hemoglobin (Hb) and oxygenated hemoglobin (HbO) in tissues. During hemodynamic and metabolic processes, these light values change in association with neural activity in the brain [16]. These metabolic changes can then be detected through the measurement of the diffusively reflected light pulsed into the brain cortex [16] [3] [17]. The use of fNIRS includes measuring a range of cognitive states while computer operators engage in tasks during normal working conditions [18] [19] [20] [6]. In the next section, we describe the construct of emotion and provide a review of the literature on the subjective and objective measurement of emotion.

2.2.2 The Construct of Emotion

Research on emotion has increased significantly over the past two decades with many fields contributing; including psychology [19] [20] [6] [21], neuroscience [22] [23], medicine [24] [25], sociology [26] [27], and computer science [3] [11]. Among this recent surge, most researchers agree that emotions are affective states that exist over a relatively short period, with durations ranging from milliseconds to minutes, and are related to an event [28] [29]. A frequently used metric for quantifying emotions is by mapping them to points in a two-dimensional space of affective valence and arousal. Valence represents overall pleasantness of an emotional experience and arousal represents the intensity level of an emotion, ranging from calm to excited [2] [30] [31]. These two dimensions enable us to differentiate between four basic categories of emotions. Some models of emotion identify nine categories of emotion by including a neutral section. In principle, an infinite number of other categories can be defined [32], but Russell's [2] circumplex model of affect is a widely used and well-vetted [33] [34] [35] model in contemporary HCI research. Fig. 2.2 depicts the arousal and valence dimensions and their relation to emotional state [2]. Other conceptual models of emotion include Ekman and Friesen's [36] model based on discrete sets of universal emotions and Plutchik's [37]. The literature on modeling emotion is reviewed further by Posner et al. [38]. They provide support for the two-dimensional model of affect with examples from empirical studies.

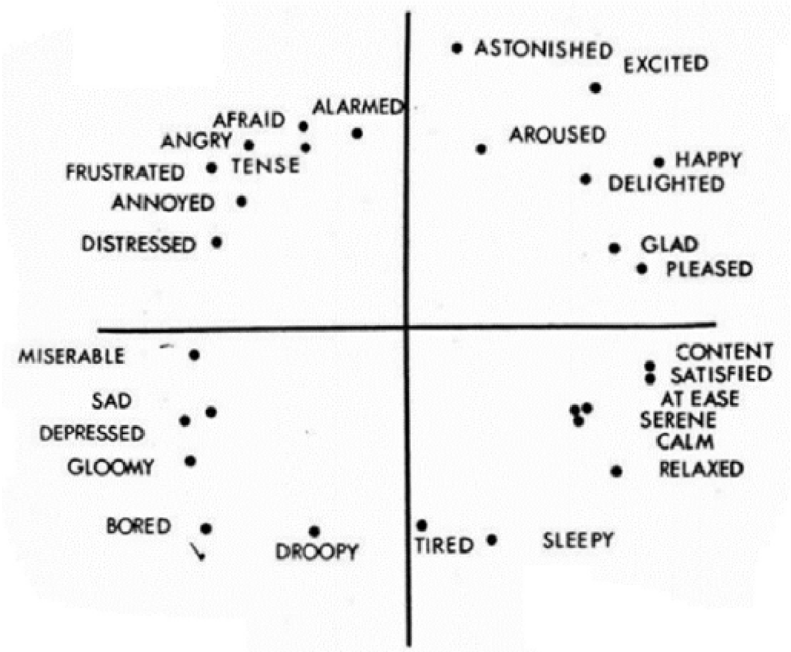


Fig. 2.2: The Circumplex model of affect. Horizontal axis shows degree of valence (pleasure/displeasure) and vertical axis shows degree of arousal [2].

Self-report surveys, such as the Self-Assessment Manikin (SAM) or the Positive and Negative Affect Scale (PANAS) survey instruments [39] are most commonly used to locate a person's perceived emotion within the circumplex model [2]. In our experiment, we use the SAM, a pictorial assessment technique for evaluating the pleasure, arousal, and dominance associated with subjects' affective reactions to stimuli.

2.2.3 Objective Measures of Emotion

To overcome the subjective limitations of self-report surveys, a range of objective sensors have been used to measure emotion. Charles Darwin formally documented variations in facial expressions associated with specific discrete emotions [40]. Schwartz and his colleagues built on the pioneering work of Darwin and recorded facial EMG of subjects engaging in pleasant and unpleasant mental imagery [41] [42] [36]. Facial Electromyography (EMG), the recording of electrical signals associated with facial muscle activity, is used extensively as a measure of emotional state [43] [44] [45]. Early research found that conscious experiences of emotion evoke specific physiological activity [46]. Another view is that distinct experiences of emotions are produced by a continuous interaction of both mind and body [47]. More recent research has shown that psychological states evoke skin conductance changes when a user is presented with emotionally charged pictures [48], computer games [49] and emotional films [50]. Extensive literature has examined the strong association of emotion with cognition and brain activity. Cacioppo et al. [51] suggested that emotion helps construct cognition and cognition helps construct emotion. Further, methods for identifying neural networks associated with different semantic emotional states in the brain were developed [52] as well as using Russell's 2-dimensional valence/arousal model [6]. With its capability to localize specific brain regions of activation, fMRI studies have measured the neural correlates of emotion in the brain. For example, Viinikainen et al. [53] showed participants affective images from the International Affective Picture System (IAPS) [30] database while in the fMRI scanner. Results showed that both arousal and valence manifested different types of responses to negative and positive stimuli in the brain and suggesting that there are different valence and arousal representations in the brain for negative and positive (unpleasant and pleasant) stimuli. They also note that the medial pre-frontal cortex (mPFC), a large region spanning the front portion of the human brain, is important in the processing of emotions. In another fMRI study, Colibazzi et al. [6] found that changes in arousal directly affected the supplementary motor cortex, and several deep brain regions that support the limbic system. Changes in valence also effected the supplementary motor cortex, as well as the dorsolateral prefrontal cortex (DLPFC), inferior parietal cortex, and the frontopolar cortex. The researchers suggest that the more unpleasant the emotion the higher the activity in the DLPFC and frontopolar cortex. Also, the researchers note that the supplementary motor cortex may constitute an interface between limbic and motor-executive systems, whereby the brain transforms affective experiences into complex motor plans. For example, a feeling of excitement that drives a desire to dance activates the motor cortex even if the movement is not actually executed. The emotion-related activation in the pre-motor cortex was complemented by recent research by Warren et al. [45], who made concurrent fMRI and EMG recordings of participants while they listened to auditory sounds designed to elicit different levels of valence and arousal. Their results showed that positive auditory-induced emotions engage the pre-motor cortex, by causing the brain to automatically prepare for responsive facial gestures to the affective stimuli. In other words, sounds that induce positive emotions engage the pre-motor cortex, as the brain is preparing to create a facial gesture, such as producing a smile. Several fMRI studies of music and emotional states also found the pre-motor cortex to be directly related to the emotional experience of music. For example, they suggested sad pieces of music contrasted with happy pieces by producing differing activations in the DLPFC, frontopolar cortex, and superior temporal gyrus. These regions have also been associated with emotional experiences, introspection, and self-referential evaluation [54]. Broca's area the central region for language processing has also been linked to the emotional experience of lyric-based music [55]. Recent work in EEG and fNIRS have also focused on measuring emotion in the brain. For example, [56] used fNIRS to measure the DLPFC region of participants' brains during their experiences of different emotional states. They found that increases in participants' subjective arousal correlates with activation in DLPFC. Rodrigo et al. [57] conducted an experiment that compared the subjects' emotional response to neutral and fearful faces using fNIRS. They found that some regions of the PFC (right medial) showed increased activity when viewing fearful faces. In another study, Balconi et al. [58] made concurrent measurements of EEG, Heartrate, and fNIRS data while participants viewed IAPS pictures and filled out the Self-Assessment Manikin for self-assessed valence and arousal ratings. Both the fNIRS and EEG results showed an increase in activation on the right side of the frontopolar region relating to negative emotions.

2.2.4 Challenges in Using Machine Learning on Cognitive Data

Several of the studies described above use single trial classification on cognitive data to predict emotional state. It is worth noting that using machine learning (ML) techniques on cognitive data (whether it be from EEG, fNIRS, fMRI, or some other measurement technique) is non-trivial in terms of the difficulty of data preparation, cleaning noise artifacts, feature generation, and algorithm selection and parameter adjustment. Although ML has the potential to help researchers maximize the use of neurophysiological sensors in a variety of domains, there are significant research challenges to using ML on cognitive data. For example, the high dimensionality of sensor data coupled with small sample sizes produces datasets that can be susceptible to model overfitting. Smaller subject populations provide less data for ML algorithms to train on, making the development of across subject model development and generalizability particularly challenging [29]. Because the brain is a highly-individualized structure, most ML on brain data trains and tests classifiers at the individual level [8] [7]. These classifiers have been found to improve dramatically as training time and model development improve. However, lengthy training sessions can be laborious. While well suited to research in areas such as Brain Computer Inter- facing (where a user spends days, and even months training his or her medical system), long training sessions may not be ideal in the HCI domain for users. One way to increase training data is to merge datasets from multiple subjects, enabling the classifier to train and test models based on data from many people and test the models on new individuals [59]. Across subject ML on cognitive data has been explored by a handful of researchers in the HCI domain, but classification accuracy tends to be lower than that achieved by training each model per individual. One reason for this finding is that each person's brain has slight differences, and the placement of sensors on each person's brain may differ spatially. Thus, one 'channel' of EEG or fNIRS data on one participant, may be quite different than the same 'channel' of data on another participant, making it difficult to generate and compare features in a meaningful way for inter-subject comparisons. Furthermore, the state of the art fNIRS-based emotion research has its own set of issues, as highlighted by [60]. They highlight the challenge of separating emotional activity from the other cognitive processes in the prefrontal cortex. Mentioning the importance of good experimental design when it comes to the study of emotion using fNIRS. They also identify the lack of sufficient experimental conditions, where some studies choose to use only positive and negative emotional conditions [61] where others include positive, negative as well as the neutral condition [62], which makes it difficult to compare results between experiments. In addition, fNIRS signals can be affected by peripheral responses such as facial muscle movements and changes in cardiovascular activity [63]. Another issue pointed out by [63]. is the possibility of the subjective emotional response and the neural responses lasting longer than the length of the stimuli. The length of stimuli needs to be decided with this consideration. They also mention how the selection of appropriate indicators of cortical activation (oxygenated blood flow vs deoxygenated blood flow) can affect the analysis of emotion. Also, the individual biological differences need to be taken into consideration when analyzing data between subjects. This effect has so far been difficult to investigate due to the small size of the fNIRS datasets available (15-60 participants).

2.3 Experiment

Our experiment goal was to induce a variety of emotional states in participants while measuring the hemodynamics of their brain with fNIRS. We aimed to demonstrate the use of the resulting fNIRS brain data to identify emotional state. Specifically, we aimed to develop across subject classifiers to accurately predict emotional state. For the sake of consistency, throughout the rest of the chapter we will consider emotion to be an affective mental state as perceived by the person, as elaborated upon in Section 2, thus quantifiable with self-report surveys. We use the SAM measure of emotion as our ‘ground truth’ measure of emotion. Twenty healthy, college age participants from a university in the Northeast took part in the experiment (13 male, 7 female). Upon arrival to the lab, participants provided informed consent and completed a pre-questionnaire to obtain their demographic data. They were then provided with instructions explaining the experiment and how to fill out the post task surveys.

2.3.1 Selection of Stimulus Material

The widely-used databases for emotion elicitation are International Affective Picture System (IAPS) and International Digitized Sound System (IADS) [64]. In this study, we chose music video clips from the Dataset for Emotions Analysis using Physiological signals (DEAP) dataset [1] because prior studies have found that visual-audio stimulus gives a better result than using either visual stimulus or audio stimulus [65]. A subset of music videos from the DEAP dataset were selected as stimuli to elicit participants’ emotions. The DEAP dataset experimenters preselected 120 videos using the emotion related tags from last.fm, and using a manual selection method. With this, their goal was to make sure to choose videos that fit in the four quadrants of the circumplex model (Fig. 2.2). Then they used a web-based subjective assessment experiment with 14 volunteers to further rate the music videos on valence and arousal scales. The resulting processed scores (mean/standard deviation) for valence and arousal were used as coordinates to place the music videos on the circumplex model. Then the final 40 videos were chosen that constituted regions in the circumplex model representing five experimental conditions of High Valence Low Arousal (HVLA), High Valence High Arousal (HVHA), Low Valence High Arousal (LVHA), Low Valence Low Arousal (LVLA) and Neutral Valence Neutral Arousal (N). We selected fifteen of these videos; three videos to represent each of the above five conditions. Videos were intentionally selected to maximize the expected emotional reaction of participants; that is, the HVLA, HVHA, LVHA, LVLA videos were handpicked from the results reported by Koelstra et al. that were as far away from circumplex model’s ‘neutral’ center as possible. The purpose of this selection was to ensure that each participant’s brain state was maximally representative of the quad- rants in the valence/arousal space.

2.3.2 Equipment Setup

The experiment was performed in a controlled laboratory environment. The fNIRS signals from participants' brains were recorded using a Hitachi ETG-4000 fNIRS device with a sampling rate of 10 Hz. The device provides 52 channels of brain activity data from the frontal region of the participant's brain. [66] Each participant was seated on the experiment chair and the chair was adjusted to his or her comfort level. The fNIRS probe (Fig. 2.3) was a 3×11 probe with 17 light sources and 16 detectors, resulting in 52 locations measured on the head. The distance between all light source and detector on the ETG-4000 is 3 cm, resulting in a measurement depth into the average adult brain of 2-3 cm [67]. Once the fNIRS probe was in place, a 3d digitizer was used to record the locations of each fNIRS channel on that subject's head.

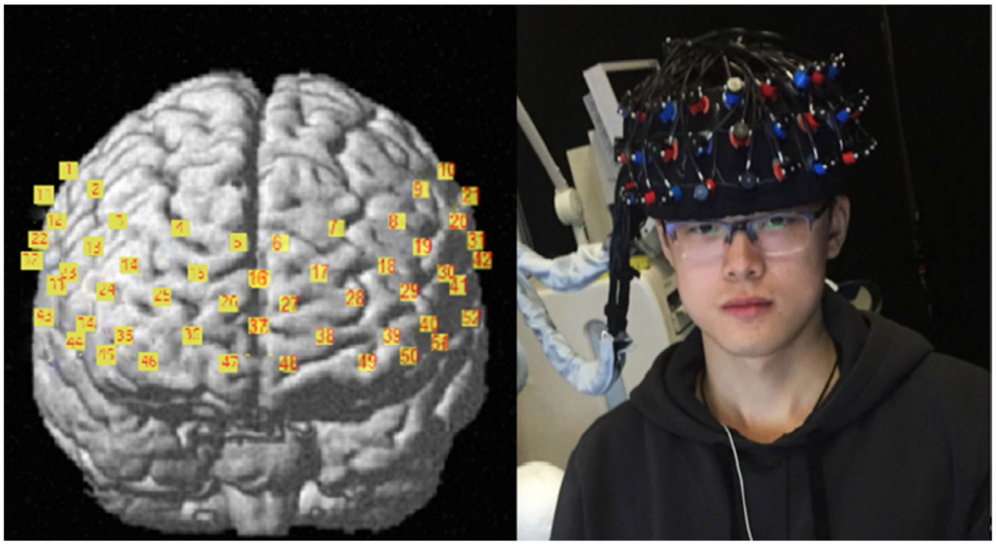


Fig. 2.3: (Left) fNIRS probe positions mapped onto brain. (right) A participant wearing the fNIRS sensors.

2.3.3 Protocol

After starting the recording of physiological data, the participant viewed a series of music videos. Each video was 60 seconds long. After the video ended, participants filled out the SAM survey for self-report assessment of valence (Likert ratings of 1-5) and arousal (Likert ratings of 1-5). Since a hemodynamic response triggered by an event typically shows an increase in signal lasting 10-12 s to rise to peak and return to baseline [68], the rest period between the videos was chosen to be 15 seconds. After this 15 s rest to allow neural activity to return to baseline, participants began watching the next video. Fig. 2.4 shows a screen shot of one of the music videos and a REST screen.

Fig. 2.4: Music video stimuli that was presented to the user and the rest screen that was shown in between trials.

The protocol followed a block design format. The music videos were separated into three blocks, each containing videos from the DEAP dataset with five unique emotion labels and included the conditions of Low Valence/Low Arousal (LVLA), High Valence/High Arousal (HVHA), Low Valence High Arousal (LVHA), High Valence/Low Arousal (HVLA), and Neutral Valence Neutral Arousal. A Note that the ‘neutral’ condition was included with music videos that were found by Koelstra et al. to be neutral on both the valence and arousal dimensions. The order within blocks was selected to ensure that within each block of videos the stimuli were presented in a random manner to the participant (so that the participants would not be able to easily guess which type of video would be played next, and to avoid the possible confounding effects of having the same-emotion-inducing video in a row, which would result in an intensified emotion effect), while still ensuring that each block in the experiment contained one video from each of the five conditions above.

Music video block 1 (videos in this block were randomized in order) 1.A fine frenzy, Almost Lover: Low Valence Low Arousal (LVLA) 2.Black Eyed Peas, My Humps: High Valence Low Arousal (HVLA) 3. Blur, Song 2: High Valence High Arousal (HVHA) 4. Smashing pumpkins, 1979: Neutral (N) 5. Stigmata, in the reflection of the eyes : Low Valence High Arousal (LVHA)
Music video block 2 (videos in this block were randomized in order) 1. Sia, Breathe me: Low Valence Low Arousal (LVLA) 2. Christina Aguilera Lady Marmalade: High Valence High Arousal (HVHA) 3. Napalm Death, Procrastination : Low Valence High Arousal (LVHA) 4. Madonna, Rain: Neutral (N) 5. Taylor Swift, Love Story: High Valence Low Arousal (HVLA)
Music video block 3 (videos in this block were randomized in order) 1. Glen Hansard, Falling Slowly: Neutral (N) 2. White Stripes, Seven nation army High Valence High Arousal (HVHA) 3.Trapped Under Ice, Believe: Low Valence High Arousal (LVHA) 4. Wilco, How to Fight Loneliness: Low Valence Low Arousal (LVLA) 5. Louis Armstrong, What a Wonderful World: High Valence Low Arousal (HVLA)

Table 2.1: Block design of experiment. Videos within each block were randomized.

2.4 Data Analysis and Results

2.4.1 Survey Data Analysis and Results

Responses to the valence and arousal items from the Self-Assessment Manikin (SAM) are made on two 5-point scales. Before beginning analyses, we looked for agreement between the SAM survey data reported by our participants and the expected results, based on the label of each video within the DEAP dataset. It is well known in the emotion literature that individual and cultural differences effect one's emotional response to a given stimulus [69] [70]. Despite these individual differences, if we look at the survey data in aggregate, we would expect that a video in the DEAP dataset with a label of high valence would, on average, result in similar ratings on the SAM when our participants watched that video. For example, the song 'What a Wonderful World' by Louis Armstrong (Table 2.1) was labeled at high valence low arousal in the DEAP dataset, and we would expect most participants to feel these pleasant and serene emotions while viewing the video. However, an individual who doesn't like that song, or who is in a hurry to complete the experiment and collect compensation, may experience the slow-paced song in a different way than others. This would result in slightly different emotional experiences due to individual differences. We were curious to see whether the subjective responses from our participants were, on average, in agreement with the labels from the DEAP dataset. So, we took all videos with a DEAP dataset label of HVHA and computed the average of our respondents' valence and arousal self- report scores reported after they saw that video. We did the same for the rest of the experimental conditions. Average results are shown in Table 2.2, with Fig. 2.5 depicting a more detailed view comparing our participants' survey responses on valence arousal and their agreement with the labels from the DEAP dataset.

DEAP Grouping	Average Valence from Survey	Average Arousal from survey
HVHA	3.38	2.95
HVLA	3.90	2.83
LVHA	2.35	3.15
LVLA	2.95	2.33
N	1.38	1.38

Table 2.2: Comparison of DEAP labels to self-report surveys from experiment

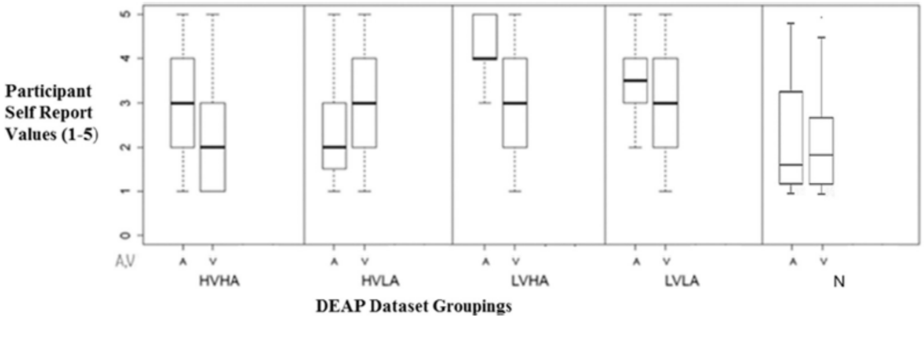


Fig. 2.5: The rating distribution of participants' responses using the labels provided by DEAP dataset (A = Arousal, V = Valence).

Our survey results showed a moderate amount of agreement between the DEAP dataset labels and our participants perceived emotional responses. Notice the wide range of responses shown in the rating distribution (Fig. 2.5), showing there were varied responses to each video. In fact, the subjects in this experiment did not report complete agreement; neither amongst themselves, nor with the DEAP dataset. Our participants disagreed with DEAP's labels in the HVHA condition participants on average said they felt 2.95 (just under 'neutral') for arousal, instead of above 3. Second, Neutral videos were reported to elicit Low Valence and Low Arousal (LVLA) instead of an average of '3' for a neutral response. In line with prior research on individual differences in emotional experiences [69] [70], this disagreement illustrates the fact that different individuals can have different, highly individualized, emotional reactions to stimuli, and it is essential to gauge each participant's self-reported reaction to stimuli, rather than assuming the stimuli will affect all individuals in the same way.

2.4.2 Label Selection

The self-report valence scale was a five-item scale (1-5), and each participant's response was rounded to the closest integer and collapsed to Low (Likert scores of 1-2), High (Likert scores of 4-5) and Neutral (Likert scores of 3) Valence. The self-report arousal scale was also a five- item scale, and the same process was used to collapse each participant's response into Low, High, and Neutral Arousal. Fig. 2.6 shows the distribution of the resulting valence and arousal labels for all participants.

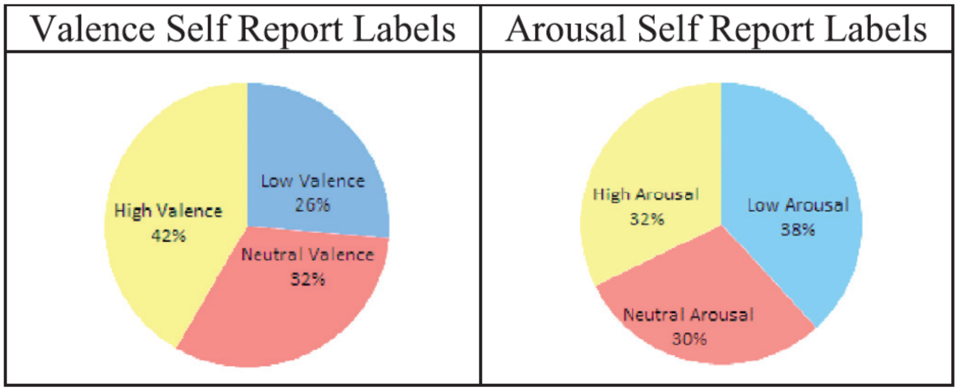


Fig. 2.6: Post task survey label distribution among the subjects. The numbers represent the percentage of trials, across participants, that were labeled as the respective arousal or valence with Likert scores of 1, 2 = low, 3 = neutral, 4, 5 = high.

These collapsed labels were used as the labels for subsequent super- vised machine learning.

It is apparent from Fig. 2.6 that there is an uneven distribution between class labels, which will cause unbalanced datasets for machine learning. This is a common issue when conducting machine learning on participants' self-report data, as individuals are likely to experience different emotional reactions to various stimuli, as reflected in their self-reports. We account for this imbalance by reporting F1-scores, because F1-scores are commonly used in lieu of overall accuracy in the presence of small unbalanced datasets. The F1 score is denoted by the following equation:

$$F1 - Score = 2 * (Precision * Recall)/(Precision + Recall)$$

$$Precision = No.ofTruePositives/(No.ofTruePositives + No.OffalsePositives)$$

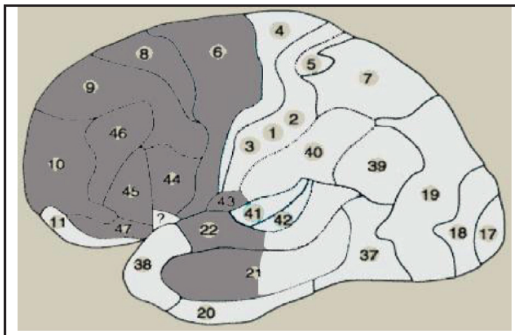
$$Recall = No.ofTruePositives/(No.ofTruePositives + No.ofFalsePositives)$$

2.4.3 fNIRS Data Analysis and Results

Data from three participants were removed from the analysis due to large motion artifacts throughout their datasets, with many channels reporting a value of 4.999, the default value used by the Hitachi- ETG when the source-detector channel has been oversaturated with light [66]. Machine Learning was carried out on the remaining 17-subject dataset. We preprocessed each participant's raw light intensity data by first down sampling our data from 10 Hz to 2 Hz. Next, we used a band pass filter to remove noise from our data, saving the frequencies between 0.5 and 0.01 Hz. We then used the modified Beer-Lambert law to convert the resulting light intensity data into relative changes of oxy- and deoxy-hemoglobin. The data was then normalized in each channel using Z-score normalization.

