Modeling Heart and Brain signals in the context of Wellbeing and Autism Applications

A Deep Learning Approach

Juan Manuel Mayor Torres

Advisor: Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science - DISI
University of Trento

This dissertation is submitted for the degree of
Doctor of Philosophy

University of Trento, Italy January 2020

I would like to dedicate this thesis to my loving parents Nain Alberto Mayor Motato, and Maria Teresa Torres Castro all your support for all these hard-working years of struggles and fighting finally gives a success in my career and my life. This is all yours

I would also like to dedicate this thesis to my beautiful son Juan Martin Mayor Herrera he is now the light and my motivation why I'm doing... My world is complete with him!

Acknowledgements

I would like to acknowledge a lot of people in this academic travel. I would like to thank first my parents Naim Alberto Mayor Motato and Maria Teresa Torres Castro for all their support and constant encouraging to finish this Doctor of Philosophy degree (PhD). They were a key element for my motivation and stamina.

I want to thank my advisor professor Dr. Ing. Giuseppe Riccardi for his patience and constant support in this project, and professor Morena Danielli, PhD for her constant support in the most difficult moments of this path. I would like to thank Evgeny Stepanov, PhD for all his help in the beginning of this PhD journey, and my good, and bad moments when I was in Trento.

I also want to thank professor Matthew D. Lerner PhD. for his incredible support, and his discussions about research and life from the academic point of view, they were so valuable, and professor Christian C. Luhmann, PhD. for all his work supporting my ideas and proposing intelligent methods and initiatives in this project.

I also want to thank my brother Carlos Fernando Mayor Torres for all his constant support and strength, all my uncles and cousins and all my near family members for their constant support and kindness in this process.

I thank God for giving me a possibility to be with my beautiful and intelligent son Juan Martin Mayor Herrera. I love him so much! He is my light into darkness!

I want to thank my good friend Heather O'Leary for all her support meanwhile I was sick, and her discussions about research and clinical trials that were so important for me in the execution and analysis of this particular project. She rules!

I want to also thank Tessa Clarkson, Kathyrin Hauschild, PhD, and Rebecca Bianchi for all their work helping me to learn more about Autism Researcher, English grammar, and research procedures Thank you very much!!

And finally I want to acknowledge my good friend Don Rubercindo. He has been an incredible support and help when I have need him in order to pursue this PhD in the difficult moments.

Abstract

The analysis and understanding of physiological and brain signals is critical in order to decode user's behavioral/neural outcome measures in different domain scenarios. Personal Health-Care agents have been proposed recently in order to monitor and acquire reliable data from daily activities to enhance control participants' wellbeing, and the quality of life of multiple non-neurotypical participants in clinical lab-controlled studies.

The inclusion of new wearable devices with increased and more compact memory requirements, and the possibility to include long-size datasets on the cloud and network-based applications agile the implementation of new improved computational health-care agents. These new enhanced agents are able to provide services including real time health-care, medical monitoring, and multiple biological outcome measures-based alarms for medical doctor diagnosis.

In this dissertation we will focus on multiple Signal Processing (SP), Machine Learning (ML), Saliency Relevance Maps (SRM) techniques and classifiers with the purpose to enhance the Personal Health-care agents in a multimodal clinical environment. Therefore, we propose the evaluation of current state-of-the-art methods to evaluate the incidence of successful hypertension detection, categorical and emotion stimuli decoding using biosignals.

To evaluate the performance of ML, SP, and SRM techniques proposed in this study, we divide this thesis document in two main implementations: 1) Four different initial pipelines where we evaluate the SP, and ML methodologies included here for an enhanced a) Hypertension detection based on Blood-Volume-Pulse signal (BVP) and Photoplethysmography (PPG) wearable sensors, b) Heart-Rate (HR) and Inter-beat-interval (IBI) prediction using light adaptive filtering for physical exercise/real environments, c) Object Category stimuli decoding using EEG features and features subspace transformations, and d) Emotion recognition using EEG features from recognized datasets.

And 2) A complete performance and robust SRM evaluation of a neural-based Emotion Decoding/Recognition pipeline using EEG features from Autism Spectrum Disorder (ASD) groups. This pipeline is presented as a novel assistive system for lab-controlled Face Emotion Recognition (FER) intervention ASD subjects. In this pipeline we include a Deep ConvNet as the Deep classifier to extract the correct neural information and decode emotions successfully.

Table of contents

List of figures

List of tables

Nomenclature

ACC	Anterior Cingular Cortex
ADI	Autism Diagnostic Interview
ADOS-CSS	Autism Diagnostic Observation Schedule - Calibrated Severity Score
ADOS	Autism Diagnostic Observation Schedule
α	α Rhythm
β	β Rhythm
γ	γ Rhythm
μ	μ Rhythm
θ	θ Rhythm
ANN	Artificial Neural Networks
AQ	Autism-Spectrum Quotient
ASD	Autism Spectrum Disorder
BCI	Brain Computer Interface
BVP	Blood Volume Pulse
ConvNet	Convolutional Neural Network
ECG	Electro-Cardiography
EEG	Electro-Encephalography
ERDS	Emotion Recognition and Display Survey

ERP	Event-Related Potentials
FER	Face Emotion Recognition
FG	Fusiform Gyrus
fMRI	functional Magnetic Resonance Imaging
GLM	Generalized Linear Model
GSR	Galavanic Skin Response
ICA	Independent Components Analysis
LPP	Late Positive Potential
LRP	Layer-Wise Relevance Propagation
LSTM	Long Short-Term Memory
MEG	Magneto-Encephalography
mPFC	Medial Prefrontal Cortex
PCA	Principal Components Analysis
QCNN	Quaternion Convolutional Neural Network
QNN	Quaternion Neural Network
QRNN	Quaternion Recurrent Neural Network
RBM	Restricted Boltzmann Machine
SCQ	Social Communication Questionnaire
SRS	Social Responsiveness Scale
SSO	Social Skills Observation
SSRS	Social Skills Rating System
STS	Superior Temporal Sulcus
SVM	Support Vector Machine

Chapter 1

Introduction

> "As long as our brain is a mystery, the universe, the reflection of the structure of the brain will also be a mystery.."
>
> *Santiago Ramon y Cajal*

The implementation of automatic classification-based pipelines on recent health-care environments has become an important research field nowadays and the near future research projects (Beam and Kohane, 2018; Iqbal et al., 2016). Multidisciplinary and Interdisciplinary are more necessary to understand deeply the influence of biosignal based systems, the reliability of neural data acquisition, and the statistical incidence of behavioral, biological, and neural outcome measures important for other measure of clinical micro-states, and diagnosis prediction (Park and Han, 2018).

Multimodality is considered an important feature in real clinical environments as well as the synchronicity, artifact detection and removal, and waveform characterization and signal morphological categorization (Athavale and Krishnan, 2017). For the purpose of this dissertation we considered the studies without a clear (Stimulus Onset Asynchrony) SOA as an "in-the-wild" study where the stimulus and the neural activity are not time-locked, and therefore not statistically linked, and in the opposite site a SOA study where the stimulus and the neural activity are locked in time and therefore the study is connecting more the stimulus elicitation with the neural activity of the TD and ASD individuals.

Neuroscientists, psychologists, and computer scientists have made a serious effort on create comfortable, adaptive, and wearable systems which are not only able to present signals, but also to analyze, and predict future outcomes measures from neurological disorder patients, and neurotypical controls. Clinicians suggest to not only improve the multimodality features on health-case agents, but also to evaluate effects between central biosignals such as EEG,

fMRI, and MEG, with peripheral such as ECG, BVP, GSR and all the subsequent correlates (Lischke et al., 2017).
In this dissertation we will describe multiple EEG pipelines using broadly-known datasets such as DEAP and TROIKA, and a lab-controlled time-locked clinical trial including Autism Spectrum Disorder (ASD) participants in order to describe how signal processing (SP), and machine learning (ML) methodologies, and novel results influence positively to no-SOA and lab-controlled environments being this latter a treatment and intervention study on ASD groups.

1.1 Motivation

Our main motivation with this project and dissertation is **to describe and explain the quantifiable incidences of robust and new state-of-the-art SP and ML techniques on clinical lab-controlled and non-clinical no-SOA environments to predict emotion, categorical labels, and other biological outcome measures.**
We propose in the first part of the dissertation 1) to evaluate our current pipelines using broadly evaluated datasets on neuro-typical participants for predicting stimul-based classes, and for the second and main part of the dissertation 2) we evaluate a Deep classifier architecture such as the Deep ConvNet over a lab-controlled clinical trial including ASD participants, and complementing the face emotion recognition tasks (Bzdok and Meyer-Lindenberg, 2018).
With the pipelines proposed here we are not only evaluating the performance of multiple SP and ML methods, but **the statistical relationship between the pipelines learnt and resulting parameters with the behavioral and neural outcome measures.**
As a novel evaluation for ML systems on clinical environments. We are using a set of robust and recent saliency methods to describe with the highest reliability the neural activity relevance in lab-controlled clinical environments assuming the Deep ConvNet as a "black-box" object (Zhang and Zhu, 2018). This technical consideration is important and critical for understanding what clinical outcome measures and its corresponding correlates are more relevant for the system to do correct encoding and subsequent emotion class decoding (Kindermans et al., 2017c).
In this study we wont emphasize in how the classifiers are trained to decode multiple behavioral labels in multiple data collection scenarios only, but **how input features, or initial biosignals's features can influence on final classification results, and thus predict correlated heart and brain outcome measures on clinical lab-controlled environment.**

1.1.1 ASD clinical trial - Motivation

Our proposed real clinical analysis included ASD participants and a corresponding Deep Classifier pipeline's tests. ASD is a common neurological disorder presented in a rate of 10/1000 individuals around the globe following World Health Organization (WHO) and American Psychiatric Association (APA) official reports, **thus affecting in the broad majority children, adolescents, and young adults almost with the same level of severity permuting normal distributions of individuals around the globe** (Stavropoulos et al., 2018).

ASD participants show critical impairments for recognizing others' emotions and emotion appraisal. **This emotion recognition impairment is associated with different connectivity patterns observed in ASD individuals' mirror-neurons and emotion processing neuromechanisms, and it is considered a crucial deficit in important human communicative and executive tasks such as FER, communication, and social interaction** (Baron-Cohen, 2016).

Recent studies confirm neural deficits evaluating FER and emotion appraisal in ASD groups comparing Event-Related Potentials (ERP) amplitudes and timing, thus showing emotion processing structures differences using functional Magnetic-Resonance Imaging (fMRI) data. **This neurological unbalance is also correlated positively with altered neural connectivity for processing emotions in ASD groups** (Ameis and Catani, 2015).

Neuroscientists, psychologists, and psychiatrists associate the behavioral deficit observed in ASD individuals with a neural deficit observed with multiple signals and sources through the last decade. However, the differences and effects observed in neural signals have been identified using average/group-level analysis. In current studies we can not individualize the neural activity in single-trials to observe potential variability per trial, especially when a ML pipeline is included.

In this dissertation we propose an initial evaluation of current state-of-the-art techniques for a biosignals encoding, and subsequent and new Deep learning model based on a Deep ConvNet for a robust EEG-based emotion decoding applied to ASD, and non-ASD groups in multiple age-ranges. **We implemented, for the first time, a Deep ConvNet using a 2D EEG feature-set collected from a FER clinically-controlled experiment preserving the critical and important early and late ERPs associated with emotion processing.**

Additionally, in this study we evaluate the comparison human v.s machine performances on ASD, and non-ASD groups finding quantifiable differences between FER human performances, and the Deep ConvNet performances across the age-ranges and groups. For every age group the Deep ConvNet outperform the human performances supporting the presence of plausible and intact emotion processing neuromechanisms in ASD population preserved by our proposed Deep learning pipeline. This pipeline thus find a complete emotion neuromechanism in ASD individuals using alternative analysis such single-trial classification is considered a big milestone in Autism and neuroscience research.

Our experience working with clinical groups working for ASD and Rett syndrome research teams in Boston Children's Hospital (BCH), and Stony Brook University (SBU), and the knowledge we got with the data collection, contact with participants, and data analysis procedures were priceless and invaluable to enhance the quality of our research work and scopes of this dissertation.

In the following sections we will describe an overview of this PhD project and the structure of this dissertation as well as the relevant publications related to this project.

1.2 Thesis Goals

As described above, we will set the thesis goals between the two parts of the thesis. The first goals will involve a preliminary evaluation of SP and ML systems on current broadly used datasets for heart-rate, object category stimuli, and emotion decoding based on biosignals:

1. To evaluate preliminary emotion decoding in arousal and valence high-low levels using the DEAP dataset.

2. To evaluate performance of two-class pictograph object category decoding using epoched EEG signals features.

3. To calculate HR and IBI measures from real-life scenario under exercise from the TROIKA dataset, and daily-life activity environment.

From this initial evaluation we proceed to apply the knowledge learnt from this evaluation to a real clinical scenario. As mentioned above the inclusion of deep classifiers such as the Deep ConvNet have been evaluated recently in clinical environments, however, in this project we propose to include the processing of 2D feature-set initial arrangement for EEG single trials on ASD clinical studies. For this clinical evaluation we propose a main comparison between human emotion recognition, and a Deep ConvNet (machine) emotion recognition performances. For this part of the thesis we set the following goals:

1. To evaluate how a Deep ConvNet-based pipeline can be considered an intervention tool and/or an online classifier for ASD population behavioral treatment.

2. To map the most relevant features from the neural activity using emotional stimuli in ASD and nonASD groups, and match this information with current connectivity studies.

3. To correlate the machine (Deep ConvNet) training and performance parameters with the Autism behavioral and severity assessment scores.

With this evaluation on a complex clinical environment such as emotion recognition autism intervention, we propose a quantifiable generalization of a Deep ConvNet classifier to decode emotion successfully and support ASD behavioral deficits.

1.3 Thesis Contributions

The thesis' main contributions are focused on the lab-controlled clinical study including ASD participants evaluating there a Deep emotion recognition pipeline. However, in our initial evaluation we contribute to performance evaluation on current no-SOA datasets. An important contribution is not to include a more signal correlated stimuli such time-locked emotion faces, but to construct a more reliable system for behavioral outcome measures decoding (Dinov and Leech, 2017).

With a proper construction of neural (central and peripheral) signals stimuli correlation we will observe not only a better performance in the behavioral outcome measure decoding, but also a more realistic pipeline and performance results correlated with the behavioral and biological conditions of the TD and ASD participants.

The main purpose of this project is to make the Artificial Intelligence (AI) advances and ML closer to clinical assessments. Particularly, for ASD this project is a big step to connect psychological intervention and treatment methodologies with robust deep classifiers extending a the impact of multimodality, and data reliability as a relevant features for new pipelines (Andreotti et al., 2018).

1.4 Structure of the Thesis

The structure of this thesis dissertation document is composed of the following chapters:

1. **Chapter 2 - Autism Spectrum Disorder, Intervention, and Treatment Background:** In this chapter we will describe the ASD neural characterization, and behavioral assessments focusing on the Face Emotion Recognition methodologies.

2. **Chapter 3 - Multimodal Neural based Emotion Decoding for ASD and non-ASD individuals:** Here we will describe the current and more important state-of-the-art for AI implementations on ASD and non-ASD clinical trials. We

will also describe the AI implementations for behavioral and neural outcome measures in intervention studies.

3. **Chapter 4 - Machine Learning preliminary Evaluation on DEAP, Object Categories, and TROIKA datasets:** This chapter is focused on the preliminary evaluation on existing datasets such as DEAP, TROIKA, and Object Category stimuli dataset. Here we will make emphasis on the pipeline compositions, training methodologies, and results.

4. **Chapter 5 - ConvNet Pipeline for EEG-based Enhanced Emotion Decoding in Autism:** This chapter will describe the performance results of the Deep ConvNet classifier using the EEG features from ASD and non-ASD participants across three different age groups. Barplots, Confusion Matrices, and Tables are included in this chapter.

5. **Chapter 6 - Correlation between Deep ConvNet parameters and ADOS-CS:** This chapter contains the statistical analysis with Multiple comparisons ANOVA, and Pearson correlation analysis to compare the machine parameters interaction and the behavioral scores on ASD and non-ASD groups.

6. **Chapter 7 - Saliency Maps Evaluation - EEG Features Relevant Measures:** In this chapter we will explain, and describe the most robust Saliency maps in the ML state-of-the-art. These Saliency methods are included in the package iNNvestigate and the evaluation on EEG-based ASD emotion decoding shows important results linked with previous neural connectivity studies.

7. **Chapter 8 - Conclusions**

In addition with the chapters enumerated above, we include two additional appendices. **Appendix A** including the Generalized Linear Model used for the statistical comparison, and correlation analysis of **Chapter 6**, and **Appendix B** including the mathematical models for each Saliency Method described in **Chapter 7**.

1.5 Thesis Relevant Publications

For this project we have published some related and relevant papers associated with the first part of the thesis, or the preliminary ML systems evaluation such as:

1. Ghosh, A., **Torres, J. M. M.**, Danieli, M., and Riccardi, G. (2015, August). Detection of essential hypertension with physiological signals from wearable devices. In 2015

37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 8095-8098). IEEE.

2. Ghosh, A., Danieli, M., and Riccardi, G. (2015, August) Annotation and prediction of stress and workload from physiological and inertial signals. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),pp. 1621-1624, IEEE.

3. **Torres, J. M. M.**, Stepanov, E. A., and Riccardi, G. (2016, May). Eeg semantic decoding using deep neural networks. In Rovereto Workshop on Concepts, Actions, and Objects (CAOS).