2.4.4 Region of Interest Analysis and Feature Generation

Our preprocessed data from above included 52 channels of data, where each channel contained the rate of change in oxy- and deoxy- hemoglobin as measured at that location over time. We converted our 3d-digitizer data (which measured the positions of fNIRS optodes on the scalp in real-space) into MNI coordinates on the brain. Next, channels were averaged together into Regions of Interest (ROI) per Brodmann areas. The Brodmann areas covered by and accessible to the fNIRS channels are depicted in Fig. 2.7 .



Brodmann Region	**Anatomical Regions Covered**
6	Pre-Motor & Supplementary Motor Cortex
8	Frontal eye fields
9	Dorsolateral prefrontal cortex
10	Frontopolar cortex
21	Middle Temporal gyrus
22	Superior Temporal Gyrus
43	Subcentral area
45	pars triangularis Broca's area
46	Dorsolateral prefrontal cortex
47	Inferior prefrontal gyrus

Fig. 2.7: Brodmann regions covered by the fNIRS probes (21, 22 only partially covered).

This resulted in 10 ROI's for analysis, where each ROI contained information about oxy- and deoxy-hemoglobin in that region. Next, for each ROI we computed several features of interest. These features were chosen because they have been successfully employed in prior machine learning research [13] or because they were employed by other researchers in recent fNIRS classification models [71]. The features were generated for both the oxy and deoxy-hemoglobin time series data in the 10 ROIs noted above. We also generated the features separately for the (i) first half (ii) second half, and (iii) for the total of each 60 s task:

- Full-width-at-half max [71]: The time difference between the two points where the signal is at half of its maximum value for the data from each 60-second-long video second.
- Slope : Slope calculated between the start and ending values of the signal.
- Mean : Average Signal Value.
- Max : Maximum Signal Value.
- Min : Minimum Signal Value. For each of the first half, second half, and total chunks of

time series data noted above, we also further split that data into six equal segments of time and we took average values across those segments:

- Piecewise Mean 1 : Average Signal Value of Segment 1.
- Piecewise Mean 2: Average Signal Value of Segment 2.
- Piecewise Mean 3: Average Signal Value of Segment 3.
- Piecewise Mean 4: Average Signal Value of Segment 4.
- Piecewise Mean 5: Average Signal Value of Segment 5.
- Piecewise Mean 6: Average Signal Value of Segment 6.

This resulted in: (10 ROIs $\times$ 2 types of data (oxy and deoxy)$\times$ 11 features (slope, min, max, etc.) $\times$ 3 time-segments (first, second half of task and total)] = 660 features to describe the brain activity during each 60-second-long video.

2.4.5 Correlation Tests on the fNIRS Features vs the SAM Survey Labels

We were curious to see which brain regions were most highly correlated with the survey labels. Therefore, we took all participants' Average Oxygenated and Deoxygenated blood concentration data for each 60 s session and correlated it with the valence labels from the surveys for those sessions. Then we obtained the same for the arousal labels. The results of Pearson correlations are shown in Table 2.3, where positive correlations indicate a direct relationship between the relative change in oxy or deoxy-hemoglobin, and the survey label. Likewise, negative correlations indicate an inverse relationship between the relative change in oxy or deoxy-hemoglobin and the survey label.

Feature	Valence SAM label correlation with the average blood conc. data (Oxy and Deoxy)	Arousal SAM label correlation with the average blood conc. data (Oxy and Deoxy)
Premotor cortex average oxy	-0.0337	-0.0797
Frontal eye fields average oxy	-0.0672	0.0072
DLPFC average oxy	-0.1247⋆	-0.0466
Frontopolar average oxy	-0.1014	-0.0335
Middle temporal Gyrus average oxy	-0.0748	-0.0376
Superior temporal Gyrus average oxy	-0.0485	-0.0098
Subcentral area average oxy	-0.1135	-0.0056
Broca’s area average oxy	-0.0679	-0.0260
Inferior prefrontal Gyrus average oxy	-0.0805	0.0064
Premotor cortex average Deoxy	-0.1227⋆	0.0124
Frontal eye fields average Deoxy	-0.1905⋆	-0.0159
DLPFC average Deoxy	-0.1251⋆	-0.1446⋆
Frontopolar average Deoxy	-0.0854	-0.0398
Middle temporal Gyrus average Deoxy	-0.0796	-0.0399
Superior temporal Gyrus average Deoxy	-0.1177	-0.0002
Subcentral area average Deoxy	-0.0521	-0.0066
Broca’s area average Deoxy	-0.0767	-0.0184
Inferior prefrontal Gyrus average Deoxy	-0.0241	-0.0479

Table 2.3: Pearson correlation between fNIRS data and the SAM survey labels for Valence and Arousal. A ⋆ is used to denote statistical significance.

The most significant correlations (Different from 0 with a significance level alpha = 0.05) obtained from this test are marked by ⋆ . It is notable that the DLPFC region had a high correlation with valence in both oxy and deoxy features. This is consistent with what Colibazzi et al. [6] found in their fMRI study. Also, a larger number of deoxy features have significant correlations especially when it comes to valence.

2.4.6 Machine Learning

Because some individual's datasets were imbalanced, we did not run a standard leave-one-participant-out cross validation. Instead, we grouped participants' data into four folds, while ensuring that each participants' data could never be split between train and test sets, as shown in Table 2.4 below. We ran a leave one-fold out cross validation to prevent any overfitting and biases that might affect the results when only using one person's data at a time for the test set. The grouping was decided simply by considering the order of participation in the study. After pulling out one-fold of data as the test set, the resulting training set was ranked using an information gain heuristic and the most predictive 15 features were selected for classification. A Support Vector Machine (SVM) classifier was trained on these features and then tested on the participants that were initially left out as the test set. This was carried out for each of the folds. The average F1-scores achieved for each fold are shown in Table 2.4.

	Valence			Arousal		
	L vs H	N vs H	L vs Ne	L vs H	N vs H	L vs N
P1, P2, P3, P4	0.736	0.600	0.651	0.652	0.650	0.670
P5, P6, P7, P8	0.741	0.583	0.690	0.690	0.621	0.587
P9, P10, P12, P13	0.745	0.578	0.589	0.694	0.652	0.651
P14, P16, P17, P19, P20	0.737	0.733	0.691	0.590	0.630	0.700
Average F1-score	0.739	0.623	0.655	0.660	0.638	0.652

Table 2.4: Average F1-Scores using the self-report survey labels for across subject classification for pairwise comparisons of high, neutral, and low valence, as well as high, neutral, and low arousal.

Since there was disagreement between the self-report labels and the original DEAP dataset labels as seen in Fig. 2.5, for a more complete comparison with the DEAP experiment, the same analysis was done using the DEAP labels instead of the self-report labels. The results of this analysis are shown in Table 2.5.

	Valence			Arousal		
	L vs H	N vs H	L vs Ne	L vs H	N vs H	L vs N
P1, P2, P3, P4	0.670	0.520	0.560	0.667	0.680	0.561
P5, P6, P7, P8	0.660	0.703	0.441	0.6507	0.710	0.730
P9, P10, P12, P13	0.682	0.523	0.600	0.670	0.732	0.634
P14, P16, P17, P19, P20	0.651	0.750	0.612	0.647	0.690	0.672
Average F1-score	0.666	0.624	0.553	0.659	0.703	0.649

Table 2.5: Average F1-Scores using the DEAP dataset labels for across subject classification for pairwise comparisons of high, neutral, and low valence, as well as high, neutral, and low arousal.

Composition of machine learning models: As noted previously, we created a feature vector with 660 features based on the fNIRS data acquired during each 60 s video. In this section, we describe the most predictive features and brain regions that contributed to our self-report label based SVM models' output. To demonstrate which brain regions were included in the feature selection process in our self-report valence and arousal models, the left side of Fig. 2.8 shows the frequency that each brain region was included as one of the top 15 features by the information gain heuristic employed by all the self-report valence models created during the leave-one-fold out cross validations. The same process was done for the self-report arousal models, with the right side of Fig. 8 showing the frequency that each brain region was included among the top 15 features created for each of the models created during the leave-one-fold out cross validations.

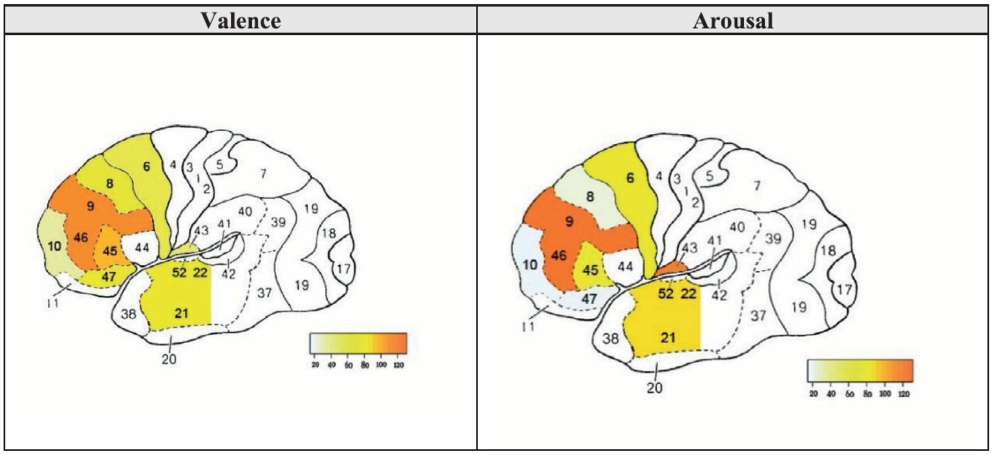


Fig. 2.8: Brain mapping of frequency counts of predictive Brodmann regions for valence and arousal classifications, across all participants

Looking beyond just the brain regions showing relevant emotional activation, we were also curious to better understand the type of features (see Section 2.4.4 for the type of features we generated) that were most predictive of the self-report valence and arousal class values across all our participants. To explore these features, we simply merged all participants' data and used the Weka 'Ranker Feature Selection' method to list the top 15 features using the information gain heuristic. We did this with all participants' data with the high and low self-report valence labels. We then repeated the process using all participants' data with their corresponding high and low self-report arousal labels. Summary data for these feature analyses is shown in Table 2.6.

For the valence comparisons, about half (8/15) of the selected features were based on oxy-hemoglobin, while the other half were based on deoxy-hemoglobin. For the arousal data, 2/3 of the top features were based on deoxy- hemoglobin. This is notable as many fNIRS studies only look at oxy- hemoglobin data, but in our analyses the deoxy-hemoglobin seems to have played an important role in the distinction of self-report valence and arousal classifications. It is not surprising to see the piecewise mean features listed often in the top 15 features, as they represented a large portion of the 660 features generated per instance (Section 2.4.4).

Valence
Deoxy Piecewise Mean 3 First Half (Frontal eye)
Deoxy Piecewise Mean 3 First Half (Inferior Prefrontal gyrus)
Oxy Piecewise Mean 4 Second Half (DLPFC)
Oxy Piecewise Mean 4 Second Half (Broca's Area)
Deoxy Piecewise Mean 4 Total (Inferior Prefrontal gyrus)
Deoxy Min First Half (Inferior Prefrontal gyrus)
Deoxy Min Second Half (Premotor)
Deoxy Piecewise Mean 4 Second Half (Frontopolar)
Deoxy Piecewise Mean 5 Second Half (Subcentral area)
Oxy Min Total (Premotor)
Oxy Piecewise Mean 4 Total (Broca's Area)
Oxy Piecewise Mean 5 First Half (Premotor)
Oxy Piecewise Mean 5 Second Half (DLPFC)
Oxy Slope Second Half (DLPFC)
Oxy Piecewise Mean 5 First Half (Frontal Eye)
Arousal
Oxy Min Total (Premotor)
Deoxy Full Width at Half Max Second Half (DLPFC)
Deoxy Max Second Half (Subcentral Area)
Oxy Piecewise Mean 3 First Half (Superior Temporal Gyrus)
Deoxy Piecewise Mean 2 Total (DLPFC)
Oxy Piecewise Mean 4 Second Half (Broca's Area)
Oxy Piecewise Mean 2 First Half (Premotor)
Deoxy Max Total (Subcentral Area)
Oxy Average First Half (Premotor)
Oxy Piecewise Mean 3 Second Half (Inferior Prefrontal Gyrus)
Deoxy Average Second Half (Frontal Eye)
Deoxy Full Width at Half Max Total (Frontal Eye)
Deoxy Piecewise Mean 3 Total (Premotor)
Deoxy Full Width at Half Max First Half (Frontal Eye)
Deoxy Piecewise Mean 3 Second Half (Middle Temporal gyrus)

Table 2.6: The most predictive features for self-report Valence and Arousal.

2.5 Discussion

As shown in Table 2.4, the self-report label based SVM model achieved average F1-scores of 0.739, 0.623, and 0.655 at distinguishing low vs high, neutral vs high, and low vs neutral valence, respectively. The arousal model achieved 0.66, 0.638, and 0.652 average F1-scores when predicting low vs high, neutral vs high, and low vs neutral arousal, respectively. Using the DEAP dataset labels

(Table 2.5) achieved F1-scores of 0.666, 0.624 and 0.553 at distinguishing low vs high, neutral vs high, and low vs neutral valence, respectively. And 0.659, 0.703 and 0.649 average F1 scores when predicting low vs high, neutral vs high, and low vs neutral arousal, respectively. When comparing the use of self-report labels vs the DEAP dataset labels, the self-report label based model provided higher F1 scores than DEAP label based model in the case of comparing Low vs High Valence, Low vs Neutral Valence, Low vs High Arousal, and Low vs Neutral Arousal. Even though there is general consensus between the F1-score results from using DEAP labels and self-report labels, the self-report label based classification performed significantly better at classifying Low vs High Valence than the DEAP label based model. This could be due to individual differences in perception of emotion as discussed in Section 2.4.1 . These results are promising, especially since the classifiers were trained across participants, highlighting the potential for creating models based on large datasets of labeled participant data for training classifiers. Considering that Koelstra et al. reported an F1-score of 0.61 at classifying between low and high valence, our self-report F1-score of 0.739 for low and high valence distinctions suggests that the fNIRS may acquire unique information relating to the measurement of valence in the brain. The fNIRS results showed that the DLPFC and Broca's region were particularly useful at distinguishing between valence levels, which is in line with prior fMRI research on valence, especially in the context of listening to music. We posit that fNIRS' high spatial resolution enables the device to measure specific brain regions (such as the DLPFC and Broca's area) which are essential in the measurement of valence. When distinguishing between high and low arousal, our self-report F1- score of 0.66 is comparable to the 0.62 F1 score achieved with the EEG in Koelstra et al.'s study. These results suggest that fNIRS could be used to complement other physiological data to measure emotion, especially when distinguishing between levels of valence. It is worth noting that our classifier was strong at distinguishing between high and low levels of valence, but F1-scores were lower when the comparisons included a 'neutral' level of valence. This makes sense as neutral valence lies between high and low valence, and it is likely harder to distinguish. Our classifiers were built from the data from seventeen participants; it would be interesting to see the accuracies from classifiers trained on a dataset with double, or even triple, the number of participants. The high density fNIRS used in this experiment (with 52 regions of the brain measured) allows for localization of brain activation. Fig. 2.8 showed the brain regions that were most predictive for the valence and arousal classification models, and these findings shed light on the neural correlates of valence and arousal. It is notable that features generated from deoxy-hemoglobin played an important role in the distinction of valence and arousal, and we urge the fNIRS research community to include deoxy-hemoglobin in their machine learning models to benefit from this information rich data. It appears that the model's prediction of valence relied heavily on Brodmann regions 9 and 46, which correspond to the dorsolateral prefrontal cortex (DLPFC). As noted in the literature review, the DLPFC region has been repeatedly linked to emotion regulation. The second most predictive region for valence was Broca's area (Brodmann region 45). Broca's is involved with processing of language, and it has been found to play a role in the processing of music lyrics, as would be the case in the music video stimuli. Perhaps participants engaged Broca's area more when they felt emotionally connected to a song and were fully engaged in lyric interpretation. The models for predicting arousal level also relied heavily on the DLFPC, which makes sense, given the region's strong link to emotional experience. When comparing the regions of the brain most predictive of valence with those regions most predictive of arousal, it is interesting to note that prediction of valence seems to engage an interconnected net- work of brain regions, whereas the prediction of arousal does not require such a large region of brain real estate. For example, valence engaged the frontopolar region (BA10) and the Frontal Eye Fields (BA8), while the arousal models did not rely as heavily on those regions. The frontopolar region is heavily involved in processing of information, and it is interesting that this region has a stronger tie to the valence dimension than the arousal dimension of emotion. The Frontal Eye Fields have been found to play a role in visual attention, and recent research from fMRI has suggested that emotionally charged visual information does influence activation in this region [72]. Perhaps the music videos that were more emotionally charged on the valence dimension also drew more visual attention from participants as they watched the videos. It's also interesting to note that both valence and arousal engaged the premotor cortex (BA6), which has been linked not only to conscious planning of movement that will be executed, but to more subconscious thoughts about movements. As suggested by [45], the emotionally charged stimuli may have engaged participants' pre-motor cortex as their brains subconsciously prepared to make a facial gesture, such as producing a smile. This also dovetails with the previously noted activation in the DLPFC, as participants would have engaged that region while regulating their reaction to the emotional content. These results are encouraging, and they rely upon localization of functional brain regions, which, before the introduction of fNIRS, could only be done with fMRI scanners. Thus, we claim that fNIRS is a strong choice for non-invasive brain measurement of emotional states, when there is a need for fMRI-quality measurements made under normal working conditions. Of course, fNIRS is limited by the fact that it has low temporal resolution and it does not measure the entire brain, making it unable to directly measure deep brain regions like the amygdala, which are heavily involved in emotional processing. We do not propose that fNIRS is the only modality for this type of measurement, but that fNIRS can be incorporated into experiments using EEG and other multi-modal sensors for measurement of emotion. It would be interesting to run an experiment with multiple sensors such as fNIRS, EEG, ECG, and GSR. Each sensor modality provides a different physiological measure, and when combined, they may provide enough pieces to the puzzle that is emotion; to get a full and accurate picture of one's emotional state. Particularly, the combination of the EEG and fNIRS [73] [74] [75] [76], can take advantage of the best aspects of each system. For instance, a combined system benefits from both the high temporal resolution of the EEG and the high spatial resolution of fNIRS. This would enable researchers to pinpoint which parts of the brain are activated by a task while also measuring quick changes in neural activity.

2.6 Conclusion and Future Work

In this chapter, we demonstrated the capability of classifying and distinguishing between affective states on the valence and arousal dimensions using fNIRS, a practical non-invasive device that can localize activation in functional brain regions with spatial resolution comparable to fMRI and superior to EEG, the most common non-invasive brain measurement modality. Our fNIRS results show that specific functional brain regions are recruited during changes in valence and arousal, and these regions are in line with prior fMRI research on emotion. We use our fNIRS data to build models to make predictions across subjects, rather than training each model individually per participant, which limits the data available for model training and testing, and can lead to overfitting. Developing accurate across-subject machine learning models is necessary to build the large datasets of data necessary for training robust classifiers that can be applied across different participants and task types. Our stimulus materials and protocol build off prior research by Koelstra et al. [1] where music videos were used as stimulus material for inducing emotional states. They found that fusing data from multiple sensors (EEG and other physiological measures) improved classification accuracy. We propose fNIRS as an additional measurement modality to further improve the predictive accuracy of emotion models. Future work should run this protocol while participants wear fNIRS, EEG, and other physiological sensors, to determine the improvement in accuracy achieved with the addition of the fNIRS modality. This line of future work would complement the growing body of literature using fNIRS in the measurement of cognitive states.

CHAPTER 3

CLASSIFICATION OF AFFECT USING DEEP LEARNING ON BRAIN BLOOD FLOW DATA

3.1 Introduction

In the psychological study of emotion, discrete feeling states such as anger, fear, delight, contentment, and surprise [21] are converted to superordinate dimensions, the most widely recognized being valence and arousal [77, 30]. Valence is also represented in Russell's circumplex model of affect as the horizontal axis (Fig. 3.1).

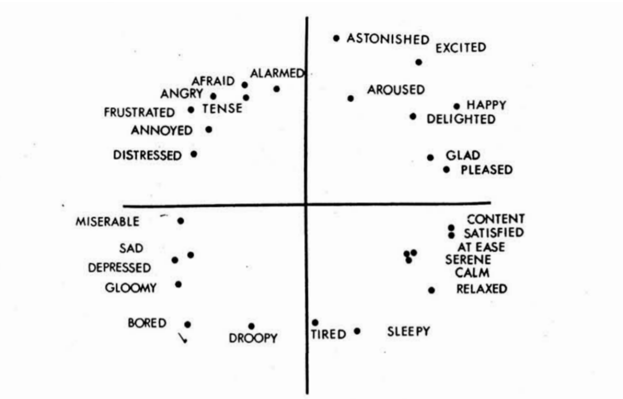


Fig. 3.1: Horizontal axis shows pleasure/displeasure and vertical axis shows degree of arousal. [2]

Accurately measuring valence has become an important aspect of human-computer interaction (HCI) research, as valence is an integral part of affect, and is useful for designing affect-based adaptive systems [1, 37, 15, 3]. Valence has typically been measured using self-report surveys such as the Self-Assessment Manikin (SAM). However, these methods have limitations, such as the inability of some subjects to accurately assess their own emotions, or intentionally censoring their responses due to an observer effect. Also, these self-report techniques are hard to implement in a real-time manner, because it would interrupt the user experience. Due to this, HCI researchers have attempted to measure affect using objective physiological sensors such as facial EMG [40, 78], functional magnetic resonance imaging (fMRI) [16, 79], Electroencephalography (EEG) [11, 80, 81, 82, 83, 84], Galvanic Skin Response (GSR) [10, 12], and heart rate variability [13, 85], to name a few.

Functional Near-Infrared Spectroscopy (fNIRS) is an emerging modality for non-invasively measuring the blood flow, specifically changes in oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (Hb), in the brain. Although attempts have been made at valence classification using fNIRS, the traditional machine learning methods, such as Support Vector Machine (SVM) and Naïve Bayes, have had limited success [56, 58, 74, 86, 87].

In this chapter, we present a new approach for analyzing brain blood flow data by incorporating Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs), to improve upon the classification accuracy. Deep belief networks and CNNs have been used to learn representations from functional Magnetic Resonance Imaging (fMRI) and Electroencephalogram (EEG) [88, 89, 90] in some previous work with moderate dataset sizes. CNNs make use of the spatial structure of image data to learn features from the image. Since fNIRS has high spatial resolution compared to EEG, it makes the CNN analysis better suited for analyzing fNIRS data. Although most fNIRS datasets are collected on a lower than ideal number of participants for proper machine learning, the ease of set-up on participants makes the device a good choice for researchers looking to build adequately sized datasets that are large enough for machine learning analysis.

Another very important aspect of fNIRS data is the time series behavior of the data. There is continuing evidence, shown by task-based fMRI research, that the temporally related brain regions have been implicated in many types of cognitive processes [91] such as emotional processing [92], reward and decision making [93], and social cognition [94]. LSTMs have been found to be very successful in modeling the long and short-term behavior of data. Therefore, we introduce the combination of CNN and LSTM to extract the spatio-temporal information in the fNIRS data.

In the first part of the analysis, we convert the Oxy Hemoglobin (HbO) and Deoxy Hemoglobin (Hb) values from the fNIRS into matrix values comparable to the pixel values in an image. This matrix is then fed into a CNN based model for classification. In the second part of the analysis, we explore the use of the combination of CNN and LSTM to further improve the classification accuracy. Our primary contributions include the following: (i) we explore the effects of removing data points that have a high likelihood of being mislabeled, which is a common challenge in research that involves labeling time series data with ground truth cognitive and emotional state labels; (ii) we investigate the effects of spatial information and dependency, and effects of reordering data channels on classification accuracy. One of the key contributions of this work is addressing the spatial dependency in the fNIRS data through CNN-based analysis. This is in contrast with the traditional machine learning methods like SVMs (which are not well suited to capture spatial information among multivariate time series data);(iii) we investigate the impact on classifier performance, and the gained accuracy when using both HbO and Hb data in the analyses, rather than just the HbO data, which is often the only biometric data source used in statistical and machine learning analyses of fNIRS data.(iv) Finally, we present the combination of CNN and LSTM to capture the spatial and temporal nature of fNIRS data and further improve overall accuracy.