4. **Torres, J. M. M.**, Ghosh, A., Stepanov, E. A., and Riccardi, G. (2016, August). Heal-T: An efficient PPG-based heart-rate and IBI estimation method during physical exercise. In 2016 24th European Signal Processing Conference (EUSIPCO) (pp. 1438-1442). IEEE.

5. **Torres, J. M. M.**, and Stepanov, E. A. (2017, August). Enhanced face/audio emotion recognition: video and instance level classification using ConvNets and restricted Boltzmann Machines. In Proceedings of the International Conference on Web Intelligence (pp. 939-946). ACM.

6. Ghosh, A., Stepanov, E. A., **Torres, J. M. M.**, Danieli, M., and Riccardi, G. (2018, September). HEAL: A Health Analytics Intelligent Agent Platform for the acquisition and analysis of physiological signals. In 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom) (pp. 1-6). IEEE.

For the second part of the thesis, we have published three papers, and now we are preparing two closing journal papers related to the ASD population clinical study extending the two main contributions described above. This work was a part of SCTL Stony Brook University internship and they will be published soon:

1. **Torres, J. M. M.**, Clarkson, T., Stepanov, E. A., Luhmann, C. C., Lerner, M. D., and Riccardi, G. (2018, July). Enhanced Error Decoding from Error-Related Potentials using Convolutional Neural Networks. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 360-363). IEEE.

2. **Torres, J. M. M.**, Libsack, E.J., Clarkson, T., Keifer, C.M., Riccardi, G., and Lerner, M. D. (2018, May). EEG-based Single trial Classification Emotion Recognition: A

Comparative Analysis in Individuals with and without Autism Spectrum Disorder. International Society for Autism Research, (INSAR), 2018, 27651

3. **Torres, J. M. M.**, Clarkson, T., Luhmann, C.C., Riccardi, G., and Lerner, M. D. (2019, May). Distinct but Effective Neural Networks for Facial Emotion Recognition in Individuals with Autism : A Deep Learning Approach. International Society for Autism Research, (INSAR), 2019, 30962

Chapter 2

Autism Spectrum Disorder, Intervention, and Treatment Background

The prevalence around the world of Autism Spectrum Disorder (ASD) has been evaluated by American Psychiatric Association (APA) and World Health Organization (WHO) reaching record levels in recent decades where an approximate average of more than 10/1000 children are diagnosed with ASD (Lord et al., 2000; Mahdi et al., 2018; WHO, 2014).

An individual diagnosed with ASD using the Autism Diagnostic Observation Schedule v2.0 (ADOS-2) (Gotham, Pickles, and Lord, 2009) the most accepted and evaluated assessment, can be categorized as Low-Functioning ASD individuals being affected by cognitive and motor disabilities and impairments related to the Autism spectrum such as communication impairments, important changes on brain structures such as Midbrain, Corpus Callosum, and Hippocampus, and other additional disorders such as Fragile X, Down Syndrome, Anxiety, and Intellectual disabilities. In most cases this group is considered Non-verbal autism.

On the other hand, Autism individuals can be categorized as a high-functioning where the cognitive and physical impairments are not so profuse or severe, but the communication skills and social competence are still widely affected between the ASD population, thus constraining ASD individuals in daily life tasks, verbal communication, and social (Fletcher-Watson et al., 2014).

New studies project an increased prevalence of ASD in children and adolescents age-ranges, which represents a considerable high cost for treatment and intervention even for countries with a high-quality of life such USA, Canada, and northern European countries (Lundström et al., 2015; Lyall et al., 2017).

In terms of Public Health research ASD and its related neurological syndrome have been studied profusely to describe critical and related biological and behavioral factors (Caronna, Milunsky, and Tager-Flusberg, 2008; Waddington et al., 2018). Recent epidemiological studies relate the emotion recognition impairments (Fletcher-Watson and Happé, 2019; Wijnhoven et al., 2018) with some Autism prevalence factors such as the Attention-Deficit/ Hyperactivity Disorder (ADHD), Rett, and the Fragile X syndrome comorbidities (presence of both disorders or syndromes) (Lyall et al., 2017; Ng, Heinrich, and Hodges, 2019). Figure 2.1 summarize ASD prevalence research items such as *Descriptive Epidemiology*, *Genetic Epidemiology*, and all the *Environmental factors* measured in the recent ASD prevalence studies.

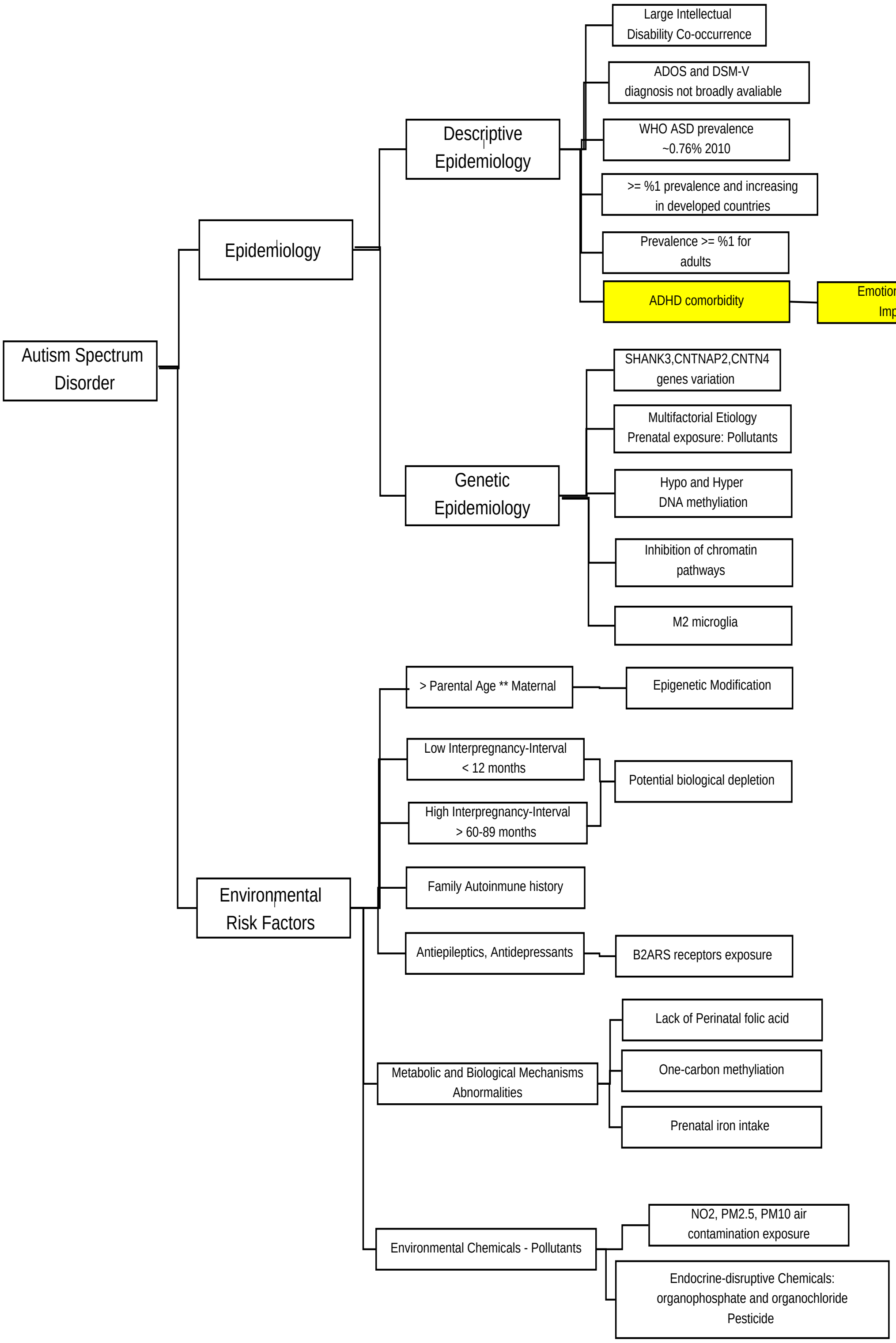

Large Intellectual Disability Co-occurrence
ADOS and DSM-V diagnosis not broadly avaliable
WHO ASD prevalence ~0.76% 2010
>= %1 prevalence and increasing in developed countries
Prevalence >= %1 for adults
ADHD comorbidity
Emotion Recognition Impairments
Descriptive Epidemiology
Epidemiology
Autism Spectrum Disorder
SHANK3,CNTNAP2,CNTN4 genes variation
Multifactorial Etiology Prenatal exposure: Pollutants
Hypo and Hyper DNA methyliation
Genetic Epidemiology
Inhibition of chromatin pathways
M2 microglia
> Parental Age ** Maternal
Epigenetic Modification
Low Interpregnancy-Interval < 12 months
Potential biological depletion
High Interpregnancy-Interval > 60-89 months
Family Autoinmune history
Environmental Risk Factors
Antiepileptics, Antidepressants
B2ARS receptors exposure
Lack of Perinatal folic acid
One-carbon methyliation
Metabolic and Biological Mechanisms Abnormalities
Prenatal iron intake
NO2, PM2.5, PM10 air contamination exposure
Environmental Chemicals - Pollutants
Endocrine-disruptive Chemicals: organophosphate and organochloride Pesticide

Fig. 2.1 Organigram/block diagram showing the scientific reference points for prevalence study of ASD population. Mutiple enviromental factors and descriptive/genetic studies focus the study of behavioral, biological, and neurological aspects of the ASD population on particular items such as modification of SHA, and CNTN prefix genes, prenatal age, parents age, inter-pregnancy interval, M2 Microglia, inhibition of chromatin pathways and so on.