3.2 Background and Related Work

3.2.1 Machine Learning challenges for Brain Data

In this section, we describe several of the challenges that have plagued research using machine learning on cognitive data. Decoding of brain states such as cognitive load and emotion are difficult for a number of reasons, including a poor signal to noise ratio in the sensor data, the dimensionality of datasets with relatively small training instances, and high variability between different trials and individuals [95]. To further complicate research on decoding of brain states, many âĂŸtraditionalâĂŹ machine learning techniques can lead to inflated accuracies, and experts in machine learning on cognitive data have begun to caution would-be reviewers and other researchers about the dangers of un-knowingly running âĂŸcannedâĂŹ machine learning algorithms on brain data, which can yield inaccurate and overly-optimistic results. See Table 1 in [96] for a full list of machine learning pitfalls to avoid with respect to brain data [96, 97]. One notable issue is that traditional machine learning models assume that all training data samples are independent and identically distributed (i.i.d.), which is not the case with brain data. Many of these models do not adequately consider the spatial and temporal dynamics at play in the multivariate timeseries data. Also, these models usually require that the brain data be represented by a feature-set, which must be defined a priori, usually using some level of domain expertise. However, with the complexities of the human brain, identifying the âĂŸcorrectâĂŹ features a priori may not be realistic. Another limiting factor is that most models are built per-participant, which results in very small datasets that dwarf in comparison to the extremely high feature spaces of the brain data, which leads to model overfitting. Since the process of collecting training data with brain measurement devices is costly and time-consuming, researchers have noted the need to build models across participants, where an adequate amount of data can be used for model training [98, 99], and it is becoming more commonplace to build and evaluate models using ‘leave-one-participant-out’ or ‘leave-one-fold-out’ cross-validation [99, 100, 101]. However, individual brain differences regarding the spatial and temporal complexities of the human brain remain a formidable challenge for this research area.

3.2.2 Objective Measurement of Emotion

Human emotion is an extremely complex phenomenon, which is in essence an interaction between the human body, made up of physiological systems, and an embodied human brain that produces all our conscious experiences [102]. The role of the physiology in this interplay has been the subject of much discussion. Cannon [103] was one of the early opponents of the idea that the peripheral nervous system played such a key role in the experience of emotion. However, research since then [51] has shown that the peripheral nervous system does play a role in the experience of emotion, while a greater role is played by the brain. Bradley and Lang [77] attempted to identify the peripheral nervous system patterns associated with discrete emotion responses with mixed results. Another important factor to consider is that depending on the environmental and individual characteristics [104], the response to a stimulus might differ. For example, a political video inducing anger to someone with an opposing political view, might cause a very different emotion in someone with aligning political views. Despite these challenges, brain measurement holds promise for assessing the neural aspects of emotion in the brain.

An additional challenge around the objective measurement and prediction of emotions is the difficulty in obtaining 'ground truth' information of emotional state [105]. This is a well-known challenge in HCI and the social sciences. When self-reports are used to gauge emotional states, the surveys are administered after a task has been completed, lacking real-time information about changing emotional state during a given time period. Furthermore, people are known to be poor at judging their own cognitive and emotional states, which can result in mislabeled data. To overcome issues relating to self-report of emotional state, another technique used to assess ground truth of emotions is to label data based on the type of stimulus presented. For example, if a particular song is labeled by experts (or by some sort of large scale study) to elicit excitement in most people, then all people that experience that song can have their data labeled based on the pre-determined emotional label. Of course, this takes a 'one-size fits all' approach that neglects individual differences or other environmental factors that could cause a person to experience a different emotional state than the pre-determined state while viewing the stimulus.

3.2.3 Functional Near-Infrared Spectroscopy (fNIRS)

The fNIRS device provides a means of measuring blood flow in the brain while being more portable than an fMRI. It provides better spatial resolution than EEG and is thus better suited for pinpointing localized brain activation in ecologically valid settings [16]. fNIRS technology uses near infrared light in two wavelengths (690-830 nm) that are pulsed into the brain (Fig. 3.2).

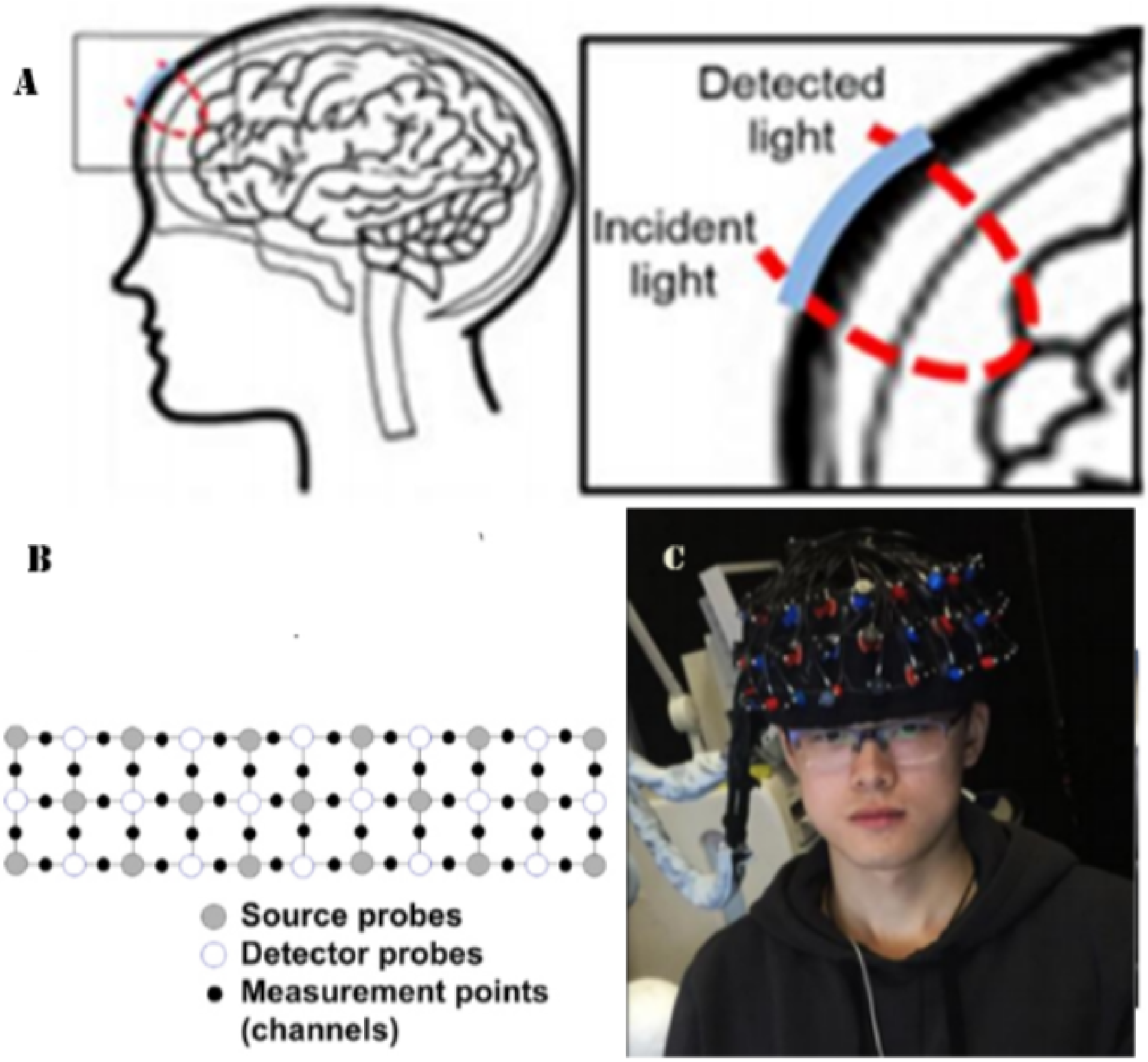


Fig. 3.2: Theoretical background and measurement settings of fNIRS. (A) A source and detector pair, which detects the Near Infrared light reflected from the cortex. This reflected light intensity varies due to the absorption by hemoglobin in the blood, and this provides a measure of blood flow in the brain, (B) Spatial configuration of the probes and channels, (C) An fNIRS sensor cap on a subject.

Hb and HbO are the main absorbers of near-infrared light in tissues during hemodynamic and metabolic changes associated with neural activity in the brain [16]. These changes can be detected by measuring the diffusively reflected light that has probed the brain cortex [16, 3, 17]. fNIRS has risen in popularity as an ecologically valid sensor that measures blood flow in the brain and thus the underlying brain activity [18, 106, 19, 20, 6].

Complexity of the fNIRS Signal

After the neurons fire in an activated brain region, which triggers a large increase in cerebral blood flow in that region, an increase in the metabolic rate of oxygen, and an increase in the volume of blood flow occur. All of these factors contribute to the blood oxygen level dependent (BOLD) signal, which can be detected (in various forms) by a number of brain measurement techniques such as fMRI and fNIRS [107]. The BOLD signal is correlated with increases and decreases in blood flow, blood oxygenation, blood deoxygenation, and blood volume [107] [108]. Research has found that HbO increases and Hb decreases in areas that are activated. Furthermore, HbO continues to increase as the load on that brain region increases until the subject becomes cognitively overloaded or switched tasks, and a good deal of fMRI and PET studies support this finding [109, 110, 111]. From a biological standpoint, research has shown that when a particular brain region is activated, there is a very fast initial decrease in HbO in that region due to the firing of the brainâĂŹs neurons. Next, surrounding regions of the brain detect this depletion in HbO and these surrounding regions provide an increase in blood flow to the region of activation. This results in the area of activation showing an increase in HbO (with accompanying decrease in Hb) and the surrounding areas showing a decrease in HbO (with an accompanying increase in Hb) [112]. However, the statement of HbO increasing in activated regions is *not* always true. There is a body of research that explores what is called the negative BOLD signal, which suggests that there are situations when increases in brain activation area actually associated with decreases in the blood volume and the HbO in the region of activation [108]. Unexpected BOLD patterns like the negative BOLD signal become even more apparent when more complicated stimuli are presented to subjects. A number of fNIRS studies have reported conflicting findings about the HbO and Hb patterns associated with brain activation. Some have noted decreases in HbO during brain activation [113, 114, 61], others have noted that the levels of HbO and Hb rise and fall during a given task, and that these changes reflect the mental processes occurring in the brain while subjects complete a mental task that involves several mental resources over the span of several seconds (i.e., at one point they could be using their spatial memory storage, then they could use their executive functioning, then they may use verbal memory storage, etc.) [114]. Since blood moves slowly in the brain, it takes several seconds for meaningful cognitive activity to be recorded with fNIRS. Research has found that it can take approximately 6-8 seconds for the increase in HbO, indicative of brain activation, to take place after stimulus onset [115, 116, 117]. This level of complexity of the fNIRS signal makes it unfeasible to analyze using simple statistical measures. Since techniques like LSTM can account for time lags in the data and the long-term cause and effect dependencies, they are a good fit for modeling such complex behaviors.

In addition, the current research in the fNIRS area treats it as a multi-channel sensor, but mostly disregards the spatial positioning of the channels. This is mainly due to the inability of the traditional machine learning algorithms like Naive Bayes, Support Vector Machines (SVMs) to incorporate spatial data without some specialized preprocessing [100]. Furthermore, the majority of fNIRS statistical and machine learning techniques are applied to just the HbO data, as HbO is known to have the strongest response to neural activation.

Convolutional Neural Networks

Convolutional Neural Networks are neural networks that specialize in processing image data or other data with a spatial structure. This assumption provides the ability to improve the performance of the forward pass of the Neural Network, and vastly reduces the number of trainable parameters in the network. CNN's are used in many domains with impressive gains in accuracy over the state of the art. Even though there have been a few recent works exploring the use of CNNs with EEG data [89, 118, 90], the application of deep learning to fNIRS data is relatively new. Hennrich et al. [119] compared brain activity to the baseline while subjects did mental arithmetic (MA), word generation (WG) and mental rotation (MR). They used a binary labeling scheme along with a Deep Neural Network (DNN) model to obtain consistently higher accuracies than the traditional classification methods. Huve et al. [120] conducted a similar study wherein they tried to classify predefined activities such as subtractions, word generation, and rest using CNNs. In [120], the input to the CNN is a matrix, such that one dimension represents the spatial location and the other dimension represents time.

Long Short-Term Memory

LSTMs have been proposed in 1997 by Hochreiter and Schmidhuber [121]. It began being widely used with the advent of speech recognition AI in the 2010s. LSTMs address one or the main drawbacks of RNNs which is the ‘vanishing’ gradient problem. Due to the chain rule multiplication done in the backpropagation phase of the RNNs, the weight changes in the RNN become more sensitive to the temporally local information and insensitive to long term dependencies. The LSTM solves this problem by introducing several logic gates that control the flow of information through the network (fig.3.3). LSTMs use addition to update itself during backpropagation, thereby eliminating the ‘vanishing’ gradient problem caused by the multiplication in the chain rule. Another key advantage of the LSTM is its ability to selectively ‘forget’ or ‘remember’ certain information. This enables it to model both long and short-term dependencies in the data [122]. LSTMs have proven stable and powerful for modeling time series behavior of data in many studies [123, 124] [125]. Our work applies the LSTMs into the fNIRS analysis domain.

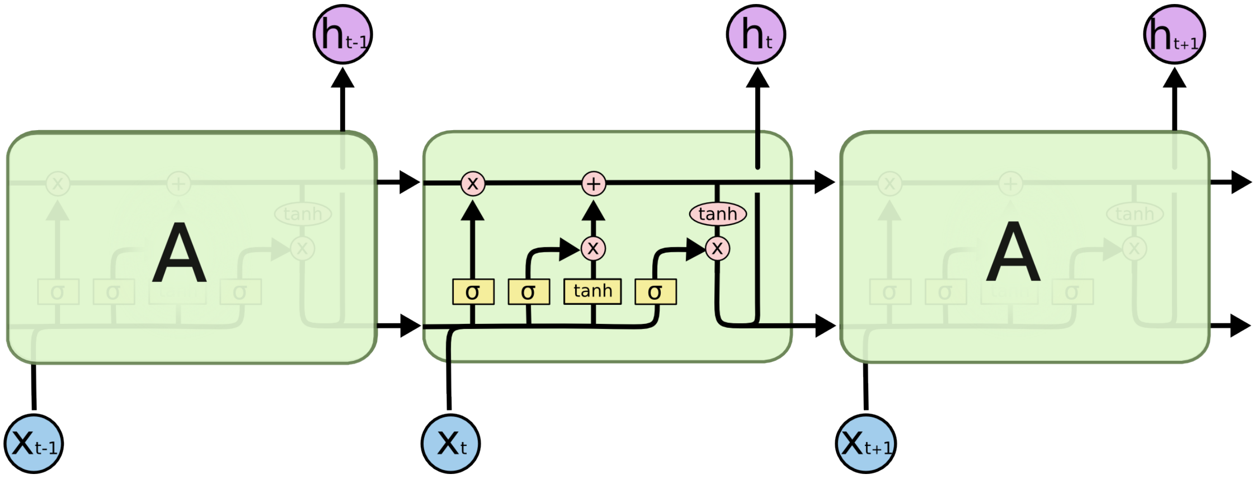


Fig. 3.3: Diagram of LSTM cell [126]

3.2.4 Related Results

Database for Emotion Analysis Using Physiological Signals

In combining the CNN and LSTMs we convert this classification problem into a video classification problem. The input being a set of images consecutive in time. Similar models have been previously used in gesture recognition [127]. We aim to apply the same idea to brain activity classification.

There have been previous attempts at emotion classification using fNIRS data. For example, Heger et al. [128] used the International Affective Picture System (IAPS) [129] and International Affective Digitized Sounds (IADS) [130] stimuli to classify emotion in a binary classification. We use a publicly available music video dataset called Database for Emotional Analysis using Physiological Signals (DEAP) in our experiments as emotional stimuli. It contains the stimulus videos and EEG data collected from 32 subjects and is widely analyzed in EEG studies [131, 132, 133, 134]. Our experiment is the first time that DEAP stimulus materials are being used in an fNIRS experiment. Thus, we now present a summary of the other existing work that uses the DEAP EEG data set as the input data. It should be noted that these methods are quite varied, they employ EEG data, and it is hard to draw parallels between these studies due to the large number of variables involved.

Rozgić et al. [131] used the DEAP EEG dataset for binary classification, with k-PCA followed by an RBF-SVM, to achieve 76.9% accuracy. Li et al. [132] used a deep belief network followed by an RBF-SVM to classify nine emotion levels of the DEAP dataset to achieve 58.4% accuracy. Another work by Wichakam and Vateekul [133] used channel selection followed by an SVM on the DEAP dataset with nine emotion levels, achieving 64.9% accuracy. Candra et al. [134] used the DEAP dataset for a binary classification problem. By calculating optimal window size and applying an SVM, they achieved 65% accuracy on the binary classification.

In this work, in contrast to the prior work on binary classification, we focus on a multi-class classification problem, and use CNNs (and later on, LSTMs) to classify the input fNIRS time series blood flow change data into three valence levels (Low/Neutral and High). In our analysis, we show the spatial dependency of the fNIRS channel data, and present results that illustrate the importance of addressing this spatial and temporal dependency. In addition, we show the accuracy gained when using multiple data types, namely HbO and Hb data, simultaneously.

3.3 Experiment Design

3.3.1 Description of Dataset Collection

A subset of music videos from DEAP [1] was selected as stimuli to elicit participant emotions. The DEAP stimuli materials consisted of 40 music videos that had been found to induce consistent self-report scores in the outer boundaries of the four quadrants of the circumplex model (Fig. 3.1), representing five experimental conditions of High Valence Low Arousal (HVLA), High Valence High Arousal (HVHA), Low Valence High Arousal (LVHA), and Low Valence Low Arousal (LVLA) as well the neutral condition (N) that was attributed to music videos that fell on the origin of the circumplex model. We selected fifteen of these videos; three videos to represent each of the five conditions. Videos were purposely selected to maximize the expected emotional reaction of participants and were thus chosen as far away from the neutral as possible in the results reported by Koelstra et al. [1]. The videos were ranked by the average labeling value and the ones with the highest absolute label value in the Arousal and Valence dimensions were chosen for this experiment. The purpose of this selection was to ensure that each participant's brain state was maximally representative of the quadrants in the valence/arousal space. Even though our goal was to classify valence levels, we chose to use videos from all quadrants of Russel's model to ensure robustness of our method.

The participants were seated in front of a computer in a normal computer user sitting position. The fNIRS signals from participants' brains were recorded using a Hitachi ETG-4000 fNIRS device with a sampling rate of 10 Hz. The device provides 52 channels of brain activity data from the frontal region of the participant's brain (Fig. 3.2). After starting the recording of fNIRS data, the music videos were presented in a randomized block design order. After the video ended, the participant filled out the self-assessment manikin (SAM), for self-report assessment of valence (Likert ratings of 1-5). After resting for 15 seconds, to allow blood flow activity to return to baseline, the participant began watching the next video. Figure 3.4 shows a screen shot of one of the music videos and a REST screen, and Fig. 3.5 shows the valence survey question that participants answered after each video session.

Fig. 3.4: Music video stimuli that was presented to user and the rest screen that was shown in between videos.

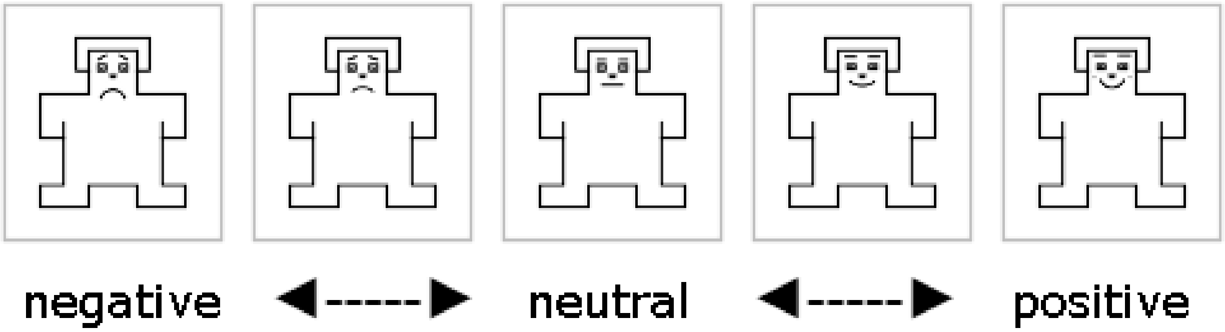


Fig. 3.5: Valence survey item that was administered after each video session.

The experimental protocol followed a block design format. The music videos were separated into three blocks, each containing videos with labels from the DEAP dataset. The order in the blocks was selected to ensure that within each block of videos the stimuli were presented in a random order to the participant (i.e. they would not be able to easily guess which type of video would be played next) while still ensuring that each block in the experiment contained video from each of the three valence conditions. The three music video blocks are shown in Table 3.1.

Music video block 1 (videos in this block were randomized in order) 1.A fine frenzy, Almost Lover: Low Valence Low Arousal (LVLA) 2.Black Eyed Peas, My Humps: High Valence Low Arousal (HVLA) 3. Blur, Song 2: High Valence High Arousal (HVHA) 4. Smashing pumpkins, 1979: Neutral (N) 5. Stigmata, in the reflection of the eyes : Low Valence High Arousal (LVHA)
Music video block 2 (videos in this block were randomized in order) 1. Sia, Breathe me: Low Valence Low Arousal (LVLA) 2. Christina Aguilera, Lil' Kim, Mya, Pink, Lady Marmalade: High Valence High Arousal (HVHA) 3. Napalm Death, Procrastination on the Empty Vessel: Low Valence High Arousal (LVHA) 4. Madonna, Rain: Neutral (N) 5. Taylor Swift, Love Story: High Valence Low Arousal (HVLA)
Music video block 3 (videos in this block were randomized in order) 1. Glen Hansard, Falling Slowly: Neutral (N) 2. White Stripes, Seven nation army High Valence High Arousal (HVHA) 3.Trapped Under Ice, Believe: Low Valence High Arousal (LVHA) 4. Wilco, How to Fight Loneliness: Low Valence Low Arousal (LVLA) 5. Louis Armstrong, What a Wonderful World: High Valence Low Arousal (HVLA)

Table 3.1: Block design of experiment. Videos within each block were randomized.

3.4 Data Analysis and Results

There were 20 subjects in the experiment, and each subject took part in 15 video sessions. However, some subjects did not finish all sessions due to fatigue, therefore we ended up with 294 sessions in all. Initially, a moving average band pass filter was applied to each fNIRS channel and we used the modified Beer Lambert Law Hirshfield et al. (2014) Chance et al. (1998) to convert the light intensity data to measures of the change in HbO and Hb in the brain. After creating the HbO and Hb datasets, we purged instances that were likely to be mislabeled from our dataset by looking for agreement between the SAM survey data reported by our participants and the expected results, based on the label of each video within the DEAP dataset. We would expect that a video in the DEAP with a label of high valence would result in similar ratings on the SAM when our participants watched that video. However, the video session survey labels did not always agree with the DEAP dataset labels as shown in Fig. 3.6. The numbers next to the circles show the number of video sessions having the corresponding DEAP dataset label on the horizontal axis, and survey label on the vertical axis. For instance, there were 45 video sessions for which DEAP label is 1 and the survey label is 4, which is contradictory. To perform the training reliably with conflict-free labels, we purged the mismatches (data enclosed in the triangular sections) from the dataset as shown in Fig. 3.6.

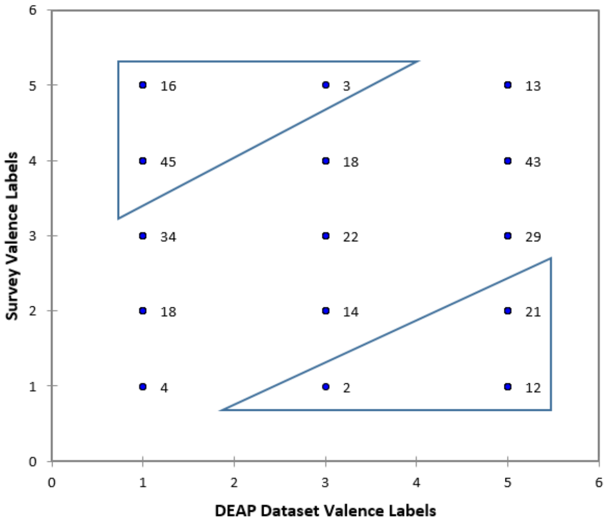


Fig. 3.6: Comparison between the survey labels and the stimulus labels in all the video sessions.

After purging the video sessions that the survey labels disagreed with DEAP dataset labels, we ended up with 195 video sessions. Since each session was 60 seconds long, and with the fNIRS being 10 Hz., the resulting dataset has 195X60X10=117000 data points.

When we looked at the label distribution from the resulting dataset, we noticed that the distribution followed an approximately normal distribution with a skew towards the high valence labels (Figure 3.7).