The positive correlation between the comorbidity of emotion recognition impairments and other common syndromes and disorders extend the incidence of Genetic Epidemiology, and its corresponding correlation with the emotion recognition impairments to more individuals with ASD these with specific emotion recognition deficits. Recent epigenetic studies (Wendt et al., 2019) have examined the variation and deletion of PHB014 with results of human Face Emotion Recognition (FER) tasks in individuals diagnosed with Autism. As well as previous FER behavioral analysis (Dawson et al., 2002), the recent studies involving epigenetics point out multiple important features such as *Accurate Recognition of Emotion*, *Speed of Emotion Recognition*, and *Differentiating Emotion Intensity*.

Emotion recognition on different executive levels can be represented as an important item of the ASD group's Descriptive Epidemiology section. Emotion recognition is now broadly studied from controlled and epoched experiments such as auditory, especially voices, and visual, especially faces (Harms, Martin, and Wallace, 2010; Hobson, Ouston, and Lee, 1988). During the last decade the study of face recognition, and the brain regions encoding face units has been developed broadly (Bal et al., 2010; Dawson, Webb, and McPartland, 2005; Tian, Kanade, and Cohn, 2001).

In this study we will incidence profusely on at least two emotion recognition metrics such as *Accurate Recognition of Emotion* and *Speed of Emotion Recognition* specifically on visual-facial emotion recognition performance of ASD groups. The inclusion of Deep classifiers as we will mentioned in the following chapters will be presented as a robust and adequate option for neural-based emotion recognition online classifier for ASD individuals. In the following sections and subsections we will discuss how Autism researchers have analyzed finding effects between behavioral, biological, and neurological variables, as well as correlating these effects with possible intervention and treatment methodologies.

2.1 Face Emotion Recognition (FER) - ASD

Preliminary reviews analyzing emotion recognition in ASD populations (Dawson, Webb, and McPartland, 2005; Dawson et al., 2002) Autism researchers set a three critical and important neurological outcome measures related to emotion recognition metrics such as *Accurate Recognition of Emotion* and *Speed of Emotion Recognition*. The first measure denoted by Autism researchers and neuroscientists is *Gaze and eye contact* which is important for influencing a posterior social motivation and subsequent emotion appraisal. Most studies analyze the gaze and eye contact converting the FER task as executive task, where the face stimuli is divided into units such as eyes, nose, and mouth.

Recent studies measure the gaze and link this gaze fixation with memory and neural correlates using eye tracker studies (Murias et al., 2018; Vernetti et al., 2018). The first effect measured was a different neuromechanism observed between Controls and ASD groups. This is also associated with early cognitive Theory-of-Mind non-processing (Senju et al., 2009) in ASD groups.

A second neurological measure observed in FER studies is the ability for *Face Memory* to occur. Recent studies measure the performance for face discrimination, and face recognition. For children and adolescents the performance is different but is not considerable between Controls and ASD grpups. However, in more recent studies this difference is more plausible between Controls and ASD in Adults age ranges (Vettori et al., 2019) relating the emotion recognition impairment within a selective behavior and thus supporting the activation of an altered neuro-mechanism for emotion recognition in the ASD group.

The third measure found in the early behavioral studies is *Abnormal strategies* for those with ASD to process a face prior a social interaction. Specifically, these effect are observed when the altered neural network activation used to identify and recognize faces and their corresponding emotions better from parts rather than a holistically (Lynn et al., 2018).

Despite multiple behavioral and neural correlations studies published in the current ASD state-of-the-art. There is still discrepancies in results that can relate altered neural structure from the ASD groups with the emotion recognition impairments. Multiple neural structures such as Amygdala, ACC, and STS show low activations (Ameis and Catani, 2015; Black et al., 2017), and few connectivity patterns in ASD groups in comparison with Controls. Others regions such as mPFC, and pFC, Precuneus, Stratum, Insula, and other white matter areas with hyper-myleniation show higher levels of connectivity in individuals with ASD supporting again the different strategies on face and emotion perception in Autism individuals.

In the following subsections we will describe clinical studies including Autism population with and without comorbidities which have studied the statistical relevance of important EEG neuro-correlates such as N170, N200, P300, P600, and Late Positive Potential (LPP) with its corresponding timing windows.

2.1.1 Electrophysiological Face Processing in Autism Individuals

The basic units for measuring the prevalence and activation of neural activity in clinical studies is the Event-Related-Potential (ERPs). An ERP is defined by basis (Luck and Kappenman, 2011; Perry et al., 2015) as an electrical measure showing a dendritic post-synaptic aggregation associated with a specific neurobiological stimulus recorded from an EEG clinical device. An ERP can be detected as a positivity, negativity or a big slope elicited from neurological stimulus. The neural potential associated with the stimulus can be observed as a difference in amplitude, time, and frequency outcome measures such as

Event-related De-synchronization (ERD) conforming a critical measure for epoched trials as Event-Related Spectral Perturbation (ERSP) (Wang et al., 2015). From this point of view the ERSP can be catalogued as an important outlier across time from this type of time-locked stimuli.

Multiple other measures are used for stimuli without time locking. Most of these measures are used in order to understand brain activation in a non-invasive way for resting-state, or longer tasks associated to neural micro-state (Wang et al., 2013). Multiple studies use frequency analysis for epoched trials for face processing because it is easier to relate face stimuli in a time locked environment. The power of the spectrum from infra-slow ranges from δ [1-3] Hz, θ [4-7] Hz, μ [8-10] Hz, to low frequencies such as α [8-13] Hz, Low β [15-20] Hz, and High β [20-40] Hz, to a high frequencies such as γ [> 40Hz] and ripples higher than 100 Hz.

The neurological correlates associated in both types of experiments have been analyzed in Controls and individuals with ASD finding important correlation between face processing, and the subsequent emotion processing associated with the face units.

ERPs associated with Face processing

Cognitively speaking neuroscientists in Autism research have tried to segment the ERPs observation in early and late neural stages (Webb et al., 2006). The initial analysis have explored the face structures in parts, and holistically finding differences in amplitude and latencies between modalities.

In adult humans a negativity is observed constantly in 170 ms after the stimulus onset. This negativity is denoted by N170 slope. The N170 is observed with a certain level of latency varying and comparing faces with objects, or faces with faces structural variations such as interventions, eyes-only, or deconstructed faces (Dawson et al., 2002).

When the Autism individuals are included in the initial EEG-based face processing experiments bigger latencies are observed in the N170 component. This effect is associated with an early face processing disruption (McPartland et al., 2004), and the Autism altered network connectivity. Furthermore, the N170 different neural modulation is also observed in different cortical regions such as more central activation in comparison with the basal activation in temporo-occipital regions observed in those without ASD.

In general the shorter response of the N170, and the precursor of the N170 prN170 observed in the typically developed individuals are also associated with a higher speed on face processing and recognition in this group in comparison with ASD (Webb et al., 2011). With the prN170 there is also morphological observations, for instance quicker responses of prN170 in autistic children in comparison with adults, and smaller amplitude of prN170 for autistic children too (Webb, Neuhaus, and Faja, 2017). Across age ranges, the N170 is more negative in adults than in children or adolescents. Studies with neonates confirms not only differences in amplitude but also in terms of latencies to process faces, thus relating this effect with cortical brain structures maturity, and other subcortical structures such as STS and amygdala (Webb et al., 2006). However, other studies suggest no differences between N170 amplitude and latencies across age groups.

Recent magnetic resonance studies correlate atypical lateralization of the N170 in ASD groups (Ji et al., 2019) assuming again a different and altered connectivity pattern between Control and ASD groups. Other EEG-fMRI studies correlate the N170 with a-posteriori emotion processing structures such as ACC, amygdala, and insular cortex only for controls (Bayer et al., 2019), and showing again differences with ASD groups.

In summary studies including the analysis of face processing and specifically N170 and nearby components have found an important early and subsequently pre-cognitive disruption of face processing. Further, low and small neural activation presented in autistic individuals.

Alpha-α and Gamma-γ measures - Face processing

In recent ASD studies involving EEG signals researchers have found frontal γ activity involved in early face processing stages Naumann et al., 2018. Individuals with autism show no behavioral differences recognizing holistically structured faces in comparison with a non-structured face. However, individuals with ASD shown an elevated γ power from P1 to the Late Positive Complex (LPC) range around 800ms after the stimulus onset.