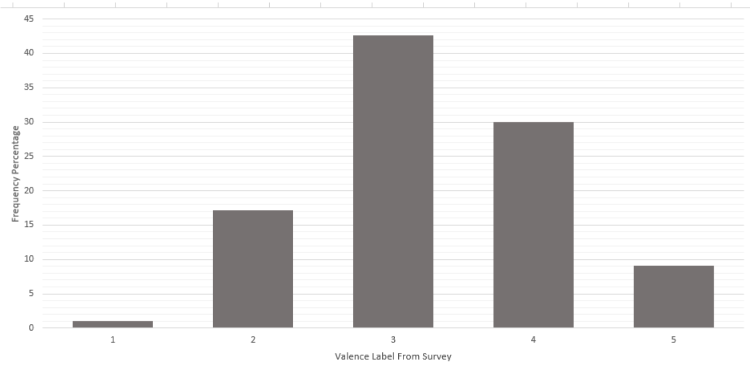


Fig. 3.7: Frequency distribution of survey labels.

To achieve a more even distribution of labels, we decided to combine the 1 and 2 labels into a single classification label and 4 and 5 labels into another classification label. The resulting label distribution is shown in Figure 3.8.

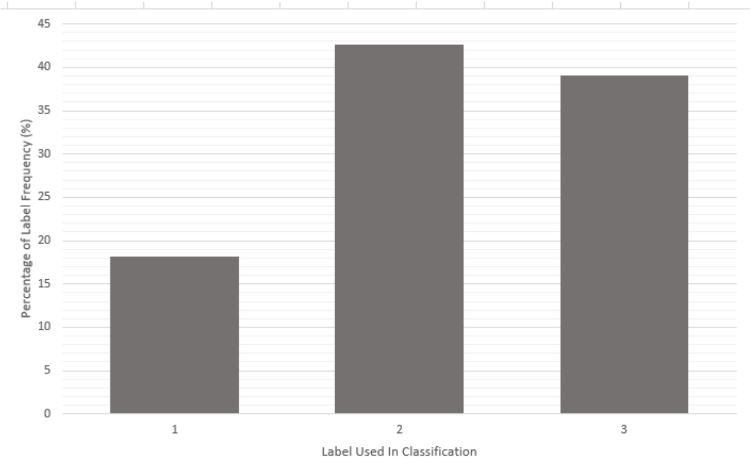


Fig. 3.8: Frequency distribution of survey labels after merging the labels.

The mapping from survey labels to classification labels is shown in Table 3.2.

Survey Response	Assigned Label
Highly Negative Valence	Negative Valence
Negative Valence	Negative Valence
Neutral Valence	Neutral Valence
Positive Valence	Positive Valence
Highly Positive Valence	Positive Valence

Table 3.2: Mapping of Survey responses to Classification Labels

To evaluate our method, we used a 5-fold cross validation with 4 of the subjects separated as the test set and the rest of the subjects used to train the classification model. After running 5 such classifications on each of the folds, and then running this test 10 times to check the stability of classifier, we report the average accuracy and aggregate confusion matrices as well as the standard deviation of the classification accuracy.

3.4.1 Preprocessing

In this study, the input data is time series cerebral blood flow change ($\triangle$HbO versus time, $\triangle$Hb versus time and the combination of $\triangle$HbO and $\triangle$Hb versus time) measured at each fNIRS channel, In our experiment, we used a Hitachi ETG-4000 fNIRS device which consists of a sensor matrix with infrared emitters and detectors placed in a 3 by 11 configuration per the diagram in Fig. 3.9(A).

The sensor probes can be placed on the subject's head using a cap (as shown in Fig. 3.2(C)). This enables the subject to use the computer in a sitting down position while the fNIRS cap is on their head. The 3 by 11 configuration of emitters and detectors results in 52 channels located in the configuration shown in Fig. 3.9 (B). In Fig. 3.9 (C), the measurement channels which are shown in green, were right aligned, and the leftmost channels were dropped from row 2 and row 4, resulting in a 5×10 matrix. Then this matrix was pooled temporally and fed into the model as input (Figure 3.10)

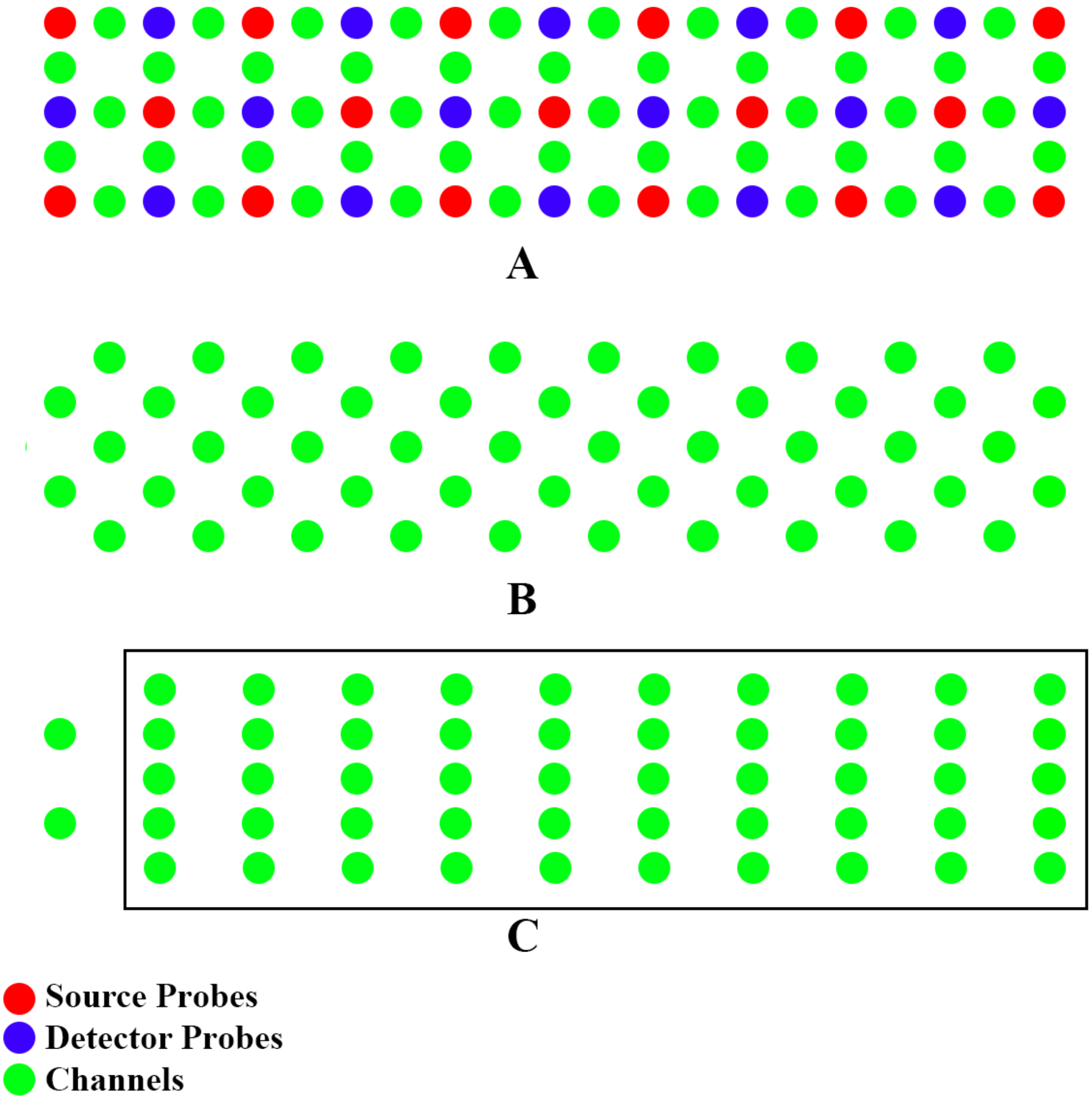


Fig. 3.9: (A)Layout of the fNIRS probes and channels. (B) Channel layout of the fNIRS sensor. (C) Right-aligned channel layout.

3.4.2 Classification Model Development

A two-layer CNN and LSTM model was used along with max pooling after each convolution. The input to the CNN for each data point was a T×5×10 matrix with the survey data as the labels

(Negative Valence, Neutral Valence, Positive Valence); the T here being the number of previous timesteps fed into the model for classifying the current timestep). The learning rate and the number of training epochs of the model were decided upon through a grid search of the parameter space. Dropout was used between the dense layers to prevent over-fitting. The high-level structure of the network used can be seen in Fig. 3.10. A more detailed view of the CNN structure is shown in Figure 3.11 The parameter values are as follows:

- No of training epochs:100
- Learning Rate: 0.002
- Regularization: L2
- Dropout probability: 0.25
- Activation Function in the Hidden Layers: Rectified Linear Units
- Optimizer: Adam

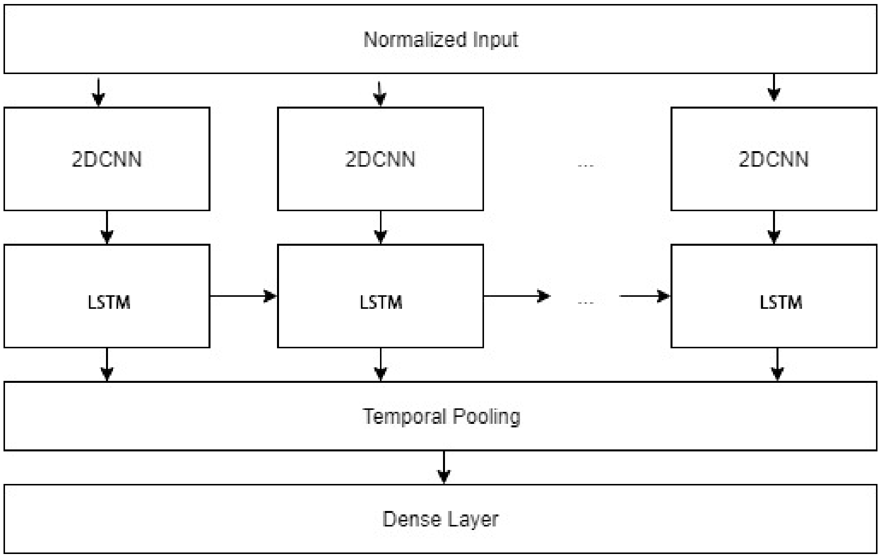


Fig. 3.10: Structure of the CNN + LSTM Model

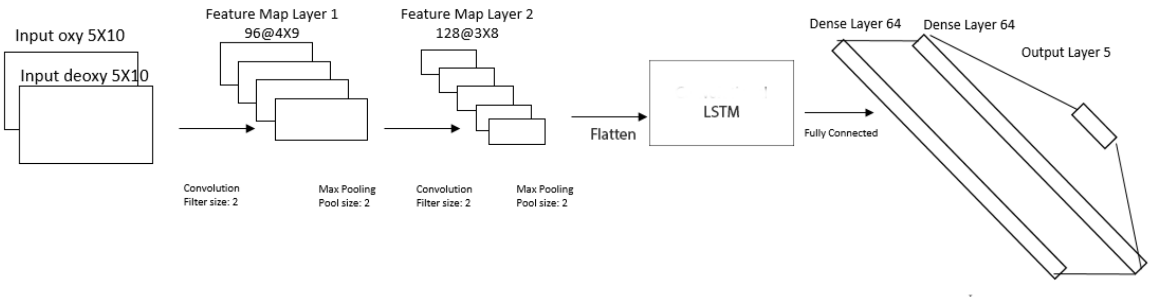


Fig. 3.11: Detailed structure of the CNN + LSTM model

3.4.3 Spatial Dependency in Input Data

In this work, we propose a new way to address the spatial dependency between fNIRS channel measurements and incorporate both HbO and Hb blood concentration data by using a CNN and LSTM based approach. The hypothesis in using a CNN for fNIRS data is that there is spatial dependency between fNIRS channel measurements, and CNNs are uniquely suited to capture this spatial dependency when measurements are fed to the network preserving the spatial proximity. In existing approaches relying on traditional models, all channels of fNIRS data are treated equally and any sort of spatial information is lost. If our hypothesis is correct, incorporating spatial information, and addressing spatial dependency would provide an increase in classification accuracy. To evaluate this effect, the network structure shown in Fig. 3.11 was used on the HbO data from fNIRS channels arranged in three different ways by randomizing the input matrix entries.

- Case 1. Original channel order
- Case 2. Rows of the input fNIRS data matrix were randomly shuffled
- Case 3. Columns of the input fNIRS data matrix were randomly shuffled

The same CNN+LSTM classifier mentioned above was run on for each of the cases to examine how the classification accuracy changed for each randomized case. The number of timesteps (T) was set to 1 in order to negate the effect of the LSTM layer and isolate the CNN layer. A drop in classification accuracy would indicate that the spatial positioning of the channels indeed holds important information towards the goal of high accuracy classification. The results of this analysis are summarized in Table 3.3, which confirm our hypothesis that there is spatial dependency, and addressing this does in fact improve classification accuracy.

	Original channel order	Shuffled Rows	Shuffled Columns
Average Accuracy	67.36% ± 0.87	67.1% ±1.15	66.26% ± 0.85

Table 3.3: Classification accuracy using original channel data vs randomly shuffled channel data. These results are aggregated from 10 runs of the 5-fold cross validation for each case

To ensure the statistical significance of our result, we did a two tailed t-test on the Average Accuracy results from each of the 10 runs of the cross validation. The comparison between Original channel order and the shuffled rows scenario provided us a p value of 0.5959 which is not statistically significant. However, when we compared the original channel order with the shuffled columns scenario, we obtained a p value of .0146 which is statistically significant at $p < 0.05$. We found that shuffling of the rows does not result in as much of a drop in accuracy compared to the shuffling of columns. This might be due to the fact that when columns are shuffled, the uniqueness of right and left brain activity is lost, thus creating a larger drop in accuracy. Also, the rows are spatially closer together compared to the columns. In summary, these results show that the higher the randomness or the higher the loss of spatial data, the higher the decrease in accuracy. This provides strong evidence that our hypothesis that spatial information is important for the model, is correct.

3.4.4 Combining HbO and Hb Data

In previous fNIRS studies, a lasting increase of HbO was taken as an indicator of cortical activation because this parameter is found to be sensitive to emotional stimulation [58, 135]. Additionally, Suh et al. [136] have reported that the direct cortical stimulation induced spatially localized increase in Hb within 12 s after stimulation. Researchers have found that HbO and Hb are negatively correlated [67]. It is therefore important to take both types of data into consideration when doing classification on the fNIRS data. In this work, we combine both HbO and Hb data to improve classifier accuracy. Another advantage of using a CNN for fNIRS analysis is the ability to seamlessly incorporate both HbO and Hb data that is produced by fNIRS.

In our previous work, we attempted to classify fNIRS data using just HbO data and Hb data by themselves and also by combining both [106]. In that particular work, the HbO and Hb data were converted into features and the best features were selected based on an Information Gain metric. In the case of the CNN-based approach proposed here, there is a unique opportunity to combine HbO and Hb data, since these two forms of data can be combined as separate channels (akin to the RGB color channels of an image) in the input data by using the network shown in Fig. 3.11. This is an intuitive way to incorporate HbO and Hb data. To confirm that this in fact helps boost the classifier performance, we tested our CNN-based method with the following combinations of HbO and Hb data, and recorded the classification accuracy:

- Only HbO data
- Only Hb data
- Combination of HbO and Hb data

The comparison of HbO and Hb classification accuracies in Table 3.4 indicates that combining HbO and Hb data provides higher accuracy than either HbO or Hb data alone. A two tailed paired t-test between the HbO and combined HbO+Hb classification performance gave a p value of 0.00001 which is statistically significant at $p < 0.01$. Same test between Hb and HbO+Hb provided a p value of 0.00001 which is also statistically significant at $p < 0.01$. These results support our assumption that both HbO and Hb data can provide important information about the valence state of a subject.

âĂć	HbO	Hb	Combined HbO and Hb
Average Accuracy	67.36% ± 0.87	67.28% ± 1.05	70.18% ± 0.87

Table 3.4: Comparison of classification accuracies obtained using HbO only, Hb only and combined HbO and Hb fNIRS data.

3.4.5 Predictiveness of fNIRS Channels

As the brain is a complex system, we would expect that certain parts of the frontal cortex would be more predictive of valence than other parts. To explore the contribution of each individual feature towards the classification, we devised a test which zeroes out one entry of the input 5×10 matrix and get the accuracy result using the same 5-fold cross validation described above. We did the test 50 times, one for each channel to come up with accuracy values for when that position was taken out of the analysis. If this 'zeroing' of the channels caused a higher drop in Average Accuracy, it can be shown that the position is more predictive and less if vice-versa [137]. Figure 3.12 shows a heatmap representation of the results of this analysis for oxy, deoxy and combined oxy+deoxy datasets. Note that the lower number and darker color in these heatmaps represent higher predictive power.

3.4.6 Effect of the Number of Timesteps on Model Performance

To explore the effect of different timestep sizes on the model performance, we chose the oxy+deoxy dataset and varied the number of timesteps to calculate average classification accuracy as described in previous sections. The resulting graph 3.13 shows that at T=0, the model behaves just like a CNN. And when T is gradually increased, the Average Accuracy keeps going up. This is evidence that the temporal nature of fNIRS data is being captured by the LSTM, After peaking at about T=10 or 1 second mark, the Average Accuracy starts dropping again, Therefore T=10 is the optimal value

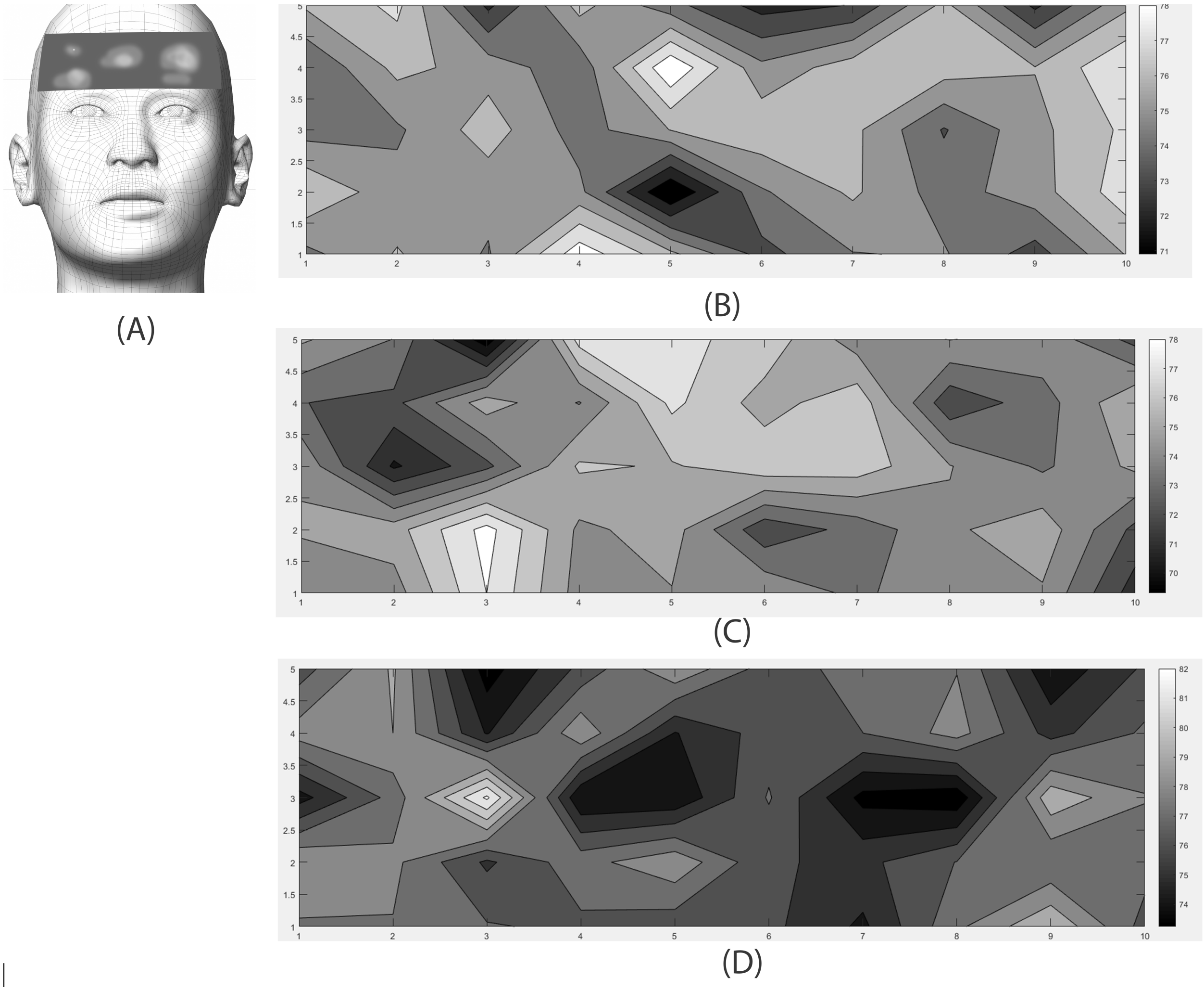

(A)
(B)
(C)
(D)

Fig. 3.12: (A) Location of the fNIRS on the head. (B) Oxy channel predictive power heatmap (C) Deoxy channel predictive power heatmap (D) Oxy+Deoxy channel predictive power heatmap

of T for high classifier performance.

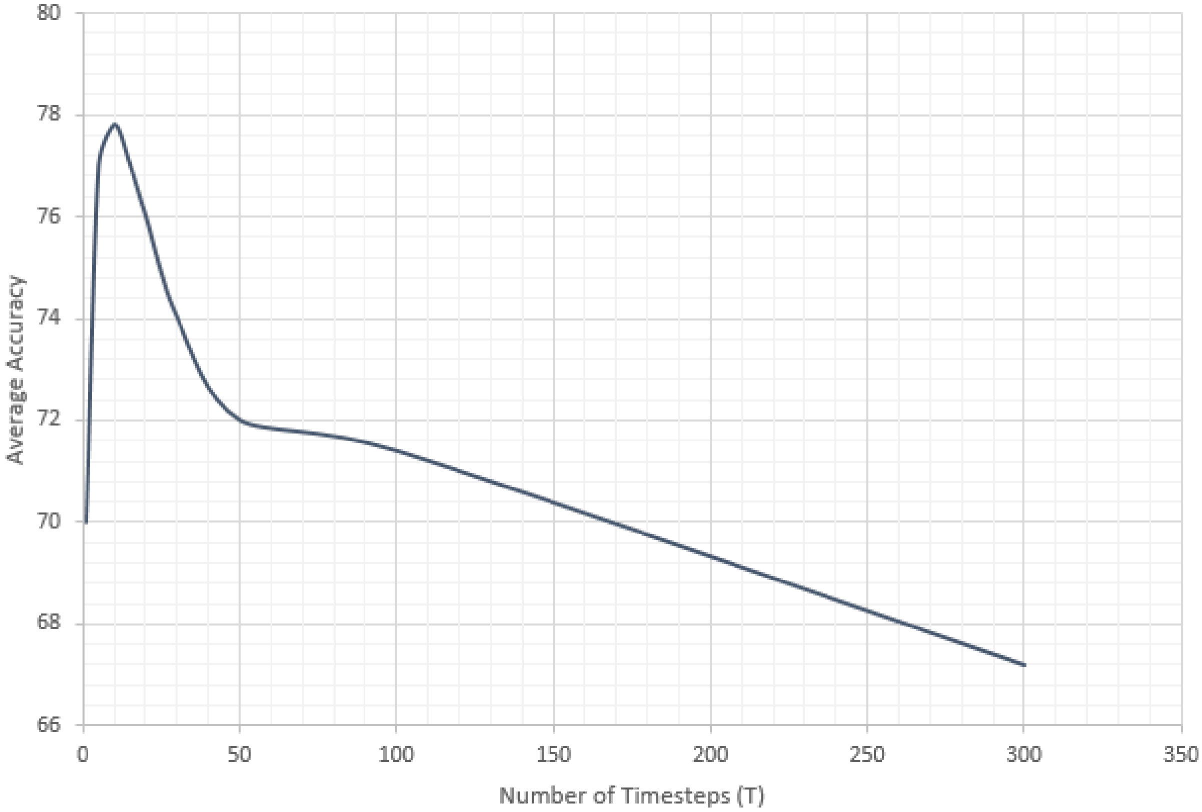


Fig. 3.13: Variation of Classification Performance with Number of LSTM timesteps (T)

3.4.7 Label-wise Model Performance

In order to investigate further into the label-wise prediction performance of our classifier, the classification results of combined HbO and Hb data from the most predictive case (HbO+Hb with T=10) are detailed in an aggregated confusion matrix presented in Fig. 3.14. The aggregated confusion matrix was obtained by summing up all the individual confusion matrices of the 5-fold cross validation. From the confusion matrix it is evident that the model has a slight tendency to classify the High Valence label more frequently than the Low Valence label. This could be due to the slight skew in the data towards the high valence label as seen in the Figure 3.8.

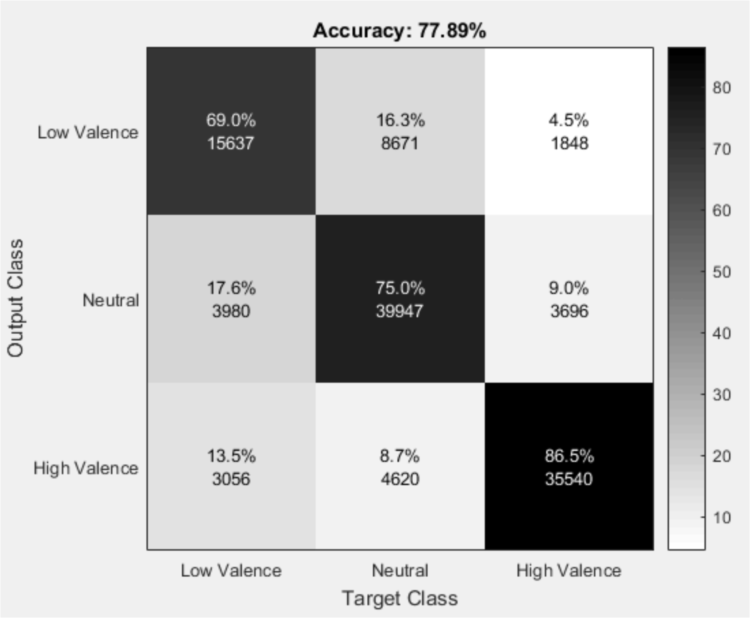


Fig. 3.14: Aggregated Confusion Matrix for HbO + Hb with T=10

3.5 Analysis of Results

The results presented above show that there is spatial dependency between fNIRS data channels and addressing this can provide a boost in classification performance. Since our fNIRS device has 52 channels, we were able to obtain enough spatial data to employ a CNN-based approach and address this dependency that previous methods were unable to address.

Also, the use of HbO and Hb data together increased classification accuracy further, thus showing that both HbO and Hb data contribute to the picture of valence. By using the DEAP dataset stimuli we were able to make reasonable assumptions about which instances were likely mislabeled. Another important result is the fact that the inclusion of time domain data using LSTMs increased the Average Accuracy. And we found that there is an optimal number of T=10 timesteps for the model to perform best. This also shows that time series behavior of fNIRS data seems to provide valuable information towards the valence classification.

3.6 Conclusion

We have presented a CNN and LSTM-based method to classify the valence level of a computer user based on fNIRS data. The model is trained to classify the input fNIRS time series blood flow change data into three valence levels. An experiment has been conducted with 20 participants, wherein they were subjected to emotion inducing stimuli while their brain activity was measured using fNIRS. Self-report surveys were administered after each stimulus to gauge participants' self-assessment of their valence. We explored the effect of spatial information on the accuracy of classification. We learned that addressing spatial dependency does in fact improve the classification accuracy. The proposed method provides 67.1% and 66.26% average accuracy when using HbO data and Hb data, respectively. We were able to increase the classification accuracy further to 70.18% by using HbO and Hb data together. By setting number of timesteps (T) to 10 (1 second) we could further improve the accuracy level to 77.29%. Thus, our proposed method provides a significant increase in average classification accuracy when compared with traditional classifiers. This opens up new opportunities to improve classification in the field of fNIRS research and Human Computer Interaction (HCI) research in general. These results show the power of CNNs and their applicability in domains far removed from image recognition. CNNs have a variety of properties that make them well suited to be applied to fNIRS data. The high spatial resolution of the fNIRS as well as the complementary measurements (HbO and Hb) obtained by the fNIRS device make a strong case for this model as an fNIRS data analysis method. Capturing the time series behavior using LSTM further strengthens the model. Therefore, it is evident that both spatial and temporal information play a role in the emotion classification process. The ability of the method to extract relevant features (instead of using hand-crafted features) is well suited to the problem of brain activity classification since it is still a challenge to find reliable underlying models of how the human brain works. An interesting implication of this research is to show that preprocessing on the data based on predictivity of the features could further improve the classifier performance.

CHAPTER 4

IDENTIFICATION OF POTENTIAL TASK SHEDDING EVENTS USING BRAIN ACTIVITY DATA

4.1 Introduction

As computing devices become more ubiquitous, the need for greater human machine symbiosis becomes an important factor. This concept, of human and computer agents working together to accomplish a goal, or set of goals, is referred to as Human Machine Teaming (HMT). In any teaming environment, whether it is a team of all human agents, a team of all machines, or a combination of the two, it is important that resources within the team be allocated as efficiently as possible such that the team may achieve its goal while simultaneously putting the least amount of strain possible on any of the team's individual members. The finite resources of an HMT, such as the processing power of a machine, or the limited cognitive capacity of a human agent, could be viewed as potential bottlenecks within an HMT system, where the team's ability to accomplish a task may falter. Though the processing power of a machine may have been the primary factor that stopped HMTs from achieving optimal performance in the past, processing power is now an easily obtain-able resource. As such, recent efforts to improve the performance of HMTs have shifted focus to improving the communication modalities between humans and machine agents. As human agents have limited cognitive capacity within a time sensitive task environment, ideal task performance is dependent on optimizing humans' information processing capabilities, which are affected by the complex interplay between their perceptual processing load, mental workload, and emotional state. Task performance within an HMT is also dependent on the ability of machine agents to detect and interpret the signs of potential overload on the part of the human agent, and to have the ability to take meaningful action to assuage, or at the very least reduce, the load placed on the human agent.

In order for a machine agent within an HMT to properly adapt to the task at hand, the machine must not only possess knowledge relevant to the task that needs to be completed, but also the amount and type of *mental workload* that is currently being placed on the human agent(s) within the HMT. The machine must also be able to discern if this amount of workload is significant enough to induce degradations in the human agent's or team's performance that might limit the HMT's ability to complete the current task. *Mental workload*, in this context, is the brain's finite amount of processing capacity to allocate to a given task. As theoretical and experimental work by Wickens' Multiple Resource Theory (MRT) [138] has shown, there are different types of cognitive resources that the brain is able to allocate simultaneously, and the overload of one type of cognitive resource does not necessarily lead to the overload of another [138]. When a person is required to perform multiple tasks that require the same type of mental resource, that resource may become overloaded, and as a result, that person's performance at the given task will degrade. If overload is adequately high, the individual may eschew the task altogether, an event known as "task shedding" [139]. A common area where this concept expressed is in the field of piloted aircraft and UAVs. In those cases, the autonomous system and pilot cooperate to achieve objectives [140]. For this type of cooperation to work, the machine needs to be able to sense and intervene when the human's performance starts dropping due to increases in task load.[141] [142]. Parasuraman and Hancock have showed that task shedding can be triggered by high workload and low certainty [143]. Though researchers have previously attempted to model this task shedding behavior using simple and basic tasks [144, 145, 146, 147] as well as more complex real-world scenarios. [148, 149], more accurate predictions about when task shedding events are likely are needed if these models are to be implemented in real time.

In this chapter, we introduce a novel method by which one can predict task shedding instances using across-task, across-subject, machine learning methods. The goal is to introduce an agnostic framework, not tied to any specific task that a particular HMT would try to complete. To achieve this goal, we created a model, using psychophysiological data recorded from a Functional Near Infrared Spectroscopy (fNIRS) device, to detect when task load on a human participant was high, and task shedding instances were therefore likely to occur. To isolate specific types of mental workload (working memory, visuospatial attention), we trained models on multiple cognitive benchmark tasks used widely in the fields of cognitive psychology and cognitive neuroscience, and tested those models on more ecologically valid tasks. We show that such a system can provide the reliable prediction of such events such that an autonomous agent with access to this type of physiological data would be able to predict when moments of mental overload might lead to performance decrements or task shedding events, and as a result would be able to take over for, or provide assistance to, the human agent within the HMT.

4.2 Background

4.2.1 Human Machine Teaming and Task Shedding

The importance of finding a way for an autonomous agent in an HMT to detect when the human agent may be subjected to events of higher cognitive workload, and thus task shedding events, has been shown in past systems design research [150]. Past work in the fields of human factors and cognitive engineering have provided evidence that when a human agent's performance is supplemented with a machine agent's ability to assist on tasks, that the reported workload of the human agent decreases. This decrease in perceived workload coincides with an increase the human agent's self-confidence and trust in the HMT and well as an increase in the overall performance of the HMT [151]. Despite these advances, there are still many issues that need to be addressed to ensure HMTs can be optimized to task performance [152]. With these issues in mind, we use predictive modeling on multiple cognitive resources to detect when increases in mental workload are likely to lead to performance decrements.

4.2.2 Using Psychophysiological Sensors to Measure Workload

Task shedding detection through brain activity requires sensors that are robust to noise, portable and non-invasive. The fNIRS device works well for this application as the device can be setup quickly and is able to target specific areas of the brain that are implicated in cognitive resources that are prone to becoming overloaded when engaged in cognitively demanding tasks [153, 154]. The fNIRS device works by using multiple pairs of optodes that are placed on the scalp and pulse infrared light (690nm and 830 nm) through the skull and into the brain. The reflected light intensity that is received by the detector is dependent on the amount of oxygenated and deoxygenated hemoglobin in the incident area over which the optodes are placed [155]. Since oxygen is consumed during the metabolic processes involved in brain activity, the concentration of hemoglobin is correlated with increased brain activity. fNIRS has in the past been used for classification of workload levels [156]. More specifically, the fNIRS' ability to measure brain activity in the frontal cortex of the brain gives it a unique ability to predict workload levels, as greater activation in the frontal cortex has been found to be associated with higher levels of mental workload [157, 158].

The ability to measure mental states, and thus workload, has already been well documented in the human factors literature [159]. Other work has found great success in being able to predict mental workload in real-world computer environments using fNIRS [160, 161]. As advances in both fNIRS technology and portability increases, fields such as brain computer interfacing (BCI) have argued that the fNIRS' increased spatial resolution would make fNIRS an invaluable tool for collecting real time psychophysiological data and building systems that incorporate and adapt to that data in real-time [162]. The portability of the fNIRS system, combined with its ability to measure workload and communicate those measurements to a machine agent could provide a solution to make strides towards correcting documented issues in intelligent system designs [163].

4.2.3 Machine Learning Classifiers on fNIRS data

Researchers have used traditional machine learning classifiers such as Support Vector Machines [164], Artificial Neural Networks [165], Hidden Markov Models [166] as well as other statistical methods [167, 168] to preprocess and classify fNIRS data. However, more recent work has demonstrated the ability of using deep learning algorithms [119] to capture the characteristics of the fNIRS signal. One category of deep learning algorithms; Convolutional Neural Networks (CNNs) [169] are typically used in the image processing domain because of their ability to capture the spatial structure of image data. They are well suited for fNIRS analysis due to the same reasons. Our previous work [170] showed that the oxygenated and deoxygenated data provided by fNIRS can be used similar to the RGB channels of an image when fed into a CNN classifier. We also used a Long Short-Term Memory (LSTM) network, to capture the time series behavior of the fNIRS data. LSTMs are a version of Recurrent Neural Networks, which can capture long and short-term dependencies in the data [171]. LSTMs have become popular for machine learning on Electroencephalography data [172, 123]. They have also recently been used in fNIRS analysis [173]. in [170], the above described model was used in across subject classification by dividing the subject pool into folds. In this pape, we use the same model on across subject, across task, three label classification tasks.

4.3 Methods

4.3.1 Experimental Protocol

fNIRS data was collected from 45 participants (14 female, 31 male, mean age = 26, min age =21, max age =36) who were selected from the undergraduate and graduate student population at a University in the Northeast. The data was collected using a Hitachi ETG-4000 fNIRS device at a sampling rate of 10Hz. The optodes were arranged into an fNIRS cap with a 3X11 probe configuration and were placed on the participant's forehead area in a symmetrical manner (Fig.4.1). Using this configuration, the fNIRS device is able to capture information about the oxygenated and deoxygenated hemoglobin levels in the frontal cortex of the participants. The fNIRS was calibrated to ensure that all probes were recording proper readings and adjusted to account for ambient light. After setting up the fNIRS device on the participant's head, a Patriot Polhemus 3D digitizer device was used to measure the location of each source/detector to account for variances in head size and shape. All participants gave informed consent under the restrictions and guidelines of the University's Institutional Review Board.

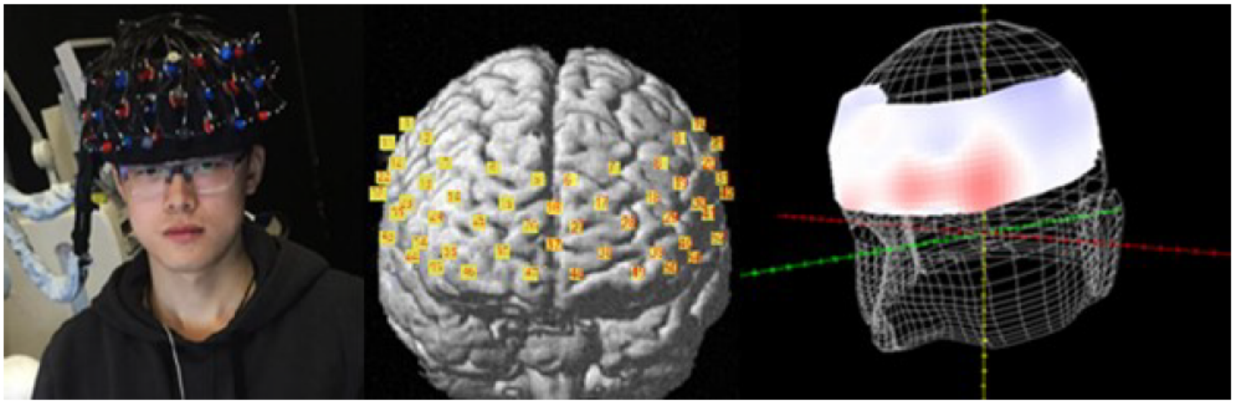


Fig. 4.1: The 3 x 11 fNIRS probe configuration.

From the set of 45 participants, a subset of these participants (n=25, 7 female, 18 male) completed the cognitive benchmark tasks described in section 4.3.2, as well as the triage cyber analyst task described in section 4.3.3. Both the cognitive benchmark tasks as well as the triage task used a variable interstimulus interval (ISI) between the offset of a trial and the onset of a new trial, during which a cross fixation point in the center of the screen was displayed on the screen. The length of the ISI was an exponential distribution (mean= 4s, min=2s, max=8s). The remaining 20 participants performed the Multi Attribute Task Battery (MATB) test bed described in Section 4.3.3. Since an adjustment in screen size has been shown to be correlated with certain physiological responses [174], all tasks were displayed to the participants on a 22-inch monitor with a screen resolution of 1280 x 1024 pixels. Participants were seated in a stationary chair so that the distance between their eyes and the monitor was 65cm. All participants would begin each fNIRS session with a 30 second session of controlled rest during which the participant fixated on a plain black plus symbol in the center of a white display. Participants would then perform ten trials of a reaction time task before they began the rest of the cognitive benchmark tasks. To mitigate the fatigue effects involved in cognitive testing [175], the benchmark task order was randomized between subjects.

4.3.2 Stimulus Materials: Training Data

Visual-Lexical Processing, Adaptive Words

This adaptive words task was developed to induce workload on participant's visual lexical processing resources. In this task, the words for the numerical values of the digits one through eight were displayed vertically for a variable amount of time in the center of the screen. The participant's goal was to determine whether the word that was displayed on the screen corresponded to either an odd or even numerical value.

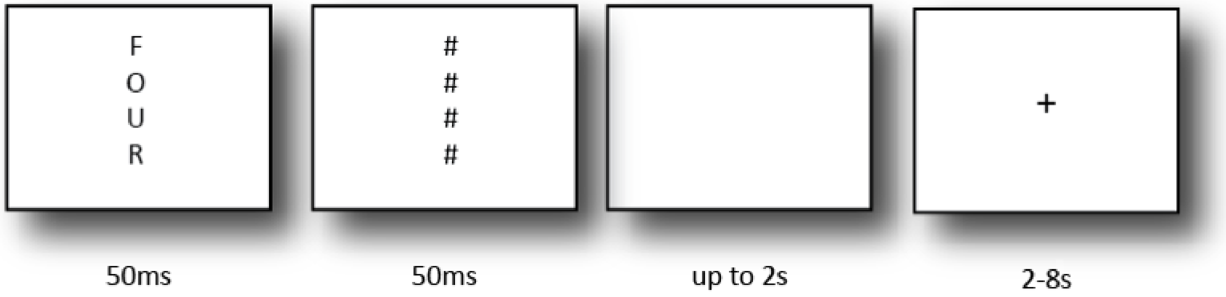


Fig. 4.2: Adaptive Words Task Presentation.

Visual Search Task, Visual Search

The visual search task is designed to cause cognitive load on people's visual processing resources, and was modeled after the task design developed Wang, Cavanagh, and Green [176]. A circular array of nine letters consisting of a distractor (backwards Ns) and a target (normal facing Ns) was displayed to the participant for a variable amount of time. The participant's task was to determine as to whether or not the target was displayed within the array.

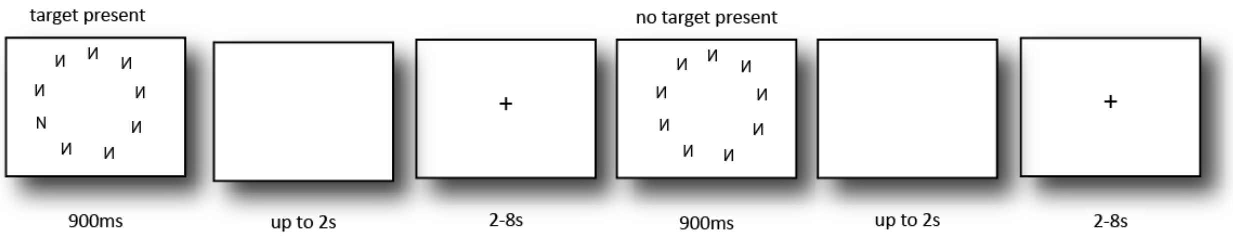


Fig. 4.3: Visual search task presentation.

Response Inhibition, Go No-Go

The response inhibition task was the go no-go task, which involved one target stimuli (a large blue circle), and one distractor stimuli (a large blue square). The development of stimulus materials was guided by Huettel, Mack, and McCarthy [177].The participant was tasked with responding to the stimuli if the target was presented, and not responding to the stimuli when the distrator was presented.

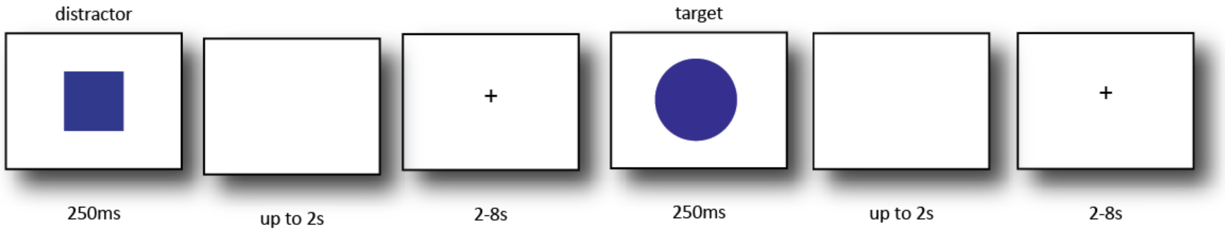


Fig. 4.4: The go no-go task presentation.

Working Memory, N-Back Task

The N-Back task (Fig. 4.5) is designed to cause cognitive load on people's working memory resources by requiring participants to hold a stream of characters in their working memory and responding when a new character that is presented to them matches one of the characters they are currently holding. The task development was based on Harvey et al [178]. The task presented participants with a series of letters, a single letter at a time, for a duration of 500ms each. The letters would appear in the center of the screen with a plain white background. Only the letters B, D, G, T, V along with their lower-case variants, b, d, g, t, v, were used. Before each block, participants were given an ‘n’ value of either one, two, three or four. The participant’s goal was to determine if the current letter presented to them matched the letter that was ‘n’ presentations behind the current letter that was displayed (case insensitive). For example, in fig. 4.5, if an ‘n’ of two had been given to the participant, then the correct response would be ‘yes’. If the participant was given an ‘n’ of one, however, the correct response would be ‘no’.

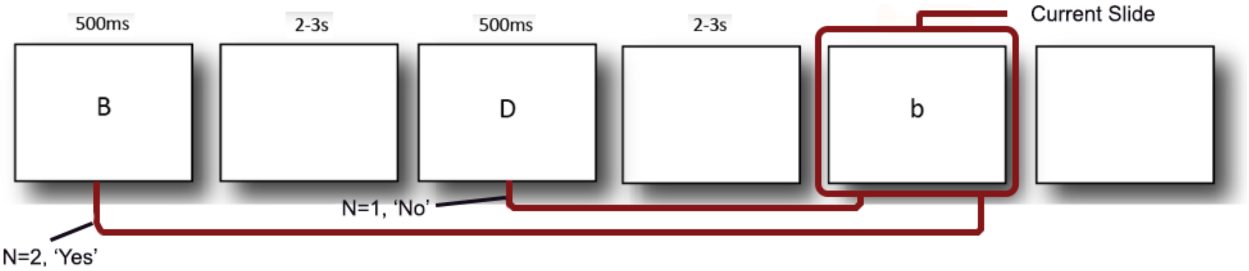
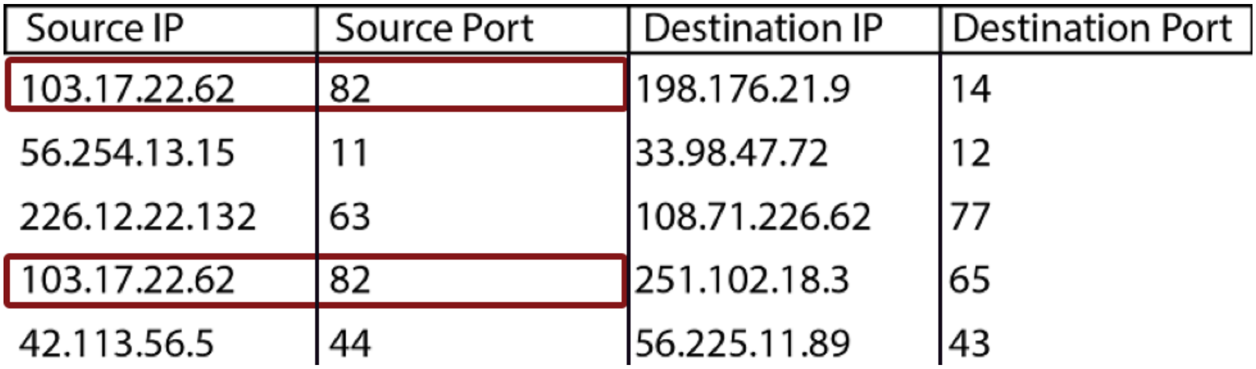


Fig. 4.5: N-back task presentation.

4.3.3 Stimulus Materials: Test Data

Triage Task

The Triage Analyst task acts as an ecologically valid representation of a cyber-security network analyst’s position, and is based on the work of Greenlee et al. [179]. The task involved the participant viewing what is at first an empty table in the center of the screen. The table headings were ‘Source IP’, ‘Source Port’, ‘Destination IP’, ‘Destination Port’. Participants were informed prior to the task beginning that they did not require working knowledge of the terminology involved in order to complete the task. The table would then be populated with incoming ‘transmissions’ on the ‘network’ the participant was monitoring. Starting from the top of the table, new ‘transmissions’ would fill the table until a maximum of five ‘transmission’ were on the screen. After five ‘transmissions’ were shown on the screen, the bottom transmission would be removed from the screen to make room for a new incoming transmission at the top, bumping the rest of the table down one slot. The participant was tasked with detecting ‘intrusions’ on the network. These ‘intrusions’ were defined as either two different ‘transmissions’ on the table having the same destination information (both ‘Destination IP’ and ‘Destination Port’), or two different ‘transmissions’ on the table having the same source information (both ‘Source IP’ and ‘Source Port’). The participant was only asked to identify if the newest (topmost) ‘transmission’ was or was not an ‘intrusion.’ Figure 4.6 shows an example intrusion. The triage testbed tracked participant response times as well as their performance (logging correct, incorrect, or no response events) throughout the task.