Other very sensitive outcome measures from γ rhythm is the lateralization index defined as any measure normalizing additive differences to follow the expression $\frac{RH-RL}{RH+LF}$ where RH is the right-hemisphere measure, and LF the left-hemisphere corresponding measure. Selecting any channel from fronto-central regions a lateralization index is showing significant differences inside the γ rhythm after P1 showing a different synaptic pathway related to face processing (Keehn et al., 2015).

MEG studies (Leung et al., 2018) also support the over-activation of temporal and insular cortical and subcortical structures such as STS, and Fusiform Gyrus in Autism individuals in comparison with typically developed who correlates the N170 peaks with activation of the occipital and frontal structures such as ACC, Amygdala, and mPFC. This correlation also suggest an alternative strategy not only for the face processing and memory tasks but also for emotion processing.

Other studies have associated higher frequencies such as α and β rhythms with a higher social competence measure, and a lower ASD severity (Courellis et al., 2019) for those with ASD group. A recent study has complement the previous desynchronization results with a

high amplitude frontal activations of α, β and γ in individuals with ASD. This fact points again to an altered connection for processing faces in ASD group.
To link the early face processing impairments found in Autism individuals, a recent study (Almeida et al., 2016) found increased α rhythm related to an early arousal perception with no variation across face structures such as upright, inverted, and holistic type of emotional faces. Therefore, the question is open to know if the altered network found in individuals with ASD only correlates with early face processing or with emotion processing and other post-cognitive ERPs.

2.1.2 Neurocorrelates related to Emotion processing- Faces- Autism

As previously we mentioned we itemize multiple neurocorrelates related with early face processing deficits. However, we can enumerate four different implications related to the face processing speed and the subsequent emotion appraisal:

1. The face processing deficit observed in early stages is also considered a **perceptual/cognitive deficit** in individuals with ASD.

2. The early face processing deficit **is preventing ASD groups to extract important information from face features** and **detect the corresponding emotional content.**

3. The neural pathway associated with face processing and all the corresponding neural structures such as FG, STS, Amygdala, and mPFC are considered **dysfunctional** in ASD population.

4. Future intervention oriented to face/emotion processing in Autism individuals should be focused on the electrical stimulation of the dysfunctional pathways, or teaching autism individuals **to focus on central features of facial emotion**, and the inclusion of **information-processing facilitation strategies to enhance emotion apprehension**.

These items do not only enumerate deficits for face and emotion processing found in previous studies, but also suggest the possibility to functionally enhance the face and emotion perception from intervention and technologies to help in ASD intervention. In this study we focus on the importance to **use alternative analysis per trial, using Deep classifiers, and creating new emotion-recognition based intervention tools that can help to restore the dysfunctional neuromechanisms presented in ASD individuals.**
In terms of the principles of face processing it is important to understand how the neural pathways process the corresponding emotion category associated and labeled to the face. The arousal segment from the emotion category associated with the face stimulus is processed

early by controls but not by individuals with Autism. Therefore, neuroscientists have tried to look for more neurocorrelates in the late stage of the face processing potentials. The main idea with emotion processing analysis using EEG or neural outcome measures is to differentiate the neurocorrelates associated with particular emotions, specially the negative and the high-arousal categories representing the performance detriment in individuals with ASD (Adolphs, Sears, and Piven, 2001).

Emotion Face processing ERPs - Autism

The first emotion category analyzed in emotion recogniton studies look into impairments in ASD population was *fear*. Representing the most critical and negative emotional state and ASD population can not perceive. Again these *fear* emotion based studies (Dawson et al., 2002) suggest an abnormal strategy, and a considerable detriment in emotion speed recognition and face processing associated with the altered neural connection for negative emotion processing.

Subsequent studies have found important later ERP components in the N300 slope comparing the performances between controls and ASD individuals. Specifically, Controls show a shorter latency for N300 in comparison with ASD individuals and even with Controls who performed worse in the emotion recognition task (Dawson, Webb, and McPartland, 2005).

However, in comparison with the early studies the N300 latencies and amplitude are not correlated with non-social stimuli. These findings linked social interaction impairments with the FER speed performances, and the slower latencies of N300 in ASD groups. (Dawson and Bernier, 2007; Lerner, McPartland, and Morris, 2013).

Recent studies conclude the observation of **a systematical information-processing deficit in individuals with ASD are supported on lower amplitude and a higher latencies from early ERPs in ASD individuals.** In imaging studies we observe structural abnormalities in emotion processing structures related with the previously mentioned temporal ERPs inaccuracies (Safar et al., 2018).

Later stage Emotion Face processing in Autism

To conceptualize the neurological emotion temporal processing on individuals with and without ASD the neural components such as ERPs inside N1, and N2 ranges can be categorized as early perceptual components. The negativities or positivities found after the previous mentioned ranges can be considered as later stage information processing or post-cognitive neural components (Dawson, Webb, and McPartland, 2005).

The corresponding emotion category associated to the face is processed properly when the face information is processed early focusing on eyes, mouth, and nose face units (Dawson and Bernier, 2007). The first intuition about emotion recognition is related to familiarity appraisal. Toddlers and children show an increased P400 amplitude when the mother's face is presented in comparison with a unfamiliar face (Luyster et al., 2014; Webb et al., 2011). Other studies involving children with Autism did not show differences in P400 amplitudes comparing familiar and unfamiliar faces. However, they show differences evaluating between a toy object and a face (Dawson et al., 2002; Jones, Dawson, and Webb, 2018) suggesting again an abnormal activation to process faces and the corresponding emotions.

LPP incidence in emotion recognition - Autism

On the other hand, a longer waveform slope such as LPP is associated with differential emotion regulation, emotion appraisal, theory-of-mind, and empathy processing (Dennis and Hajcak, 2009). The LPP waveform was first analyzed dividing into three windows *Early* [300-600] ms, *Middle* [600-1000] ms , and *Late* [1000-2000] ms. Some studies differ around 20-50 ms before and after the time values stated above (Ferri, Weinberg, and Hajcak, 2012; Foti, Hajcak, and Dien, 2009) but the three windows are critical to evaluate and process emotional faces and scenes in Autism population.

To analyze the LPP sensitivity across emotion categories the initial studies set a group of EEG electrodes across the scalp in a symmetrical way. For the LPP studies the electrodes are divided into four main clusters *Left-Anterior*, *Right-Anterior* ,*Left-Posterior*, and *Right-Posterior*(Schupp et al., 2004). These cluster division will increase the options to find significant effects using multiple comparisons and extended factors as we will describe in Appendix A.

The first emotion variable included in the analysis is the emotion dysregulation comparing neutral and *fear* emotions and it shows a positive correlation with the *Middle*, and *Late* LPP windows in the posterior clusters. Other studies correlate the difference between neutral and negative high/arousal emotions from faces, and a subsequent non-evidence of this difference in ASD (Ferri, Weinberg, and Hajcak, 2012; Luckhardt et al., 2017; Schupp et al., 2003).

More recent studies have linked single-trial classification using LPP features validating them as useful features for emotion recognition, and thus obtaining different neural connectivity patterns between Controls and ASD groups (Mayor Torres et al., 2018). Some studies find that LPP low amplitudes is positively correlated with fewer social interactions and a consequent lower emotion recognition performance in ASD and Control groups (Benning et al., 2016; Clarkson et al., 2019).

Understanding Face/Emotion impairments in ASD

With previous research attributing neural potentials to a post-cognitive emotion processing and all the alternative connectivity patterns found in ASD groups. It is possible to conclude two main statements about the cognitive/affective deficits observed in Autism individuals.

1. The neural dysfunctionality observed in early and late time-ranges is considered an important perceptual/cognitive impairment in ASD.

2. The disfunctionality of brain structures such as FG, STS, Amygdala are not only related to face processing deficits but also with emotion and motivational deficits.

3. A hypothesis is created after ASD cognitive/affective impairments is observed in subsequent neural potentials. **A lack of social motivation is associated with inefficient face/emotion information processing is observed in Autism individuals.** Autism researchers suggest two approaches to improve face/cognitive/emotion information processing capabilities.

    - To implement intervention methodologies based on cognitive *rewards* in which individuals with ASD can infer face/emotion information using feedback from neural structures which process face/emotion information per se (Whyte, Smyth, and Scherf, 2015).
    - To inhibit or stimulate the neural circuitry processing face/emotion information. Dopamine and GABAergic circuits involved in the structures mentioned above are important to enhance FER performances (Barak and Feng, 2016; Chakrabarti and Baron-Cohen, 2011).

The stated hypotheses and the influence associated with the social motivation, and altered connectivity are general across age ranges and the comorbidities. Intervention methodologies can be developed using ERPs and such as N170, N300, P400, and LPP as a construction basis for online intervention in emotion recognition. In the next section and subsections we will describe the resources and methodologies used for psychological intervention based on behavioral and biological outcome measures for individuals with Autism.