Source IP	Source Port	Destination IP	Destination Port
103.17.22.62	82	198.176.21.9	14
56.254.13.15	11	33.98.47.72	12
226.12.22.132	63	108.71.226.62	77
103.17.22.62	82	251.102.18.3	65
42.113.56.5	44	56.225.11.89	43

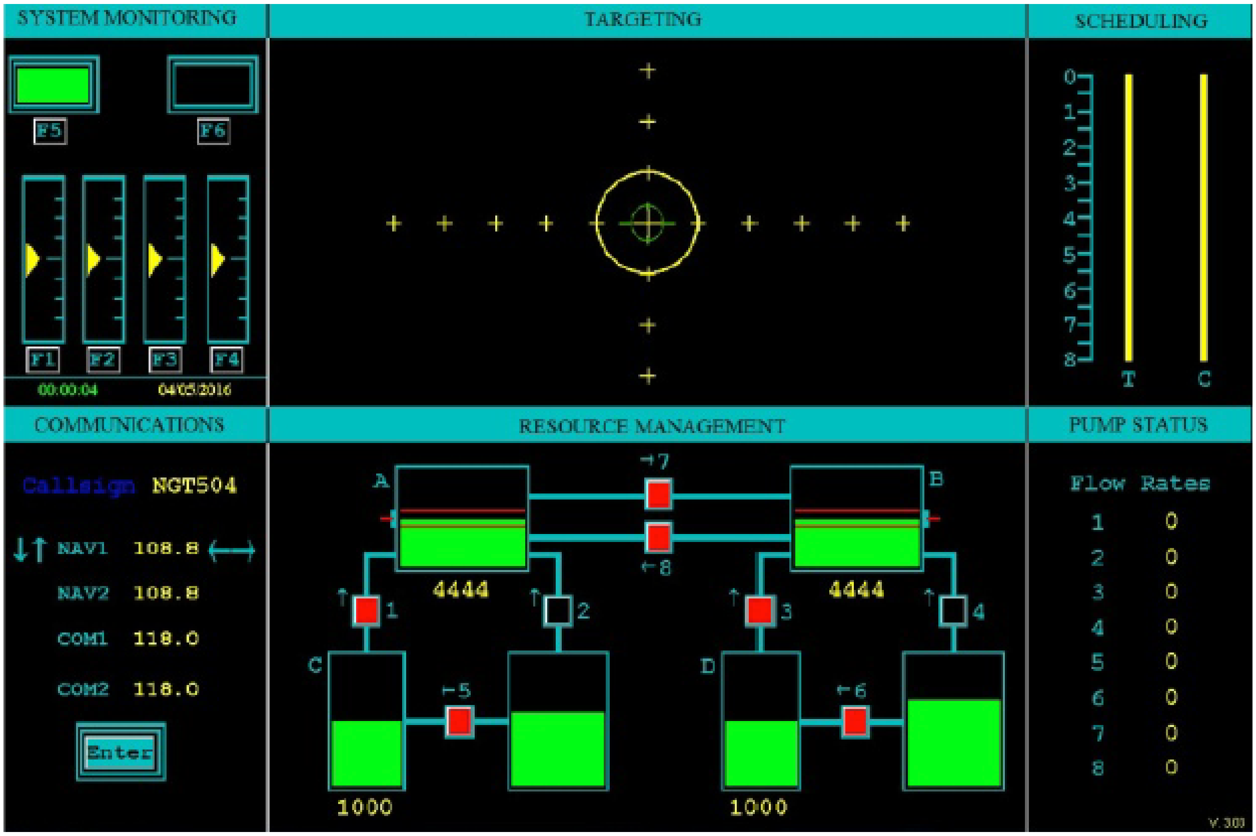
Fig. 4.6: Source Intrusion.

Multi-Attribute Task Battery

This task involved a complex multi-tasking scenario using a variation of the Multi-Attribute Task Battery (MATB) [180]. This difficult task made it imperative that high achieving users not become overloaded or overly stressed, forcing them to prioritize their actions based on the most time sensitive or important needs of the task at the time. We used the Air Force’s updated version of the Multi-Attribute Task Battery (AF_MATB) [181], and we chose a difficulty level that required a good deal of mental effort and multi-tasking.

With the difficulty of the task and the high level of multi-tasking required, the task was nearly impossible to complete perfectly, and our pilot tests showed that all subjects had to remain extremely engaged during the entire task to receive an adequate performance score. Like the original version, AF_MATB consists of six windows, which provide information about four different subtasks (see Figure 4.7) these subtasks include: System Monitoring, Communications, Resource

Management, and Tracking. The ‘tracking’ subtask was disabled during this experiment in order to reduce physical motion of the participant activating the motor cortex in the brain as well as reducing motion artifacts from the fNIRS data. The last two windows, which contain Scheduling and Pump Status information, are resources that the user can use to improve performance during the task. The first requirement in the system monitoring subtask (top left pane) is to keep track of the two lights on the top of the window and keep them at their original status by toggling them on/off using the buttons below them. The second requirement in the system monitoring subtask requires the user to be aware of the four scales and press the corresponding button if any of the scales deviates from the center by more than one tick mark. The communication subtask (bottom left pane) involves the subject listening for verbal requests to change the frequency of specific radios. The verbal requests include a call-sign and if the call-sign is not the call-sign of the subject, the request is to be ignored. The resource management subtask (bottom center pane) requires the subject to keep the fuel levels in tank A and B within 500 units of the initial level of 2500 units each. The pumps connected to the tanks can be used to pump fuel from the lower supply tanks to tanks A and B. The pumps can be turned on/off by clicking on the particular pump. However, these pumps can malfunction for periods of time during the experiment. If a pump is malfunctioning, it can not be turned on. The AF_MATB keeps track of, and outputs, a thorough report of each subject’s performance data on the various subtasks. In our pilot studies, we selected a MATB difficulty level that would result in the majority, if not all, of the subjects having difficulty executing every task perfectly. They would have to multi-task, prioritize, and accept that while their performance would likely be imperfect, they must keep from becoming frustrated in order to complete the demanding MATB scenario.



Callsign NGT504

NAV1 108.8

NAV2 108.8

COM1 118.0

COM2 118.0

Flow	Rates
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0

Fig. 4.7: The Multi-Attribute Task Battery.

4.4 Data Analysis

4.4.1 fNIRS Data

The fNIRS provides 104 channels (52 channels of oxygenated hemoglobin + 52 channels of de-oxygenated hemoglobin) of data at 10 Hz. This data was low-pass filtered with a cutoff frequency of 0.1 Hz to remove cardiac, respiratory and high frequency unwanted noise [182]. The time series data from fNIRS was divided into 5-second blocks and the average value of the fNIRS data was obtained for each of the 104 data channels for that 5-second period.

4.4.2 Label Calculation

In this experiment, we relied on objective performance data to calculate our labels. The labels were calculated based on data logged by the software that presented the task itself. This way, we were able to match the exact times the stimuli was presented and the duration of the stimuli using the log files from the task software. The log files for the benchmark and triage tasks were generated by capturing information about when an event was triggered via the stimulus software, as well as information about when the participant responded to the event. Every event logged by the software was tagged with a UNIX timestamp. The participant's reaction times were then calculated, in milliseconds, for each trial included in the logfile. The logfile also indicated as to whether or not the participant responded correctly to the trial. The MATB logfile collected data at 10 Hz. The datapoints collected include the number of accurate and inaccurate user responses within each time segment and the difference between the current tank level from the target tank level in the resource management task.

Using these performance and taskload metrics (further discussed in section 4.4.2) we introduce a measure called 'Task Shedding Index (TSX)', which is a combination of taskload and performance. The TSX is used as an indicator of the potential of the user for task shedding.

Task Shedding Index (TSX)

As described in section 4.2.1, when a user is put in a high workload scenario, and the workload of the task is over a certain threshold, they tend to shed that task or switch to another task on which they may be able to perform better [183]. At this point the performance on the prior task would degrade, and would therefore be an apt point of intervention on the part of the machine agent in the HMT[141]. To account for this, a hybrid Workload and Performance model is needed to predict task shedding tendencies. Using the distance between normalized task load and normalized performance (detailed next) as our ground truth in this analysis, the Taskload (T) and Performance (P) of each subject are normalized and the Task Shedding Index (TSX) is defined as the subtraction of normalized P from normalized T (Equation 4.1).

$$Task\ Shedding\ Index\ (TSX) = Normalize(T) - Normalize(P) \tag{4.1}$$

The ranges of the T, P and TSX are shown in figure 4.8

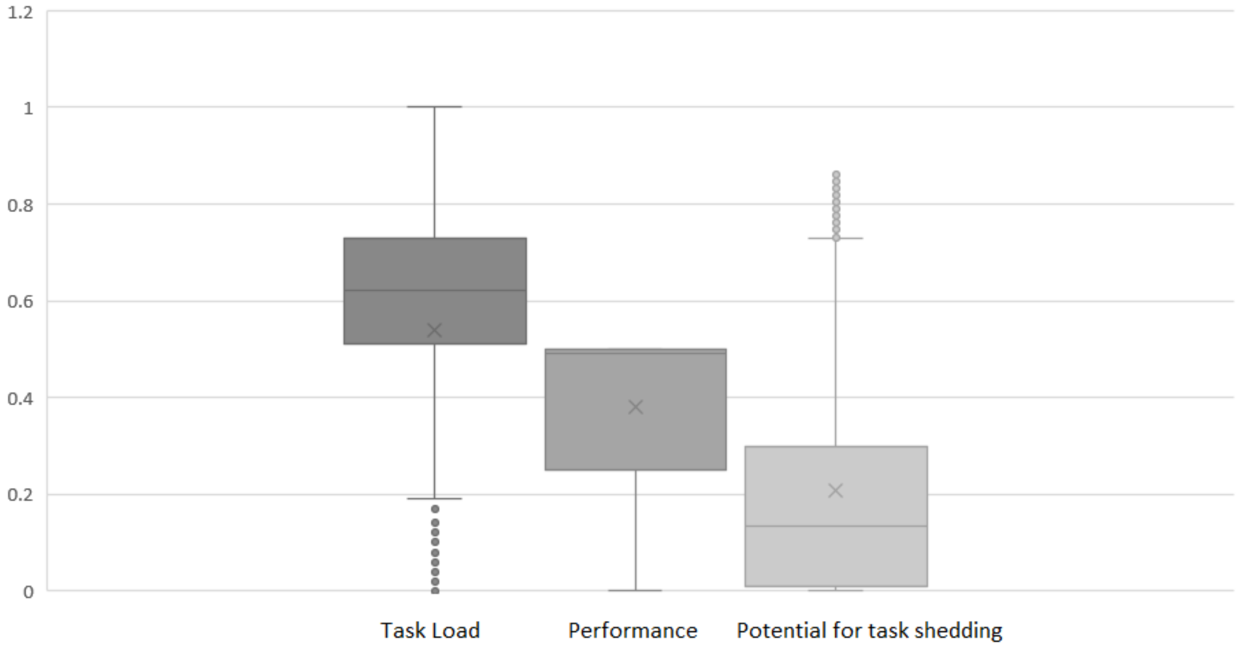


Fig. 4.8: Data ranges for Task Load, Performance and Task Shedding Index.

Task load was obtained by adding together the number of tasks that were presented to the user during the 5-second time period. If there were certain tasks that started or ended during the 5-second period, they were apportioned according to the amount of time they were present during the time frame.

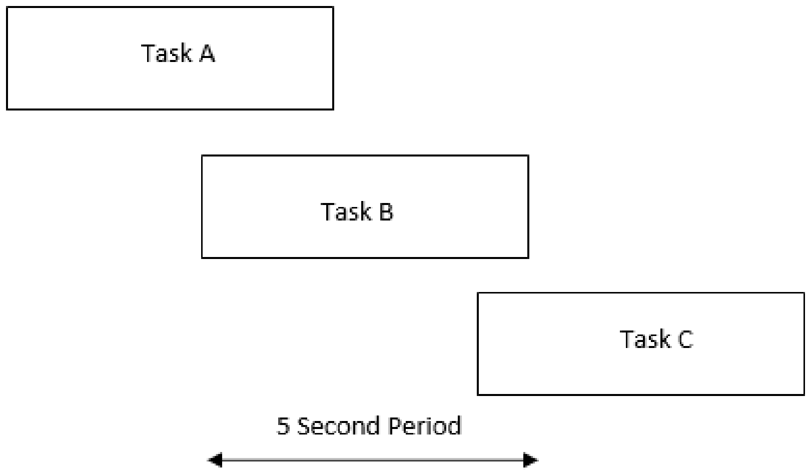


Fig. 4.9: Apportioning tasks to the time segments.

In Figure 4.9, Task A would contribute 50% to the current time period's task load. Task B would contribute 100%, and task C would contribute 25%. Therefore, the task load would be calculated according to Equation 4.2,

$$Apportioned\, TaskLoad(T) = 0.5 + 1 + 0.25 = 1.75 \tag{4.2}$$

Performance is apportioned similarly as in Equation 4.3,

$$Apportioned\, Performance(P) = 0.5 * TaskA_{Perf} + 1 * TaskB_{Perf} + 0.25 * TaskC_{Perf} \tag{4.3}$$

To calculate performance in the benchmark and Triage tasks, the onset, duration and accuracy of user response were saved to a file during the experiment by the test bed. Calculating user performance for Benchmark and Triage tasks was done using Equation 4.4 for each 5-second time period.

$$Triage_{performance} = \frac{(Correct\ Responses) - (Incorrect\ Responses)}{TaskLoad(T)} \tag{4.4}$$

Since MATB is a multi-attribute task, the performance calculation for MATB was done by taking the compound performance of all the subtasks. The MATB contains following subtasks,

- Time point
- Number of accurate light toggle responses (L)
- Number of accurate gauge responses (G)
- Number of accurate communications responses (C)
- Number of inaccurate light toggle responses(l)
- Number of inaccurate gauge responses (g)
- Number of inaccurate communications responses(c)
- Difference of Tank A value from desired target value (AD)
- Difference of Tank B value from desired target value (BD)

Equation 4.5 was used to calculate the user performance for the MATB task.

$$MATB_{performance} = \frac{(L+G+C)-(l+g+c)+FuelTankA_{performance}+FuelTankB_{performance}}{TaskLoad(T)} \tag{4.5}$$

Where Fuel Tank A performance is calculated by using the following equation,

$$\begin{cases} 0 & ABS(\Delta AD) > 0 \\ 1 & ABS(\Delta AD) \leq 0 \\ & \end{cases}$$

Same method was used to calculate Fuel Tank B performance.

4.4.3 Discretization of ground truth Labels

The TSX labels follow a positively skewed distribution. We are interested in breaking down the distribution into three levels of TSX, namely ‘low’, ‘moderate’, ‘high’. This categorization has been chosen because it allows us to look at the ‘moderate’ level of TSX as the desired level, the low level of TSX as when user is ‘idling’ and ‘high’ level of TSX as when user is overloaded. This enables different strategies for dealing with each of these cases. The TSX labels were discretized using Equal Frequency Discretization over all the datasets. The Equal Frequency Discretization algorithm sorts all values of continuous variable in ascending order, and divides the range into three intervals so that every interval contains the same number of sorted values. The resulting ranges from Equal Frequency Discretization for the TSX labels were:

- Low TSX : Less than 0.12
- Moderate TSX : 0.12 to 0.323
- High TSX : 0.323 upwards

The resulting discretized label distribution for each of the datasets is given in Table 4.1

	Low TSX	Moderate TSX	High TSX
Adaptive Words	130/36%	115/32%	120/33%
Visual Search	298/45%	205/31%	157/24%
Go no go	240/37%	253/39%	153/24%
nback	210/31%	277/42%	180/27%
Triage	2050/34%	1802/30%	2106/35%
MATB	402/23%	678/40%	614/36%

Table 4.1: Ground truth label distribution for each of the datasets.

4.4.4 Classification Model

We trained a classification model that we developed in our earlier work to classify fNIRS data using a combination of a Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) as shown in figure 4.10. Since CNNs are well-suited for capturing the spatial nature of fNIRS data and LSTMs are good at capturing the temporal behavior of fNIRS data, this model provided performance improvement over traditional machine learning methods [170]. We evaluated the accuracy of the classifier for different LSTM timestep sizes and found that the optimal timestep size for LSTM was 3 timesteps (15 seconds).

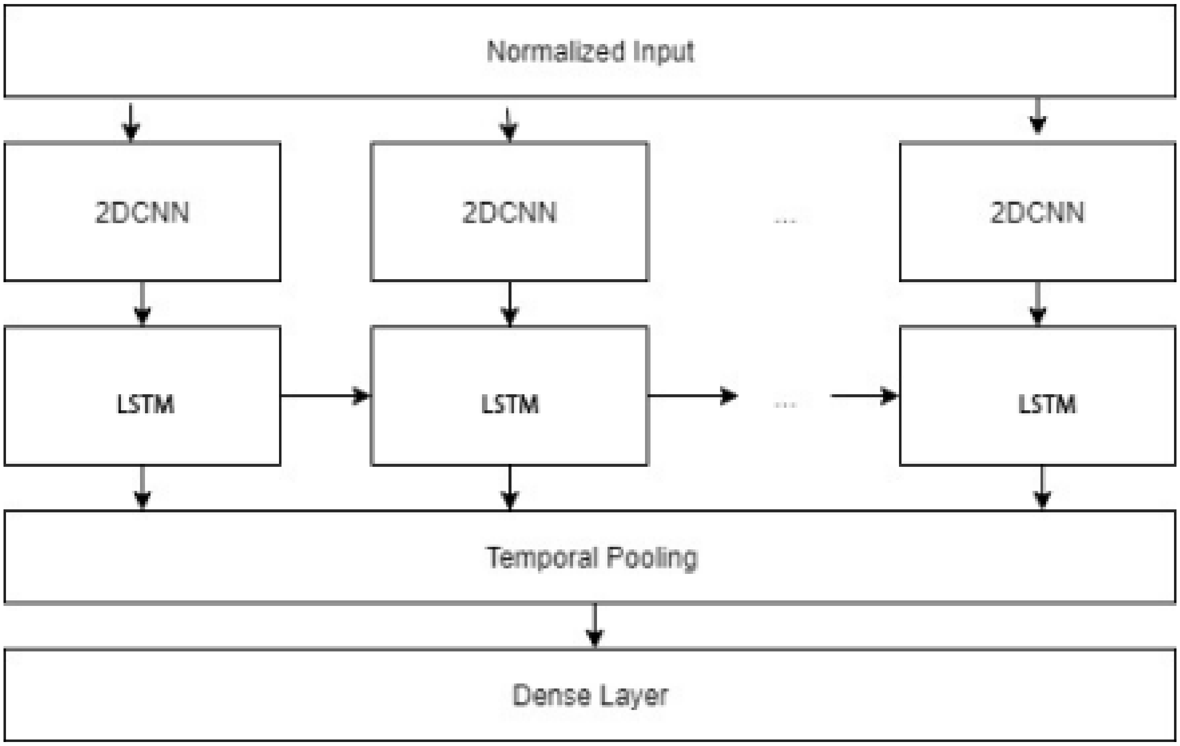


Fig. 4.10: CNN+LSTM classification model.

4.4.5 Model Evaluation

Data from the benchmark tasks, detailed in section 4.3.2, were used to train four separate models. Each model was trained using the respective benchmark task and was tested against both the Triage Task and MATB task. Each benchmark task involves a different cognitive resource, with each resource capable of being independently overloaded. We hypothesize that the different models trained on these different benchmark tasks will perform differently based on the type of task shedding that it is predicting. For instance, a model trained on the n-back task, would perform better for a task that involved the utilization of short term memory, such as remembering a string of digits for a variable period of time, whereas the visual search might perform better on a task that involved finding a salient stimuli in a cluttered visual environment. In addition to the models trained on individual benchmark tasks, a combined model was also trained on all benchmark tasks using a voting classifier, which takes the predicted probabilities for each class from each model and averages them. The predicted labels are calculated as follows:

$$Adaptive\ words\ class\ probabilities = [p_{low1},\ p_{medium1},\ p_{high1}]$$

$$Visual\ search\ class\ probabilities = [p_{low2},\ p_{medium2},\ p_{high2}]$$

$$Go\ no\ go\ class\ probabilities = [p_{low3},\ p_{medium3},\ p_{high3}]$$

$$Nback\ class\ probabilities = [p_{low4},\ p_{medium4},\ p_{high4}]$$

$$Ensemble\ class\ probabilities = [Average(p_{low}),\ Average(p_{medium}),\ Average(p_{high})]$$

$$Ensemble\ voting\ prediction = argmax[Ensemble\ class\ probabilities]$$

We tested the models on both a sequential task (Triage) and a concurrent Task (MATB) [184]. The report on our model performance in these different tasks is detailed in section 4.5.

4.5 Results

We were interested in the overall classification performance of our models as well as the performance on each of the label types because each of the ‘low’, ‘moderate’ and ‘high’ TSX labels can be used to keep system in an optimally productive state (Figure 4.13). Overall, The 5 trained models had better accuracy on the MATB task than the Triage task ($p<0.007$), where accuracy is defined by,

$$Accuracy\ =\ (Correctly\ Classified\ Instances/Total\ Instances)$$

Accuracy results from the tests are described in table 4.2.

	Triage	MATB
	Accuracy	Accuracy
Adaptive Words	60%	63%
Visual Search	60%	60%
Go no go	59%	60%
nback	61%	61%
Ensemble	61%	63%

Table 4.2: Accuracy results of the Models on the test sets.

As shown in the table, the overall accuracy of the models was around 60%, with the best overall accuracy results obtained by the Ensemble model, with 61% accuracy for Triage data and 63% accuracy for MATB data. This is promising considering that random guessing would result in 33% accuracy on the 3-class problem. Next we look at the confusion matrices and the precision, recall, and f1-scores of the models to better understand model performance in the context of task shedding events. Below we present the confusion matrices from the testing done on Triage data and MATB data using the ensemble model. (Figures 4.11 and 4.12). In these figures, 'Low' TSX is represented by 1, 'Moderate' TSX by 2 and 'High' TSX by 3.

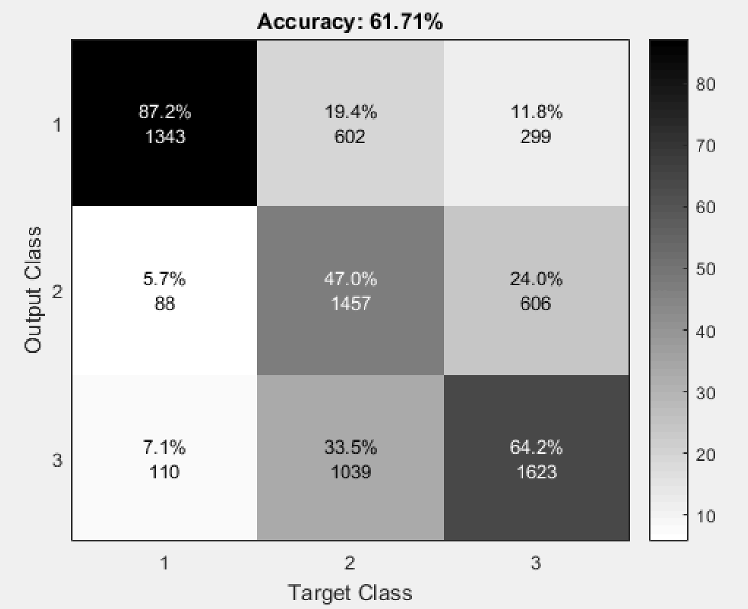


Fig. 4.11: Confusion matrix for Ensemble model tested on Triage task.

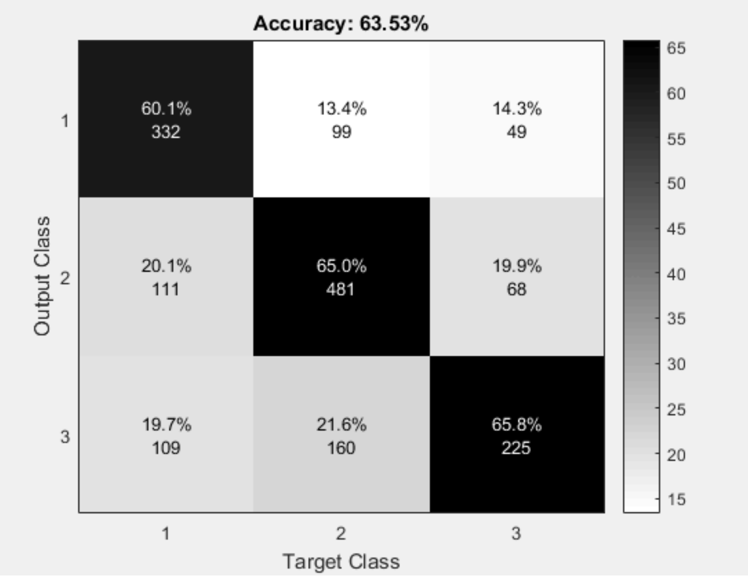


Fig. 4.12: Confusion matrix for Ensemble model tested on MATB task.

The results indicated in the confusion matrices (Figs: 4.11 and 4.12) show that the ensemble model performed well when identifying instances of ‘Low’ TSX for the Triage task. However, the model had difficulty in correctly identifying the ‘Moderate’ and ‘High’ TSX instances within the Triage task. On the MATB test, the ensemble model performed evenly well when predicting TSX values. This difference in model performance could be due to the sequential nature of the triage task, which has the user performing one single task multiple times in a row, constantly engaging one subset of cognitive resources, not having to switch to any other type of cognitive processing. As a result of this sequential presentation of the stimulus, the same brain region may have been continuously activated throughout the completion of the entire task, and the physiological response may have been so gradual that the model was unable to predict when sharp changes in TSX occurred. This may account for differences seen between 'Moderate' and 'High' TSX labels being not as easy to detect as the difference between 'Low' to 'Moderate' labels. On the contrary, the MATB, as a concurrent task, activates multiple brain regions due to task presentation happening all at once, with the user having to switch between using multiple cognitive resources in order to successfully complete each task within the battery. Though it is a possibility that the differences in accuracy between the model is due to the nature of the tasks, more ecologically valid datasets would be required to test out whether or not the presentation of the task, sequential or concurrent, show a similar effect on the model accuracy.