2.2 Intervention and Treatment for Emotion Recognition in Autism

The main details for starting a behavioral intervention for ASD groups are a) For children increase the affective exchange and the eye contact towards the face stimulus, making reinforcement through toys and familiar objects associating them with the presented stimulus, b) Increase the face-to-face contact to motivate social interaction in children with ASD, extending it for a cognitive training a the individual preferences, and c) For adults with ASD it is more convenient to describe the face units and features and relate the corresponding emotion with familiar facts and daily life (Webb, Neuhaus, and Faja, 2017; Webb et al., 2011).

From the previously mentioned studies ASD researchers have created multiple FER training assessments in order to affect the attention of socio-emotional cues, enhance social awareness, and Theory-of-Mind (ToM) capabilities in cognitively and affectively sides.

The computer based clinical practices more commonly used by Austim researchers were stimuli sets designed for intervention such as *The Frankfurt Test and Training of Facial Affect Recognition* (FEFA) (Bölte et al., 2006) focusing on the training on face photographs and the corresponding eyes regions, The Transporters video series is also a series of face stimuli taken from the video series for Autism intervention (Adams and Robinson, 2011), The *Emotion Trainer* focusing also on real faces and the stimulation of the FG (Silver and Oakes, 2001), *Let's face it* which teach children and adults ASD individuals to enhance identity and emotion perception (Tanaka et al., 2010), The computer based assessment used for Autism ToM-based intervention *FaceSay* (Whyte, Smyth, and Scherf, 2015), and *Mind Reading* stimuli set included in the Cambridge Mind-Reading Face-Voice battery (Golan, Baron-Cohen, and Hill, 2006) which familiarize autistic individuals introducing the concept of attributing/perceiving emotion state to each face stimulus (Lacava et al., 2007).

Recent reviews and survey studies (Berggren et al., 2018) have stated the information collected in experiments using multiple FER intervention methodologies and also have itemized the challenges and future directions of Autism researchers must go to enhance these methodologies.

2.2.1 Intervention Methodologies - Autism

The methodologies that have found interesting effects and results on the FER performances are itemized in the following types of intervention methodologies: a) interventions that enhance emotion recognition using a treatment wait-list with controls (Fletcher-Watson et al., 2014), b) interventions with "placebo" contact control method such as cartoon based video engines (Grossard et al., 2017), or c) interventions with no therapeutic content to bias the intervention content such as leisure groups, pizza parties, etc. (Russo-Ponsaran et al., 2018). (Berggren et al., 2018) summarized nine important studies that we will include in this chapter to illustrate how psychologists and neuro-scientists are investigating how to enhance the FER performances from key behavioral outcome measures obtained from *Pre* and *Post*, before and after the intervention. To clarify this we re-organize the Table 2.1 on (Berggren et al., 2018) adding a more summarized information about recent intervention methodologies focusing on FER tasks.

As a brief overview of the review, the intervention period documented here varies from 8 days to 20 weeks. The intervention intensity is also variable and not necessarily linked with a shorter intervention period. Most evaluations has reported by PhD students and graduate assistants.

Some of the studies included in this section proposed two types of behavioral outcome measures. One type of measure was defined to quantify the level of emotion recognition enhancement, and other type of measure related to the level of social competence. Any significant increasing effect observed for the particular outcome measure between the *Pre* and *Post* and/or an extra *Follow-up* period is considered a significant and a positive effect which supports the intervention process.

Emotion Recognition - Outcome Measures

The assessments used for extracting the emotion recognition outcome measures can be seen in Table 2.1. These include for instance the NEPSY-II affect recognition skills battery composed of 32 neuro-psychological subtests (Brooks, Sherman, and Strauss, 2009) to evaluate emotion recognition researchers used the face matching test.

Other examples are the Emotion Recognition and Display Survey (ERDS) (Thomeer et al., 2011) which include a set of face stimuli which are categorized using the CAM-C and the 35 emotion states defined by (Golan et al., 2010). The Receptive and Expressive subscales from ERDS are used for some intervention studies complementing the initial cognitive evaluation.

An intervention tool which is used broadly in recent studies not only for emotion recognition but for social competence measurement is the Diagnostic Analysis of Non-Verbal Accuracy 2 (DANVA-2) stimuli set (Nowicki, 2000). In this stimuli childrens' and adults' faces and voices are labeled in four emotions as we will explain in the sections below. DANVA-2 stimuli set is used in this study in order to elicit EEG neural activity from the child/adult emotional faces. For this particular case of intervention DANVA-2 can be quantified using the FER performance using error-rate and accuracy outcome measures (Lerner, Hutchins, and Prelock, 2011).

Table 2.1 Information in (Berggren et al., 2018) showing different types of interventions methodologies used for Autism researchers. We report here the based on Face-based tools used in the intervention, the corresponding behavioral outcome measures for FER tasks, and for Social Competence. ** Means an evident difference between the means of the *Pre* and *Post* times. NS is non-significant effect for the corresponding variables in the outcome measures columns,

Study-Authors	**Age Range**	**Intervention**	**N**	**Follow-up**	**Duration**	**Emotion Recognition**		**Social Competence**		**Comparator**
						Outcome Measures	**Differences between groups**	**Outcome Measures**	**Differences between groups**	
(Golan et al., 2010)	4-8 years	The Transporters video series	39 (30 /9)	Post	4 Weeks, 3 episodes per visit	Situation Facial-Expression matching: 1. Familiarity generalization, 2. Unfamiliriaty generalization, 3. Distant Generalization	1. 1.65 (0.90,2.39), 2. 1.66 (0.91,2.41), 3. 1.41 (0.70, 2.14)	N/A	N/A	No-intervention
(Hopkins et al., 2011)	6-15 years	FaceSay	49 (44/5)	Post	6 Weeks, 1 visit twice a week, 10-25 mins per visit	Emotion Recognition	LFA: Non-Significant, HFA: 1.43 (0.51, 2.35)**	1. SSRS, 2. SSO	1. LFA: 0.91 (0.05,1.75), HFA: Non-Significant. 2. LFA: Non-Significant, HFA: 1.23 (0.36, 2.11)	Wait-list tux painting, assistance for children in painting
(Rice et al., 2015)	5-11 years	FaceSay	31(28/3)	Post	10 Weeks 1 time per week, 25 mins per visit	NEPSY-II affect recognition Score (matching capabilities)	1.33 (0.54, 2.12)	1. SRS-2, 2. Positive Observations, 3. Negative Observations	1. NS, 2. NS, 3. NS	SuccessMaker, reading instructions from wait-list
(Ryan and Charragáin, 2010)	6-14 years	Emotion Recognition Training	30 (27/3)	Post	4 Weeks, Weekly session 1 hour	Emotion Recognition	Significant difference	N/A	N/A	Waitlist

(Silver and Oakes, 2001)	1-18 years	Emotion Trainer (ER)	22	Post	2-3 week, daily, 30 mins	1. Facial Expressions photographs, 2. Strange stories	1. Significant difference supporting intervention, 2. NS,	N/A	N/A	School Lessons
(Tanaka et al., 2010)	10.85±2.61 Treatment, 11.41±3.7, Control	Let's Face it	79 (62/17)	Post	20 hours playing the computer game/ 100 minutes per week	1. Emotion Recognition cartoons, Let's face it Skills battery	1. NS	N/A	N/A	Waitlist
(Thomeer et al., 2015)	7-12 years	Mind Reading	43 (38/15)	Post, 5 week follow up	12 weeks, 2 sessions per week, 90 mins per session	1. CAM-C faces, 2. CAM-C voices, 3. ERDS receptive, 4. ERDS Expressive	Post: 1. 1.19 (0.54, 1.84), 2. 1.05 (0.41, 1.69), 3. NS, 4. NS ** Follow-up: 1. 0.76 (0.12, 1.36), 2. 0.73 (0.11,1.35), 3. NS, 4. NS**	1. SRS, 2. BASC-2 social skills	Post: 1. NS, 2.NS Follow-up: 1. NS, 2. NS	Waitlist
(Young and Posselt, 2012)	4-8 years	The Transporters video series	25	Post	3 weeks, 3 episodes per week	1. NEPSY-II affect recognition Score, 2. Faces Tasks	1. 1.55 (0.63, 2.46), 2. 1.20 (0.34, 2.07)**	1. Peer Interest, 2. Eye Contact, 3. Gaze Aversion	1. NS, 2. NS, 3. NS	Thomas the tank engine, 15 selected episodes
(Lopata et al., 2010)	7-12 years	Summer program	36 (34/3)	Post	6 20 minutes session for 5 weeks	Danva-2 Child faces performance	NS	1. SRS, 2. BASC-2 social skills	1. 0.69 (0.00, 1.37), 2. 0.82 (0.14, 1.59) **	Waitlist
(Solomon, Goodlin-Jones, and Anders, 2004)	7-12 years	The Social adjustment enhancement curriculum	18 (18/0)	Post	training program	Danva-2 Child and Adults faces performance	NS	N/A	N/A	Waitlist
(Thomeer et al., 2012)	7-12 years	Summer program	35 (30/5)	Post	6 20 minutes session for 5 weeks	Danva-2 Child faces performance	NS	1. SRS, 2. BASC-2 social skills	1. NS, 2. 0.88 (0.18, 1.59)	Waitlist

Another example of emotion recognition intervention in ASD groups is a classic emotion recognition task for face and audio stimuli correlated with an additional or an extra part of the stimuli set as a Waitlist.

Overall the enhancement effect observed in the studies summarized in Table 2.1 varies the emotion recognition measurement across the type of intervention. In most cases the effect is not powerful enough to be considered a fully-recovered emotion recognition capabilities in ASD groups. A considerable effect is only observed to match faces or gestures. However, for the emotion recognition at least as a quantifiable outcome measure the effect can not be generalizable for any age, or ASD participant sample. This measure is not even observable for social competence as we will describe in the next subsection.

Social Competence - Outcome Measures

In Table 2.1 we can observe multiple different variables which are proposed to measure not only the amount of social interactions or the enhancement in social communication (White, Keonig, and Scahill, 2007).

In recent studies the social competence outcome measures are more sensitive markers to show a FER performance improvement in ASD population (Marino et al., 2019). As we mentioned in the previous section the face/emotion perception is more affected by motivational deficits and ASD individuals should receive intervention using emotional reinforcement.

Social competence outcome measures are not only very sensitive for face/emotion perception and recognition, but these variables can be considered important markers to measure the intervention efficacy for the social competence and the correlated emotion recognition tasks (Lee et al., 2018).

Although most FER intervention studies do not include social competence assessments. A social competence assessment included in the studies is the Social Skills Rating System (SSRS) (Gresham et al., 2011). The SSRS has a social skills evaluation subscale with 40 items. 10 items refer to Cooperation, 10 items to Assertion, 10 items to Responsibility, and 10 items to Self-Control all items oriented to social interactions in school, relationships with peers, and assertive behavior.

A complement for SSRS is the Social Skills Observation (SSO) which is a psychological protocol described by (Hopkins et al., 2011) as empirical social competence evaluation. The SSO methodology consists in a 2 hour practice session evaluated by two blind psychologist trained for the assessment. During the two assessment hours ASD individuals have to perform a peer interaction and meanwhile the trained reviewers annotate the performance. The psychologist/annotators should achieve a 90% of inter-rater reliability for each item in the assessment.

A widely evaluated social competence assessment is the Social Communication Questionnaire (SCQ). This evaluation was constructed initially for Autism screening as a Autism Screening Questionnaire by (Rutter et al., 2007) based on the old version of Autism Diagnostic Interview old version, nevertheless the SCQ is used for social competence evaluation from the primary caregiver focusing on reciprocal social interaction, language and communication, and stereotypes behaviors (Chandler et al., 2007).

Despite the reliability of the caregiver-based questionnaire, other social competence outcome measures are used in FER-based intervention for ASD individuals. The Social Responsiveness Scale (SRS) (Constantino, 2013) is a 65-item rating scale that measure the Reciprocal Social Behavior (RSB) or the quality of reciprocal social interaction that with ASD children involve in. This scale evaluate the severity of three possible social competence deficits such as social, language, and stereotypic behaviors.

Recently, other assessment have not only included clinical and behavioral variables in the evaluation but the adaptive responses found in children with ASD when a repetitive protocol is presented.

The Behavior Assessment System for Children - 2 (BASC-2) measures not only diagnosis aspects but also behavioral and personality aspects that can affect profusely the emotion recognition capabilities (Volker et al., 2010). BASC-2 was developed with the purpose of assisting the diagnosis of Autism using the DSM-IV. The initial part of the assesment can be filled by the participant or the reviewer, but the Parent Rating Scale (PRS) reports the scores to evaluate adaptive behavior at home and school as well as in the community. All the PRS included in the BASC-2 are composed of 150 items including subjective behaviors evaluation such as Aggression, Anxiety, Attention Problems, Atypicality, Conduct Problems, Depression, Hyperactivity, Somatization, and Withdrawal, and for each behavior BASC-2 defines five adaptive scales such as Activities of Daily Life, Adaptability, Functional Communication, Leadership, and Social Skills.

As overall summary across the social competence outcome measures used in the intervention studies are shown in Table 2.1. If the stimuli set is not adequate for social skills measurement the studies are not reporting significant effects between the *Pre* and *Post* treatment spots. The important effect observed in the intervention studies was in the BASC-2 whole score. The SSRS and the SSO show a small difference effect when the population is divided in low-functioning, and high-function groups (Hopkins et al., 2011), but not with the entire ASD group is taken into account for the evaluation.

The null effect observed in social competence outcome measures for most FER intervention studies is a critical indicator of a multidomain deficit associated with the face/emotion processing. **The neural outcome measures mentioned above such as N170, N300, P400,**

LPP, and the frequency de-synchronizations are not only related with a deficit in early face processing. Additionally, the posterior emotion recognition is also related to behavioral deficits, but to a lack of motivation and social skills including subscales such as reciprocal social interaction, and atypical behaviors.

Intervention Generalizability

In the complete set of interventions described in Table 2.1 we can enumerate the reasons supporting the systemetic lack of assessment generalizability as follows:

1. The difficulty in creating a long-term follow-up process in intervention studies due to a incremental cost and patients' eligibility criterion.
2. The studies are not generalized enough to attribute a positive effect of the current interventions methodologies into the autism social skills improvement.
3. With a combination of emotion recognition and social competence oriented intervention it is possible to obtain small effects in social skills and emotion recognition improvements.

In future studies the interventions methodologies should be modified in order to affect emotion recognition and social skills outcome measure in an integral way. **In this study we propose to assist the current emotion recognition intervention methodologies using Deep classifiers extracting the important information from a successful emotion decoding from EEG single-trials.**

Clinical Implications - Intervention

Due to the high prevalence of ASD across multiple demographics, the demand of interventions for emotion recognition has increased considerably. New technologies and resources should be included not only to reduce the time and cost to reproduce an emotion recognition intervention, but also to automatize and make the assessments easier for the participants and reviewers.

The difficulty of having statistical generalizability should be taken into account for the design of future behavioral interventions. The difficulty for finding improvement effects across multiple ASD groups make the decision for changing the intervention methodologies very challenging.

For future implementations of emotion recognition interventions for ASD individuals the decisions and preferences of the parents and caregivers are strong biases for finding significant and generalizable effects.

A double-blind for reviewer and caregiver, or a randomized effect should be introduced into the participant sample to increase the sample generalizabity and make the effects stronger in the intervention evaluation.

In the next chapter, we will complement this clinical, neural, and behavioral studies background clarifying the corresponding Autism deficits for face/emotion processing with the implementations of previous Machine Learning (ML) systems on Controls and individuals with Autism. We will introduce the ML systems dedicated to infer and learn from neural features to decode emotion and other outcome measures without including ASD individuals samples.

Chapter 3

Multimodal Neural based Emotion Decoding for ASD and non-ASD individuals

ML Classifiers based on neural features have become a new tendency in the ML state-of-the-art research. SVM, LDA, and different Gaussian Mixtures approaches have been proposed from the initial analysis of EEG single-trial classification (Blankertz et al., 2011; Pfurtscheller et al., 2006). However, the classification model is not a single piece pipeline, and for bio-signal and events classification pre-processing, artifact removal, and statistical transformations are important for an adequate EEG single-trial classification for in-vitro, and/or in-the wild data acquisition (Vaid, Singh, and Kaur, 2015).

Most EEG-based single trial classification studies for emotion decoding have included only non-ASD participants because of the flexibility of non-ASD participants and the complexity of finding formally ADOS-2 or DSM-IV diagnosed participants. This limitation is common in other neurodevelopmental disorders. Some related clinical studies have certain limitations when a purely engineering team is trying to obtain results from brain signals (Govindarajan and Kumaravelu, 2019).

Another important difficulty for EEG-based classification pipelines no-SOA is to find an adequate SNR, and an adequate synchronization between the stimulus and the neural signal itself (Mühl et al., 2014). Depending on the protocol study, it is not easy to find a clinical formal study with formally time-locked stimuli including only non-ASD participants.

There are multiple wireless BCI platforms such as Emotiv EPOC, gtec Nautilus, and Enobio Neuroelectrics (Debener et al., 2012). These and multiple other devices are included in EEG-based behavioral decoding with purpose of making EEG-based pipelines more flexible and enrich no-SOA data acquisition environments. However, these devices are not precisely included in lab-controlled environment where the preferable devices have an increased SNR, Total-Harmonic-Distortion (THD), and a Common Mode Rejection Ratio (CMRR) (Mora, De Munari, and Ciampolini, 2015).

The Deep ConvNet based pipeline proposed in this project is studied in depth in the next sections in order to understand not only if the neural information features can decode successfully the emotion or the behavioral class, but how the Deep ConvNet is weighting the importance of these particular input features. We include the evaluation of the most reliable saliency maps in the current ML state-of-the-art. With this analysis we are not only finding a complete and intact neuromechanism for emotion processing, but which time ranges and electrodes are important got ASD and non-ASD groups to decode successfully emotions from the corresponding neural activity.