As mentioned above, we were also interested in per-label performance of the models. Precision and Recall values were calculated for each of the TSX labels. Precision is a measure of how many of the 'true' classifications are relevant in each class. Recall or sensitivity refers to the true prediction of each class when it is actually true. F1-score (Harmonic mean of Precision and recall) is a measure of the accuracy of each of the tests. The F1-score is more useful than accuracy when the cost of false positives and false negatives are variable. Precision, recall and f1-score values for each of the five separate models are presented in tables 4.3 - 4.5. Each table is organized by their TSX value category.

	Triage			MATB		
Model	precision	recall	f1-score	precision	recall	f1-score
Adaptive Words	0.56	0.94	0.70	0.66	0.66	0.66
Visual Search	0.6	0.85	0.70	0.71	0.59	0.64
Go no go	0.60	0.89	0.71	0.69	0.55	0.61
nback	0.58	0.90	0.70	0.68	0.62	0.65
Ensemble	0.60	0.90	0.72	0.69	0.66	0.67

Table 4.3: Precision,vrecall and f1-score of the ‘Low’ label for models on the test sets.

	Triage			MATB		
Model	precision	recall	f1-score	precision	recall	f1-score
Adaptive Words	0.66	0.53	0.59	0.70	0.68	0.69
Visual Search	0.62	0.42	0.50	0.66	0.60	0.63
Go no go	0.71	0.44	0.54	0.72	0.61	0.66
nback	0.64	0.43	0.51	0.70	0.59	0.64
Ensemble	0.68	0.45	0.54	0.73	0.60	0.66

Table 4.4: Precision, recall and f1-score of the ‘Medium’ label for models on the test sets.

	Triage			MATB		
Model	precision	recall	f1-score	precision	recall	f1-score
Adaptive Words	0.60	0.66	0.63	0.50	0.74	0.6
Visual Search	0.58	0.60	0.59	0.43	0.61	0.5
Go no go	0.50	0.67	0.57	0.37	0.74	0.49
nback	0.41	0.58	0.48	0.41	0.68	0.51
Ensemble	0.58	0.61	0.59	0.45	0.56	0.49

Table 4.5: Precision, recall and f1-score of the 'High' label for models on the test sets.

The precision for the 'low' TSX labels was higher for the MATB task classification than the Triage task (From two tailed t-test resulting in p<0.00001). Precision for 'moderate' TSX labels was also higher for the MATB task than Triage task (p<0.01). Precision for 'high' TSX labels was higher for the Triage task than MATB task (p<0.0003). Recall for the 'low' TSX labels was higher for the Triage Task compared to MATB (p<0.00005). Recall for 'moderate' TSX labels was higher for MATB task than Triage task (p<0.00002) and recall for 'high' TSX labels was also higher for MATB task than Triage task (p<0.22). Interestingly, even though the MATB data was collected from a different subject pool than the one that the models were trained on, neither Triage nor MATB test did significantly better over all the labels. Precision and recall values for the model could stand to be improved, perhaps through the use of more training data. However, for a general model such as the one proposed here, establishing criteria for what the acceptable values are for both precision and recall scores may prove difficult as these two values tend to be task dependent. Recall, in the case of the two ecologically valid examples that were trained on for this experiment, would be an important factor for 'High' TSX label as a result of the false negative cost being too high (Figure 4.13). If the task shedding event is not detected, and the human agent is currently overloaded, the machine agent has no way of knowing that it should interfere and assist the human agent within the HMT with accomplishing the current goal. This is a moment in which task shedding will become likely due to the high mental workload of the human agent within the HMT. Precision, when considered with the ecologically valid tasks that the model was tested on, is also important to consider because a high false positive rate may lead to user behavior being interrupted unnecessarily, increasing both the human agent's frustration with the system and negative affect. Frequent and unnecessary interventions by the machine agent might carry a performance hit for the HMT both by interrupting human agents in the HMT, and the overhead involved in task switching. Models using this type of data to provide feedback for an autonomous system within an HMT environment will need to therefore strike a good balance of precision and recall when it comes to the predicting 'High' TSX values.

A good prediction performance for the low TSX label is also important because, when a user is in low TSX state, the HMT system can load some tasks from overloaded users to the 'Low' TSX user (figure 4.13), thereby keeping the overall productivity rate of the HMT high. However, false positives in this case could be harmful because the system might load more tasks to somebody who is not at 'Low' value of TSX. False negatives, however, may not prove to be as harmful as their counterpart because the system will then simply ignore an idling user. Though this would cause a productivity decline within the system, it would not cause any catastrophic failures. Therefore, precision should be prioritized over recall when it comes to 'Low' TSX values.

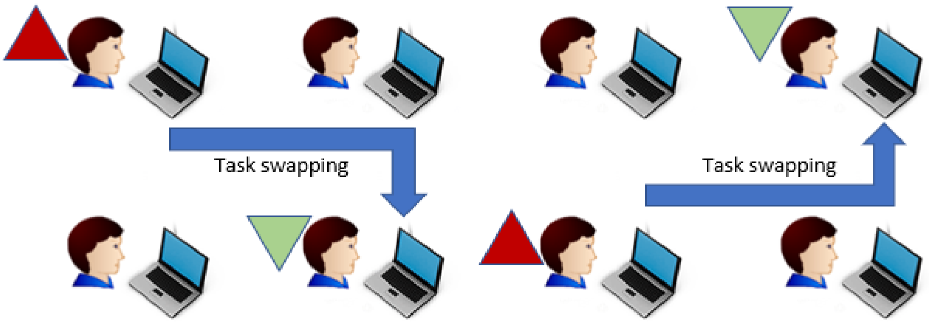


Fig. 4.13: An HMT system swapping tasks from overloaded (High TSX) users to underloaded (low TSX) users.

The ‘Moderate’ TSX is the ‘ideal’ state in HMT systems, this is the state in which the human user is productive without being at risk for being overloaded. False positives for this label mean that no action will be taken by the system. If the user is actually at a ‘Low’ TSX, then this won’t cause any major issue, but if the user is actually overloaded, or in the ‘High’ TSX category, this may lead to task shedding by the user and is not desirable. False negatives in case of ‘Moderate’ TSX label depend on the predicted state. If the user is actually at ‘Moderate’ TSX and the system predicts ‘Low’ TSX, the system might overload the user by assigning more tasks to him. On the other hand, if the user is at ‘Moderate’ TSX and system predicts ‘High’ TSX, it might unload the user putting him into idle state. Because of these aspects, precision and recall values should be considered carefully depending on both the nature of the task, and the tolerance of potential performance decrements within the HMT environment. Also of note is that, as hypothesized, the different models trained on different benchmark tasks performed differently. And the ensemble model combining all models performed better than any of the other models. This was expected due to the fact that each cognitive benchmark task on which the models were trained elicits a distinct type of cognitive load, with a signature neural correlate that is recorded up in the fNIRS data. Though further analysis is needed, and better labeled ecologically valid tasks could be used to generate even more accurate models, this could be helpful for training models used by autonomous agents within HMTs designed to accomplish tasks that rely heavily on different types of cognitive processing by the human agent. For example, a model trained on the adaptive words task could be more suitable to capture verbal working memory load, and therefore might perform better on a task in which a human agent performs a task that requires the use of working memory and visual lexical processing, such as remembering a set of instructions, than a model trained on the go nogo task, which measures one’s level of response inhibition.

4.6 Conclusion

One of the major challenges today in fNIRS literature is the difficulty of obtaining large datasets to run analysis on. In this work, we demonstrate that across-task machine learning is possible on fNIRS data with promising performance. This would indicate that researchers would be able to combine data from multiple experiments to develop models that generalize well. As mentioned above, we were able to generalize not only across tasks but also across subject pools. These have important implications for using psychphysiological data from fNIRS in real world applications and environments.

In a multi tasking situation similar to MATB, which an air force pilot faces, any operator performance dips can cause catastrophic incidents. Therefore, our method would be useful in such scenarios to prevent user error due to cognitive overload. Other such scenarios can include air traffic controller interfaces, or stock broker interfaces, where the cost of performance degradation due to cognitive overload is very high. In addition, once the operator TSX state is detected, it can be used to improve overall productivity of the HMT, by loading tasks to users who are in idle state. In this way, the same method we introduced here can be used as a productivity tool in collaborative workplaces. For example, if a group of users are doing some data entry task, the HMT can monitor users who are overloaded and swap the tasks to the underloaded users, thereby improving overall productivity. Therefore, this method can be used in both critical and non critical systems to improve system behavior.

In the above discussion, we focused on multi-user environments as an application area for this work. However, this method could be adapted to a single user as well. For example, a user doing multitasking could occasionally have some parts of their visual cognitive faculties overloaded. In this scenario, the system can offload some of the visual tasks from the user and present tasks that occupy different cognitive faculties. Such a system could improve the performance of a single user. This idea could also be extended to multi-user, multitasking environments, where during task swapping, the system could check for users who are more suitable to accept the task type based on their current cognitive faculties being used.

In this article, we have introduced a measure combining both task load and performance in order to be able to detect task shedding events within HMT environments. We have been able to classify the task shedding index into three levels: ‘Low’, ‘Moderate’ and ‘High’. We have considered two test cases which are the Triage (Cyber Analyst) Task and the MATB Task in our testing. The fNIRS enabled us to use a CNN+LSTM model designed to capture both the spatial and temporal nature of the brain activity. As stated in previous sections, this method of classification shows great promise in both the domains of HCI and Human Factors, though it also has more direct practical implications for human machine teaming and multitasking environments where various cognitive resources of a user are occupied at different times. Specifically, our across task performance shows promise for training models on simpler tasks and being able to generalize to compound tasks. In conclusion, our study demonstrates that we can obtain a generalizable classifier that performs well across multiple subject pools as well as across multiple tasks, thus enabling adaptive Human Machine Teaming across diverse real-world settings.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

We have presented machine learning methods that can be used to improve upon the state of the art in current fNIRS research. Our first approach involves segmenting the fNIRS data into 'Regions of Interest' based on the functional brain regions covered by the fNIRS channels. Our approach enables us to capture the spatial structure of the human brain as well as achieving dimensionality reduction. This method improves upon the valence classification F1 score achieved by Koelstra et al.[185] on the same music video stimuli. The next approach that we propose is to use a Convolutional Neural Network to capture the spatial dependencies in the fNIRS data. This approach has enabled us to combine the oxygenated and deoxygenated blood flow data into a single data analysis method. Here we feed in the oxy and deoxy blood flow data as two 5X10 matrices that are treated similarly to an image classification problem. HCI researchers have struggled with feature extraction when it comes to fNIRS data due to the lack of understanding that we have about the dynamics of blood flow in the brain. By applying CNNs to fNIRS data, we circumvent the need for this type of trial and error feature extraction approach. Another important aspect introduced in this work is the use of Long Short-Term Memory in order to capture the temporal behavior of the fNIRS data. We experimented with different time step sizes to find out the optimal time step size for the LSTM layer. We found that this optimal timestep size changes according to the application domain. We show that using the combination of CNN and LSTM, we can achieve higher classification performance in across subject datasets which has been a challenge in the emotion classification domain. With ability to do across subject training, it become possible to collect more training data which can be expected to improve the classification performance further. After developing these methods, we applied our CNN+LSTM model to another application area in HCI. Our goal was to predict the human performance degradations due to high task load using the brain activity data alone. This type of claasification has implications for Human Machine Teaming (HMT) tasks where multiple humans work with multiple computers on concurrent tasks. The goal of such a system is to improve the productivity of the overall system by responding to the state of each individual user. In this work, we were able to use our CNN+LSTM model to do classification across tasks. This across task classification has been by far the most challenging problem in this domain and our results show that our model can be generalized to real world applications.

5.2 Future Work

Brain activity is a highly objective measure of user state. And provides a lot of opportunities and challenges to future research.

Analyzing the Effect of Individual Differences on Brain Activity Data

In our current research, the fNIRS datasets we collected tend to be from college students that are college aged and have similar education levels. It would be interesting to study the robustness of our method when applied to homogeneous and non homogeneous demographic populations. Such a study will be a key factor in making this research applicable on a mass scale. Of course, the main challenge in such an undertaking would be that it would require a much larger subject pool.

Event Detection Algorithms for fNIRS Data

In this research we looked at capturing the macro time-scale dynamics of the fNIRS data. However, there is potential for researchers to investigate transient, event level analysis of the fNIRS data. Some examples might involve anomaly detection to detect spikes in cognitive load, using Markov models and reinforcement learning to detect sudden state changes.

Ordering of fNIRS Channels for Optimal Classification Performance

In this dissertation, we focused on capturing the spatial location (along with temporal behavior) of fNIRS channels and showed that it can improve classification performance. Even though the spatial ordering is useful for Convolutional Neural Networks, there can be more optimal channel orderings that can provide further classification performance improvement. Giving more weight to some channels over others could also be an approach to improve classifier performance.

In closing, fNIRS provides us an objective glimpse into the workings of our brain and its function. As the amount of computing devices that surround us grows at an exponential scale, those devices will constantly collect physiological and brain activity data from their users. Therefore, we will have access to an ever-growing repository of brain activity data that will make it imperative that we create algorithms that scale well and can accommodate users with varied demographics.

REFERENCES

[1] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis ;using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan 2012.

[2] J.A. Russell, “A circumplex model of affect,” *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] Kurtulus Izzetoglu, Scott Bunce, Banu Onaral, Kambiz Pourrezaei, and Britton Chance, “Functional optical brain imaging using near-infrared during cognitive tasks,” *International Journal of HumanâĂŞComputer Interaction*, vol. 17, no. 2, pp. 211–227, 2004.

[4] Robert Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[5] Raja Parasuraman and Matthew Rizzo, *Neuroergonomics: The brain at work*, Oxford University Press, 2008.

[6] Tiziano Colibazzi, Jonathan Posner, Zhishun Wang, Daniel Gorman, Andrew Gerber, Shan Yu, Hongtu Zhu, Alayar Kangarlu, Yunsuo Duan, James A Russell, et al., “Neural systems subserving valence and arousal during the experience of induced emotions.,” *Emotion*, vol. 10, no. 3, pp. 377, 2010.

[7] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven,

"Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, space, and environmental medicine*, vol. 78, no. 5, pp. B231–B244, 2007.

[8] David Grimes, Desney Tan, Scott Hudson, Pradeep Shenoy, and Rajesh Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," January 2008.

[9] Johnny Chung Lee and Desney S Tan, "Using a low-cost electroencephalograph for task classification in hci research," in *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 2006, pp. 81–90.

[10] Audrey Girouard, Erin Treacy Solovey, Leanne M. Hirshfield, Krysta Chauncey, Angelo Sassaroli, Sergio Fantini, and Robert J. K. Jacob, *Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements*, pp. 440–452, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[11] LM Hirshfield, A Girouard, ET Solovey, RJK Jacob, A Sassaroli, Y Tong, and S Fantini, "Human-computer interaction and brain measurement using functional near-infrared spectroscopy," in *Proceedings of the ACM UIST'07 Symposium on User Interface Software and Technology*. ACM Press, 2007.

[12] Leanne M. Hirshfield, Krysta Chauncey, Rebecca Gulotta, Audrey Girouard, Erin T. Solovey, Robert J. K. Jacob, Angelo Sassaroli, and Sergio Fantini, *Combining Electroencephalograph and Functional Near Infrared Spectroscopy to Explore Users' Mental Workload*, pp. 239–247, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[13] Leanne M. Hirshfield, Rebecca Gulotta, Stuart Hirshfield, Sam Hincks, Matthew Russell, Rachel Ward, Tom Williams, and Robert Jacob, "This is your brain on interfaces: Enhancing usability testing with functional near-infrared spectroscopy," in *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, New York, NY, USA, 2011, CHI '11, pp. 373–382, ACM.

[14] Margaret M Bradley and Peter J Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[15] Johnny Chung Lee and Desney S. Tan, "Using a low-cost electroencephalograph for task classification in hci research," in *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2006, UIST '06, pp. 81–90, ACM.

[16] B Chance, Z Zhuang, Chu UnAh, C Alter, and L Lipton, "Cognition-activated low-frequency modulation of light absorption in human brain.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 8, pp. 3770–3774, 1993.

[17] Il-Young Son, Markus Guhe, Wayne D. Gray, Birsen Yazici, and Michael J. Schoelles, "Human performance assessment using fnir," 2005, vol. 5797, pp. 158–169.

[18] Leanne M Hirshfield, Philip Bobko, Alex Barelka, Stuart H Hirshfield, Mathew T Farrington, Spencer Gulbronson, and Diane 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, pp. 2, 2014.

[19] Thomas Ehring, Brunna Tuschen-Caffier, Jewgenija Schnuelle, Silke Fischer, and James J. Gross, "Emotion Regulation and Vulnerability to Depression: Spontaneous Versus Instructed Use of Emotion Suppression and Reappraisal," *EMOTION*, vol. 10, no. 4, pp. 563–572, 2010.

[20] Renata M Heilman, Liviu G Crişan, Daniel Houser, Mircea Miclea, and Andrei C Miu, "Emotion regulation and decision making under risk and uncertainty.," *Emotion*, vol. 10, no. 2, pp. 257, 2010.

[21] Robert Plutchik, "A general psychoevolutionary theory of emotion," *Theories of emotion*, vol. 1, pp. 3–31, 1980.

[22] Richard J Davidson, Daren C Jackson, and Ned H Kalin, "Emotion, plasticity, context, and regulation: perspectives from affective neuroscience.," *Psychological bulletin*, vol. 126, no. 6, pp. 890, 2000.

[23] Richard D Lane and Lynn Nadel, *Cognitive neuroscience of emotion*, Oxford University Press, 1999.

[24] Sarah D Pressman and Sheldon Cohen, "Does positive affect influence health?," *Psychological bulletin*, vol. 131, no. 6, pp. 925, 2005.

[25] Peter Salovey, Alexander J Rothman, Jerusha B Detweiler, and Wayne T Steward, "Emotional states and physical health.," *American psychologist*, vol. 55, no. 1, pp. 110, 2000.

[26] Douglas S Massey, "A brief history of human society: The origin and role of emotion in social life," *American Sociological Review*, vol. 67, no. 1, pp. 1, 2002.

[27] Klaus R Scherer, "What are emotions? and how can they be measured?," *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[28] Andrew Ortony, Gerald L Clore, and Allan Collins, *The cognitive structure of emotions*, Cambridge university press, 1990.

[29] Rosalind W Picard and Roalind Picard, "A ective computing. vol. 252," 1997.

[30] P.J. Lang, "The emotion probe: Studies of motivation and attention," *American psychologist*, vol. 50, pp. 372–372, 1995.

[31] Gene Ball and Jack Breese, "Modeling the emotional state of computer users," in *Proceedings of the Workshop on Personality and Emotion in User Modelling*, 1999.

[32] Wauter Bosma and Elisabeth André, "Exploiting emotions to disambiguate dialogue acts," in *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, 2004, pp. 85–92.

[33] Ann M Kring, Lisa Feldman Barrett, and David E Gard, "On the broad applicability of the affective circumplex: representations of affective knowledge among schizophrenia patients," *Psychological Science*, vol. 14, no. 3, pp. 207–214, 2003.

[34] Lisa Feldman Barrett and James A Russell, "Independence and bipolarity in the structure of current affect.," *Journal of personality and social psychology*, vol. 74, no. 4, pp. 967, 1998.

[35] James A Russell, "Culture and the categorization of emotions.," *Psychological bulletin*, vol. 110, no. 3, pp. 426, 1991.

[36] Paul Ekman, Robert W Levenson, and Wallace V Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[37] R. Plutchik, "The Nature of Emotions," *American Scientist*, vol. 89, pp. 344, July 2001.

[38] Jonathan Posner, James A Russell, and Bradley S Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[39] David Watson, Lee Anna Clark, and Auke Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063, 1988.

[40] Charles Darwin, *The expression of the emotions in man and animals*, New York. Appleton and Co.,, 1916., http://www.biodiversitylibrary.org/bibliography/4820.

[41] Serena-Lynn Brown and Gary E Schwartz, "Relationships between facial electromyography and subjective experience during affective imagery," *Biological Psychology*, vol. 11, no. 1, pp. 49–62, 1980.

[42] Gary E Schwartz, Paul L Fair, Patricia Salt, Michel R Mandel, and Gerald L Klerman, “Facial muscle patterning to affective imagery in depressed and nondepressed subjects,” *Science*, vol. 192, no. 4238, pp. 489–491, 1976.

[43] Alan J Fridlund, Gary E Schwartz, and Stephen C Fowler, “Pattern recognition of self-reported emotional state from multiple-site facial emg activity during affective imagery,” *Psychophysiology*, vol. 21, no. 6, pp. 622–637, 1984.

[44] Ulf Dimberg, “Facial electromyography and the experience of emotion.,” *Journal of Psychophysiology*, 1988.

[45] Jane E Warren, Disa A Sauter, Frank Eisner, Jade Wiland, M Alexander Dresner, Richard JS Wise, Stuart Rosen, and Sophie K Scott, “Positive emotions preferentially engage an auditory-motor ’mirror’ system,” *Journal of Neuroscience*, vol. 26, no. 50, pp. 13067–13075, 2006.

[46] Walter B Cannon, “The james-lange theory of emotions: A critical examination and an alternative theory,” *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[47] Stanley Schachter and Jerome Singer, “Cognitive, social, and physiological determinants of emotional state.,” *Psychological review*, vol. 69, no. 5, pp. 379, 1962.

[48] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm, “Looking at pictures: Affective, facial, visceral, and behavioral reactions,” *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[49] Rachel Bailey, Kevin Wise, and Paul Bolls, “How avatar customizability affects children’s arousal and subjective presence during junk food–sponsored online video games,” *CyberPsychology & Behavior*, vol. 12, no. 3, pp. 277–283, 2009.

[50] Maurizio Codispoti, Paola Surcinelli, and Bruno Baldaro, "Watching emotional movies: Affective reactions and gender differences," *International Journal of Psychophysiology*, vol. 69, no. 2, pp. 90–95, 2008.

[51] JT Cacioppo, "Berntson. gg, & klein. dj 1992. what is an emotion? the role of somatovisceral afference, with special emphasis on somatovisceral" illusions," *Emotion and Social Behavior*, pp. 63–98.

[52] Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager, "Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies," *Neuroimage*, vol. 42, no. 2, pp. 998–1031, 2008.

[53] Mikko Viinikainen, Iiro P Jääskeläinen, Yuri Alexandrov, Marja H Balk, Taina Autti, and Mikko Sams, "Nonlinear relationship between emotional valence and brain activity: evidence of separate negative and positive valence dimensions," *Human brain mapping*, vol. 31, no. 7, pp. 1030–1040, 2010.

[54] Elvira Brattico, Vinoo Alluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Katriina Nieminen, and Mari Tervaniemi, "A functional mri study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, pp. 308, 2011.

[55] Luciano Fadiga, Laila Craighero, and Alessandro D'Ausilio, "Broca's area in language, action, and music," *Annals of the New York Academy of Sciences*, vol. 1169, no. 1, pp. 448–458, 2009.

[56] Jose Leon-Carrion, Jes$\acute{u}$s Damas, Kurtulus Izzetoglu, Kambiz Pourrezai, Juan Francisco Martin-Rodriguez, Juan Manuel Barroso y Martin, and Maria Rosario Dominguez-Morales, "Differential time course and intensity of {PFC} activation for men and women in response to emotional stimuli: A functional near-infrared spectroscopy (fnirs) study," *Neuroscience Letters*, vol. 403, no. 1-2, pp. 90 – 95, 2006.

[57] Achala H Rodrigo, Hasan Ayaz, and Anthony C Ruocco, "Examining the neural correlates of incidental facial emotion encoding within the prefrontal cortex using functional near-infrared spectroscopy," in *International Conference on Augmented Cognition*. Springer, 2016, pp. 102–112.

[58] Michela Balconi, Elisabetta Grippa, and Maria Elide Vanutelli, "What hemodynamic (fnirs), electrophysiological (eeg) and autonomic integrated measures can tell us about emotional processing," *Brain and Cognition*, vol. 95, pp. 67 – 76, 2015.

[59] Nathan W Churchill, Grigori Yourganov, and Stephen C Strother, "Comparing within-subject classification and regularization methods in fmri for large and small sample sizes," *Human brain mapping*, vol. 35, no. 9, pp. 4499–4517, 2014.

[60] Robert CA Bendall, Peter Eachus, and Catherine Thompson, "A brief review of research using near-infrared spectroscopy to measure activation of the prefrontal cortex during emotional processing: the importance of experimental design," *Frontiers in human neuroscience*, vol. 10, pp. 529, 2016.

[61] Yoko Hoshi and Mamoru Tamura, "Near-infrared optical detection of sequential brain activation in the prefrontal cortex during mental tasks," *NeuroImage*, vol. 5, no. 4, pp. 292–297, 1997.

[62] Evelyn Glotzbach, Andreas Mühlberger, Kathrin Gschwendtner, Andreas J Fallgatter, Paul Pauli, and Martin J Herrmann, "Prefrontal brain activation during emotional processing: a functional near infrared spectroscopy study (fnirs)," *The open neuroimaging journal*, vol. 5, pp. 33, 2011.

[63] Shota Nishitani, Kazuyuki Shinohara, et al., "Nirs as a tool for assaying emotional function in the prefrontal cortex," *Frontiers in human neuroscience*, vol. 7, pp. 770, 2013.

[64] Margaret Bradley and Peter J Lang, *The International affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings*, NIMH Center for the Study of Emotion and Attention, 1999.

[65] Thomas Baumgartner, Michaela Esslen, and Lutz Jäncke, "From emotion perception to emotion experience: Emotions evoked by pictures and classical music," *International journal of psychophysiology*, vol. 60, no. 1, pp. 34–43, 2006.

[66] Hitachi Corp., "ETG-4000: Hitachi Healthcare," 2016, [Online; accessed 19-April-2018].

[67] Xu Cui, Signe Bray, and Allan 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.

[68] Bruce Crosson, Anastasia Ford, Keith M McGregor, Marcus Meinzer, Sergey Cheshkov, Xiufeng Li, Delaina Walker-Batson, and Richard W Briggs, "Functional imaging and related techniques: An introduction for rehabilitation researchers," *Journal of rehabilitation research and development*, vol. 47, no. 2, pp. vii, 2010.

[69] Peter Kuppens and Eddie MW Tong, "An appraisal account of individual differences in emotional experience," *Social and Personality Psychology Compass*, vol. 4, no. 12, pp. 1138–1150, 2010.

[70] Christie Napa Scollon, Sharon Koh, and Evelyn WM Au, "Cultural differences in the subjective experience of emotion: When and why they occur," *Social and Personality Psychology Compass*, vol. 5, no. 11, pp. 853–864, 2011.

[71] Nader Karamzadeh, Franck Amyot, Kimbra Kenney, Afrouz Anderson, Fatima Chowdhry, Hadis Dashtestani, Eric M Wassermann, Victor Chernomordik, Claude Boccara, Edward Wegman, et al., "A machine learning approach to identify functional biomarkers in human prefrontal cortex for individuals with traumatic brain injury using functional near-infrared spectroscopy," *Brain and behavior*, vol. 6, no. 11, 2016.

[72] Patrik Vuilleumier and Jon Driver, "Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1481, pp. 837–855, 2007.

[73] Giorgia Salvatori, Luca Bulf, Sergio Fonda, and Luigi Rovati, "Combining near-infrared spectroscopy and electroencephalography to monitor brain function," in *Instrumentation and Measurement Technology Conference, 2006. IMTC 2006. Proceedings of the IEEE*. IEEE, 2006, pp. 32–36.

[74] Arman Savran, Koray Ciftci, Guillaume Chanel, Javier Mota, Luong Hong Viet, Blent Sankur, Lale Akarun, Alice Caplier, and Michele Rombaut, *Emotion Detection in the Loop from Brain Signals and Facial Images*, Proceedings of the eNTERFACE 2006 Workshop. 2006, ID: unige:47926; Final Project Report.

[75] Nadège Roche-Labarbe, Boubker Zaaimi, Patrick Berquin, Astrid Nehlig, Reinhard Grebe, and Fabrice Wallois, "Nirs-measured oxy-and deoxyhemoglobin changes associated with eeg spike-and-wave discharges in children," *Epilepsia*, vol. 49, no. 11, pp. 1871–1880, 2008.

[76] Yanjia Sun, Hasan Ayaz, and Ali N Akansu, "Neural correlates of affective context in facial expression analysis: a simultaneous eeg-fnirs study," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*. IEEE, 2015, pp. 820–824.

[77] M. M. Bradley and P. J. Lang, "International affective digitized sounds (iads): Stimuli, instruction manual and affective ratings," *Tech. Rep. No. B-2*, 2007.

[78] Roland Neumann, Markus Hess, Stefan M. Schulz, and Georg W. Alpers, "Automatic behavioural responses to valence: Evidence that facial action is facilitated by evaluative processing," *Cognition and Emotion*, vol. 19, no. 4, pp. 499–513, 2005.

[79] Junwei Han, Xiang Ji, Xintao Hu, Lei Guo, and Tianming Liu, “Arousal recognition using audio-visual features and fmri-based brain response,” *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 337–347, 2015.

[80] Mohammad Soleymani, Maja Pantic, and Thierry Pun, “Multimodal emotion recognition in response to videos,” *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2012.

[81] Robert Jenke, Angelika Peer, and Martin Buss, “Feature extraction and selection for emotion recognition from eeg,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[82] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi, “Real-time movie-induced discrete emotion recognition from eeg signals,” *IEEE Transactions on Affective Computing*, 2017.

[83] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis, “Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis,” *IEEE Transactions on affective computing*, vol. 1, no. 2, pp. 81–97, 2010.

[84] Shangfei Wang, Shiyu Chen, and Qiang Ji, “Content-based video emotion tagging augmented by users’ multiple physiological responses,” *IEEE Transactions on Affective Computing*, 2017.

[85] Leanne M. Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert J.K. Jacob, Angelo Sassaroli, and Sergio Fantini, “Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2009, CHI ’09, pp. 2185–2194, ACM.

[86] Ute Kreplin and Stephen 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, pp. 879, 2013.

[87] Z. Khalili and M. H. Moradi, “Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of eeg,” in *Proceedings of the 2009 International Joint Conference on Neural Networks*, Piscataway, NJ, USA, 2009, IJCNN’09, pp. 1920–1924, IEEE Press.

[88] Saman Sarraf and Ghassem Tofighi, “Classification of alzheimer’s disease using fmri data and deep learning convolutional neural networks,” *arXiv preprint arXiv:1603.08631*, 2016.

[89] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball, “Deep learning with convolutional neural networks for eeg decoding and visualization,” *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[90] H. Cecotti and A. Graser, “Convolutional neural networks for p300 detection with application to brain-computer interfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 433–445, March 2011.

[91] Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al., “Function in the human connectome: task-fmri and individual differences in behavior,” *Neuroimage*, vol. 80, pp. 169–189, 2013.

[92] Alexander Drobyshevsky, Stephen B Baumann, and Walter Schneider, “A rapid fmri task battery for mapping of visual, motor, cognitive, and emotional function,” *Neuroimage*, vol. 31, no. 2, pp. 732–744, 2006.

[93] Mauricio R Delgado, Leigh E Nystrom, Catherine Fissell, DC Noll, and Julie A Fiez, "Tracking the hemodynamic responses to reward and punishment in the striatum," *Journal of neurophysiology*, vol. 84, no. 6, pp. 3072–3077, 2000.

[94] Fulvia Castelli, Francesca Happé, Uta Frith, and Chris Frith, "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns," *Neuroimage*, vol. 12, no. 3, pp. 314–325, 2000.

[95] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[96] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert MÃijller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387 – 399, 2011, Multivariate Decoding and Brain Reading.

[97] Etienne Combrisson and Karim Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of neuroscience methods*, vol. 250, pp. 126–136, 2015.

[98] Ziheng Wang, Ryan M Hope, Zuoguan Wang, Qiang Ji, and Wayne D Gray, "Cross-subject workload classification with a hierarchical bayes model," *NeuroImage*, vol. 59, no. 1, pp. 64–69, 2012.

[99] Xuerui Wang, Rebecca Hutchinson, and Tom M Mitchell, "Training fmri classifiers to detect cognitive states across multiple human subjects," in *Advances in neural information processing systems*, 2004, pp. 709–716.

[100] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield, "Building predictive models of emotion with functional near-infrared spectroscopy," *International Journal of Human-Computer Studies*, vol. 110, no. Supplement C, pp. 75 – 85, 2018.

[101] Norberto Eiji Nawa and Hiroshi Ando, "Classification of self-driven mental tasks from whole-brain activity patterns," *PloS one*, vol. 9, no. 5, pp. e97296, 2014.

[102] R.F. Potter and P.D. Bolls, *Psychophysiological Measurement and Meaning: Cognitive and Emotional Processing of Media*, Communication (Routledge Paperback). Routledge, 2012.

[103] Walter B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[104] Lisa Feldman Barrett and Tor D. Wager, "The structure of emotion," *Current Directions in Psychological Science*, vol. 15, no. 2, pp. 79–83, 2006.

[105] Mark Costa and Sarah Bratt, "Truthiness: challenges associated with employing machine learning on neurophysiological sensor data," in *International Conference on Augmented Cognition*. Springer, 2016, pp. 159–164.

[106] Danushka Bandara, Stephen Song, Leanne Hirshfield, and Senem Velipasalar, *A More Complete Picture of Emotion Using Electrocardiogram and Electrodermal Activity to Complement Cognitive Data*, pp. 287–298, Springer International Publishing, Cham, 2016.

[107] Richard B Buxton, *Introduction to functional magnetic resonance imaging: principles and techniques*, Cambridge university press, 2009.

[108] Alex R Wade, "The negative bold signal unmasked," *Neuron*, vol. 36, no. 6, pp. 993–995, 2002.

[109] Hoi-Chung Leung, Hwamee Oh, Jamie Ferri, and Yuji Yi, "Load response functions in the human spatial working memory circuit during location memory updating," *Neuroimage*, vol. 35, no. 1, pp. 368–377, 2007.

[110] R Parasuraman and D Caggiano, "Neural and genetic assays of human mental workload," *Quantifying human information processing*, pp. 123–149, 2005.

[111] Jonathan D Cohen, William M Perlstein, Todd S Braver, Leigh E Nystrom, Douglas C Noll, John Jonides, and Edward E Smith, "Temporal dynamics of brain activation during a working memory task," *Nature*, vol. 386, no. 6625, pp. 604, 1997.

[112] Ajit Devaraj, "Signal processing for functional near-infrared neuroimaging," 2005.

[113] Angelo Sassaroli, Feng Zheng, Leanne M Hirshfield, Audrey Girouard, Erin Treacy Solovey, Robert JK Jacob, and Sergio Fantini, "Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy," *Journal of Innovative Optical Health Sciences*, vol. 1, no. 02, pp. 227–237, 2008.

[114] Yoko Hoshi, Brian H Tsou, Vincent A Billock, Masato Tanosaki, Yoshinobu Iguchi, Miho Shimada, Toshikazu Shinba, Yoshifumi Yamada, and Ichiro Oda, "Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks," *Neuroimage*, vol. 20, no. 3, pp. 1493–1504, 2003.

[115] Scott C Bunce, Ajit Devaraj, Meltem Izzetoglu, Banu Onaral, and Kambiz Pourrezaei, "Detecting deception in the brain: A functional near-infrared spectroscopy study of neural correlates of intentional deception," in *Nondestructive Detection and Measurement for Homeland Security III*. International Society for Optics and Photonics, 2005, vol. 5769, pp. 24–33.

[116] Jeffrey K Thompson, Matthew R Peterson, and Ralph D Freeman, "Single-neuron activity and tissue oxygenation in the cerebral cortex," *Science*, vol. 299, no. 5609, pp. 1070–1072, 2003.

[117] Scott C Bunce, Meltem Izzetoglu, Kurtulus Izzetoglu, Banu Onaral, and Kambiz Pourrezaei, "Functional near-infrared spectroscopy," *IEEE engineering in medicine and biology magazine*, vol. 25, no. 4, pp. 54–62, 2006.

[118] P. W. Mirowski, Y. LeCun, D. Madhavan, and R. Kuzniecky, "Comparing svm and convolutional networks for epileptic seizure prediction from intracranial eeg," in *2008 IEEE Workshop on Machine Learning for Signal Processing*, Oct 2008, pp. 244–249.

[119] Johannes Hennrich, Christian Herff, Dominic Heger, and Tanja Schultz, “Investigating deep learning for fnirs based bci.,” in *EMBC*, 2015, pp. 2844–2847.

[120] Gauvain Huve, Kazuhiko Takahashi, and Masafumi Hashimoto, “Brain activity recognition with a wearable fnirs using neural networks,” *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1573–1578, 2017.

[121] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[122] Michiel Hermans and Benjamin Schrauwen, “Training and analysing deep recurrent neural networks,” in *Advances in neural information processing systems*, 2013, pp. 190–198.

[123] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi, “Emotion recognition based on eeg using lstm recurrent neural network,” *Emotion*, vol. 8, no. 10, 2017.

[124] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic, “Analysis of eeg signals and facial expressions for continuous emotion detection,” *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.

[125] Md Musaddaqul Hasib, Tapsya Nayak, and Yufei Huang, “A hierarchical lstm model with attention for modeling eeg non-stationarity for human decision prediction,” in *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 2018, pp. 104–107.

[126] Christopher Olah, “Understanding lstm networks,” [Online; accessed 11-May-2018].

[127] Eleni Tsironi, Pablo Barros, Cornelius Weber, and Stefan Wermter, “An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition,” *Neurocomputing*, vol. 268, pp. 76–86, 2017.

[128] Dominic Heger, Reinhard Mutter, Christian Herff, Felix Putze, and Tanja Schultz, “Continuous recognition of affective states by functional near infrared spectroscopy signals,” in

Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013, pp. 832–837.

[129] P Lang and Margaret M Bradley, "The international affective picture system (iaps) in the study of emotion and attention," *Handbook of emotion elicitation and assessment*, vol. 29, 2007.

[130] Margaret M Bradley and Peter J Lang, "The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual," *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[131] V. Rozgić, S. N. Vitaladevuni, and R. Prasad, "Robust eeg emotion classification using segment level decision fusion," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 1286–1290.

[132] Xiang Li, Peng Zhang, Dawei Song, Guangliang Yu, Yuexian Hou, and Bin Hu, "Eeg based emotion identification using unsupervised deep feature learning," 2015.

[133] Itsara Wichakam and Peerapon Vateekul, "An evaluation of feature extraction in eeg-based emotion prediction with support vector machines," in *Computer science and software engineering (JCSSE), 2014 11th international joint conference on*. IEEE, 2014, pp. 106–110.

[134] Henry Candra, Mitchell Yuwono, Rifai Chai, Ardi Handojoseno, Irraivan Elamvazuthi, Hung T Nguyen, and Steven Su, "Investigation of window size in classification of eeg-emotion signal with wavelet entropy and support vector machine," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 7250–7253.

[135] Shota Nishitani, Hirokazu Doi, Atsuko Koyama, and Kazuyuki Shinohara, "Differential prefrontal response to infant facial emotions in mothers compared with non-mothers," *Neuroscience Research*, vol. 70, no. 2, pp. 183 – 188, 2011.

[136] Minah Suh, Sonya Bahar, Ashesh D Mehta, and Theodore H Schwartz, "Blood volume and hemoglobin oxygenation response following electrical stimulation of human cortex," *Neuroimage*, vol. 31, no. 1, pp. 66–75, 2006.

[137] Michael Egmont-Petersen, Jan L Talmon, Arie Hasman, and Anton W Ambergen, "Assessing the importance of features for multi-layer perceptrons," *Neural networks*, vol. 11, no. 4, pp. 623–635, 1998.

[138] Christopher D Wickens, "Processing resources and attention," *Multiple-task performance*, vol. 1991, pp. 3–34, 1991.

[139] Christopher D Wickens, Amy Santamaria, and Angelia Sebok, "A computational model of task overload management and task switching," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 2013, vol. 57, pp. 763–767.

[140] Ruben Strenzke, Johann Uhrmann, Andreas Benzler, Felix Maiwald, Andreas Rauschert, and Axel Schulte, "Managing cockpit crew excess task load in military manned-unmanned teaming missions by dual-mode cognitive automation approaches," in *AIAA Guidance, Navigation, and Control Conference*, 2011, p. 6237.

[141] James P Bliss, John W Harden, and H Charles Dischinger Jr, "Task shedding and control performance as a function of perceived automation reliability and time pressure," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 2013, vol. 57, pp. 635–639.

[142] Ruben Strenzke and Axel Schulte, "Modeling the human operatorâĂŹs cognitive process to enable assistant system decisions," *GAPRec 2011*, p. 38, 2011.

[143] Raja Parasuraman and Peter A Hancock, "Adaptive control of mental workload.," *Stress, workload, and fatigue*, 2001.

[144] Dario D Salvucci and Niels A Taatgen, *The multitasking mind*, Oxford University Press, 2010.

[145] Yili Liu, "Queueing network modeling of elementary mental processes.," *Psychological Review*, vol. 103, no. 1, pp. 116, 1996.

[146] David E Meyer and David E Kieras, "A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms.," *Psychological review*, vol. 104, no. 1, pp. 3, 1997.

[147] Harold E Pashler and Stuart Sutherland, *The psychology of attention*, vol. 15, MIT press Cambridge, MA, 1998.

[148] Duncan P Brumby, Dario D Salvucci, and Andrew Howes, "Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2009, pp. 1629–1638.

[149] Christian P Janssen and Duncan P Brumby, "Strategic adaptation to performance objectives in a dual-task setting," *Cognitive science*, vol. 34, no. 8, pp. 1548–1560, 2010.

[150] Erik J Sirevaag, Arthur F Kramer, Mark Reisweber, Christopher D Wickens, David L Strayer, and James F Grenell, "Assessment of pilot performance and mental workload in rotary wing aircraft," *Ergonomics*, vol. 36, no. 9, pp. 1121–1140, 1993.

[151] Ewart de Visser and Raja Parasuraman, "Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload," *Journal of Cognitive Engineering and Decision Making*, vol. 5, no. 2, pp. 209–231, 2011.

[152] Jessie YC Chen and Michael J Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2014.

[153] Arno Villringer and Britton Chance, “Non-invasive optical spectroscopy and imaging of human brain function,” *Trends in neurosciences*, vol. 20, no. 10, pp. 435–442, 1997.

[154] Marco Ferrari and Valentina 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, 2012.

[155] Xu Cui, Signe Bray, Daniel M Bryant, Gary H Glover, and Allan L Reiss, “A quantitative comparison of nirs and fmri across multiple cognitive tasks,” *Neuroimage*, vol. 54, no. 4, pp. 2808–2821, 2011.

[156] Leanne M Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert JK Jacob, Angelo Sassaroli, and Sergio Fantini, “Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 2185–2194.

[157] Christian Herff, Dominic Heger, Ole Fortmann, Johannes Hennrich, Felix Putze, and Tanja Schultz, “Mental workload during n-back taskâĂŤquantified in the prefrontal cortex using fnirs,” *Frontiers in human neuroscience*, vol. 7, pp. 935, 2014.

[158] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler, “Classifying driver workload using physiological and driving performance data: two field studies,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 4057–4066.

[159] Michael E Smith, Alan Gevins, Halle Brown, Arati Karnik, and Robert Du, “Monitoring task loading with multivariate eeg measures during complex forms of human-computer interaction,” *Human Factors*, vol. 43, no. 3, pp. 366–380, 2001.

[160] Evan M Peck, Daniel Afergan, Beste F Yuksel, Francine Lalooses, and Robert JK Jacob, “Using fnirs to measure mental workload in the real world,” in *Advances in physiological computing*, pp. 117–139. Springer, 2014.

[161] Erin Treacy Solovey, Daniel Afergan, Evan M Peck, Samuel W Hincks, and Robert JK Jacob, “Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fnirs,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 21, no. 6, pp. 35, 2015.

[162] Shirley Coyle, Tomás Ward, Charles Markham, and Gary McDarby, “On the suitability of near-infrared (nir) systems for next-generation brain–computer interfaces,” *Physiological measurement*, vol. 25, no. 4, pp. 815, 2004.

[163] Tom Kontogiannis and Zoe Kossiavelou, “Stress and team performance: principles and challenges for intelligent decision aids,” *Safety science*, vol. 33, no. 3, pp. 103–128, 1999.

[164] Xu Cui, Signe Bray, and Allan L Reiss, “Speeded near infrared spectroscopy (nirs) response detection,” *PLoS one*, vol. 5, no. 11, pp. e15474, 2010.

[165] Justin Chan, Sarah Power, and Tom Chau, “Investigating the need for modelling temporal dependencies in a brain-computer interface with real-time feedback based on near infrared spectra,” *Journal of Near Infrared Spectroscopy*, vol. 20, no. 1, pp. 107–116, 2012.

[166] Raphael Zimmermann, Laura Marchal-Crespo, Janis Edelmann, Olivier Lambercy, Marie-Christine Fluet, Robert Riener, Martin Wolf, and Roger Gassert, “Detection of motor execution using a hybrid fnirs-biosignal bci: a feasibility study,” *Journal of neuroengineering and rehabilitation*, vol. 10, no. 1, pp. 4, 2013.

[167] Danushka Bandara, Senem Velipasalar, Sarah Bratt, and Leanne Hirshfield, “Building predictive models of emotion with functional near-infrared spectroscopy,” *International Journal of Human-Computer Studies*, vol. 110, pp. 75–85, 2018.

[168] Kelly Tai and Tom Chau, "Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface," *Journal of neuroengineering and rehabilitation*, vol. 6, no. 1, pp. 39, 2009.

[169] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[170] Senem Velipasalar Danushka Bandara, Leanne Hirshfield, "Classification of affect using deep learning on brain blood flow data," *Journal of Near Infra Red Spectroscopy*, Under Review.

[171] Sepp Hochreiter and Jürgen Schmidhuber, "Lstm can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[172] Paul R Davidson, Richard D Jones, and Malik TR Peiris, "Eeg-based lapse detection with high temporal resolution," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 832–839, 2007.

[173] Takaya Tamaki, Satoru Hiwa, Keisuke Hachisuka, Eiichi Okuno, and Tomoyuki Hiroyasu, "Region-of-interest estimation using convolutional neural network and long short-term memory for functional near-infrared spectroscopy data," .

[174] Byron Reeves, Annie Lang, Eun Young Kim, and Deborah Tatar, "The effects of screen size and message content on attention and arousal," *Media Psychology*, vol. 1, no. 1, pp. 49–67, 1999.

[175] Dimitri Van der Linden, Michael Frese, and Theo F Meijman, "Mental fatigue and the control of cognitive processes: effects on perseveration and planning," *Acta Psychologica*, vol. 113, no. 1, pp. 45–65, 2003.

[176] Qinqin Wang, Patrick Cavanagh, and Marc Green, "Familiarity and pop-out in visual search," *Perception & psychophysics*, vol. 56, no. 5, pp. 495–500, 1994.

[177] Scott A Huettel, Peter B Mack, and Gregory McCarthy, "Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex," *Nature neuroscience*, vol. 5, no. 5, pp. 485, 2002.

[178] Philippe-Olivier Harvey, Philippe Fossati, Jean-Baptiste Pochon, Richard Levy, Guillaume LeBastard, Stéphane Lehéricy, Jean-François Allilaire, and Bruno Dubois, "Cognitive control and brain resources in major depression: an fmri study using the n-back task," *Neuroimage*, vol. 26, no. 3, pp. 860–869, 2005.

[179] Eric T Greenlee, Gregory J Funke, Joel S Warm, Ben D Sawyer, Victor S Finomore, Vince F Mancuso, Matthew E Funke, and Gerald Matthews, "Stress and workload profiles of network analysis: Not all tasks are created equal," in *Advances in human factors in cybersecurity*, pp. 153–166. Springer, 2016.

[180] J Raymond Comstock Jr and Ruth J Arnegard, "The multi-attribute task battery for human operator workload and strategic behavior research," 1992.

[181] William D Miller Jr, "The us air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior," Tech. Rep., Consortium Research and Fellows Program, Arlington VA, 2010.

[182] Hasan Ayaz, Patricia A Shewokis, Adrian Curtin, Meltem Izzetoglu, Kurtulus Izzetoglu, and Banu Onaral, "Using mazesuite and functional near infrared spectroscopy to study learning in spatial navigation," *Journal of visualized experiments: JoVE*, , no. 56, 2011.

[183] Christopher D Wickens, Robert S Gutzwiller, and Amy Santamaria, "Discrete task switching in overload: A meta-analyses and a model," *International Journal of Human-Computer Studies*, vol. 79, pp. 79–84, 2015.

[184] Christopher D Wickens and Jason S McCarley, *Applied attention theory*, CRC press, 2007.

[185] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

VITA

NAME OF AUTHOR: Mallika Arachchilage Danushka Sandaruwan Bandara

PLACE OF BIRTH: Colombo, Sri Lanka

DATE OF BIRTH: November 28, 1984

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

SYRACUSE UNIVERSITY, Syracuse, New York; August 2010-August 2018

UNIVERSITY OF MORATUWA, Katubedda, Sri Lanka; 2005-2009

DEGREES AWARDED:

B.S. (Hons.) Electrical Engineering (2009) UNIVERSITY OF MORATUWA

M.S. Computer Engineering (2013), SYRACUSE UNIVERSITY

PROFESSIONAL EXPERIENCE:

2014-2017

Lead Software Engineer, GREENVIEW ENERGY MANAGEMENT SYSTEMS, Rome, New York