In the following subsections we will describe in detail the most important studies covering Deep and shallow ML implementation for behavioral, especially emotion recognition, on ASD and non-ASD participant samples.

3.1 EEG-based classifiers Emotion Recognition - non ASD

For non-ASD or neurotypical controls it is very easy to find multiple studies using EEG for evaluating emotion recognition, dividing the classes inside the affective circumplex arousal/valence Russell's axis (Gerber et al., 2008) and recognized 21 emotion categories/states across the circumplex axis currently defined in a neural study oriented to ASD groups (Baron-Cohen et al., 1999).

Since 2000s, a new strategy was stated to decode emotion, emotion states, and circumplex states using neural features emerged (Jenke, Peer, and Buss, 2014; Mühl et al., 2014). However, most of the pipelines constructed in the initial research wave using EEG-features did not clarify the incidence of artifacts, device distortion sources, and bad channels (Lotte et al., 2018). After the pre-emphasis and artifact removal techniques have been debugged and complement for a fair neural representation of a EEG emotion-elicited single-trial (Delorme and Makeig, 2004).

Table 3.1 summarizes all the methods for the pre-emphasis processes, the number of participants, the artifact removal procedures, Supervised/Unsupervised classifiers, the cross-validations modalities used in the evaluation, and the datasets included in the most relevant EEG-based emotion/behavioral decoding studies for non-ASD groups. For these particular studies we won't find diagnosis prediction, and additional outcome measures such as social competence, and social responsiveness questionnaire.

For ASD participants we divide the multimodal pipelines in 1) pipelines designed for ASD diagnosis and/or Early diagnosis, and 2) pipelines designed for EEG-based or behavior outcome measures prediction including ASD groups which are fewer in the current literature. To illustrate the most recent and relevant studies for these two categories we report the most important pipelines for ASD diagnosis in Table 3.2, and the pipelines constructed for ASD emotion or behavioral outcome measures recognition in Table 3.3.

In the following subsections we will summarize and analyze the drawbacks and advantages observed on the recent most relevant implementations for EEG-based emotion recognition pipelines including and without including ASD participants as well as the methodologies used for pre-emphasis, artifact removal, and cross-validation modalities used for classification.

Study-Authors	Dataset /Acquisition	N	EEG type of data	Classes	Pre-Emphasis	Artifact-Removal	Features used	Classifier	Cross-validation	Performance
(Jirayucharoensak, Pan-Ngum, and Israsena, 2014)	DEAP - EEG signal 40 videos	all 32 subjects, 40 trials, 1 minute trial per subject.	long-trials watching emotional videos, online tagging	2-class high-low Arousal/Valence levels	Downsampling from 512Hz to 128Hz, Filter for 5 frequency bands theta, lower alpha, upper alpha, beta, and gamma PSD 128 samples	None	PCA Covariate Shift Adaptation - 50 best components per window	A 3 layer Deep Neural Nework (DNN), with a final Softmax layer - 50 units hidden-layers. Fine-Tuning and Backpropagation trained using the 50 PCA features-per window, and per trial. SVM used as baseline	Leave-one-subject-out (LOSO) - Following DEAP original paper modality	Accuracy -> Arousal : 52.01%, Valence: 53.34%
(Mehmood, Du, and Lee, 2017)	Data collection from IAPS emotional scenes. 180 stimuli shown per subject. 45 trials x 4 emotions	21 subject, 9 male, 11 female	Emotiv - Epoc 1.5 s time-locked trials. LPP and P300 based	4 emotions (happy, calm, scared)	Filtering, and Hjorth parameter estimation	EEGlab plug-ing for EOG artifact removal from (Gómez-Herrero et al., 2006)	Hjorth parameters: Activity, Mobility, and Complexity. These three featues are calculated per band δ, θ, α, β, and γ. 360 epochs per emotion were grouped	LDA, KNN, SVM, DNN, Bagging, Boosting, Random Forest, and Naive Bayes. Majority Voting between all the classifiers decision is applied. 360 features used per trial no PCA	10-fold cross validation across all subjects	Accuracy For Deep Learning -> 73.66% and Majority Voting -> 76.65%
(Mehmood and Lee, 2016)	Data collection from IAPS emotional scenes. 180 stimuli shown per subject. 45 trials x 4 emotions	21 subject, 9 male, 11 female	Emotiv - Epoc 1.5 s time-locked trials. LPP and P300 based	4 emotions (happy, calm, scared)	Filtering between 0-50 Hz	ICA and pop_eegfilt applied per trial to remove artifacts EEGlab	3 features from the LPP amplitude on three LPP windows Early, Middle, and Late. These features are grouped for the 5 rhythms δ, θ, α, β, and γ. All the windows ranges and all the frequency bands feature combination are used to train.	A linear kernel SVM, and a KNN (K=5) classifiers	10-fold cross validation across all subjects, and leave-one-subject-out cross-val	Accuracies: For LPP early window in LOSO SVM -> 57.6% θ, KNN-> 58.0% α. For the 10-fold cross-val LPP early shows better performance -> 59.12%
(Li et al., 2018)	4 min length 6 music videos selected from Youtube Chinese films. Divided in 12 sessions per subject. The SEED Database	15 subjects (7 male, 8 female) they filled Eysenck Personality Questionnaire (EPQ).	ESI Neuroscan - 4 mn trials 45s rest between each video presented	Meanwhile the users are watching the video, each stimuli is labeled between 3-classes positive, neutral, and negative	None	None	A grouped sparse canonical correlation analysis (GSCCA) is applied to the raw EEG signal. An a Bi-hemispheres adversial Neural Networks (Bi-DANN) semi-supervised feature extractors	TCA, KPCA, TPT, and Bi-DANN optimization one-regularization, and two regularization layers R1, R2.	9 sessions for all the subjects for training, and 3 remanent sessions for test. Baseline for the SEED database. And also a Leave-One-Subject-Out (LOSO) cross-validation is applied	Accuracies: For Bi-DANN in a SEED baseline was 92.38 ±7.04%. And for the LOSO cross-val 83.28 ±9.61%
(Zheng et al., 2015)	4 min length 6 music videos selected from Youtube Chinese films. Divided in 12 sessions per subject. The SEED database	15 subjects (7 male, 8 female) they filled Eysenck Personality Questionnaire (EPQ)	ESI Neuroscan - 4 min trials 45s rest between each video presented	Meanwhile the users are watching the video, each stimuli is labeled between 3-classes positive, neutral, and negative	Bandpass filtering between 0.3 and 50 Hz	None	Differential entropy (DE) features for each video trial from five different EEG bands δ, θ, α, β, and γ. Different feature set from the DE feature set, and DE assymetric features set such DCAU, DASM, and RASM	A set fronto temporal channels are selected for performance evaluation. For a SVM C=1 and linear kernel, and a 3-layer Deep Belief Network (DBN) 4,6,9,and 12 channels DE features are giving for training	9 sessions for all the subjects for training, and 3 remanent sessions for test.	For 12 channels and all the DE features from the whole frequency bands together the accuracy was the best 86.85 ±2.99%. For the DBN grouping all the bands is the best accuracy 86.07 ±1.47%

(Koelstra et al., 2010)	70 candidate music videos taking from subjective selection by participants, 20 videos selected by participants. Russel's axis is defined on the three 4 subplanes LAHV/HAHV, LAHV/LALV, LALV/HALV, HAHV/ HALV, plus a neutral class	6 participants, selected 20 test videos	Active Two Biosemi. Recorded at 256 Hz. A 30 s long trials recorded per stimulus	3 2-class problems. High-low arousal, high-low valence, and like-dislike	band-pass between 0.3 and 35Hz	None	PSD features are extracted with the Welch method extracting features from a 3Hz window. CSP filter features are extracted each 3 Hz with a 50% overlap. PSD, and CSP features are originally divided in the five most used EEG frequency bands δ, θ, α, β, and γ	A linear kernel SVM C=1	Leave-one-trial-out (LOTO) per subject cross-validation	Accuracies: For Valence: PSD -> 58.8%, CSP-> 58.8%. For Arousal: PSD -> 51.9%, CSP-> 55.7%. For Liking: PSD->49.4%, CSP-> 48.8%
(Petrantonakis and Hadjileontiadis, 2011)	60 Ekman face pictures presented for 5 seconds, and rated with 6 different emotion categories. (10 per emotion)	16 Healthy participants (9 male, 7 female)	g.MOBIlab engineering, Guger Technologies, portable biosignal acquisition system four EEG bipolar channels, filters: 0.5–30 Hz at 256Hz	The 6 Ekman's basic emotion states happiness,surprise, anger, disgust,sadness, and fear	band-pass for letting pass alpha and beta rhythms only 5-15 Hz	Signal Averaging (with the same trial)	Wavelet-based features, High-Order-Crossing (HOC), and additional statistical features per 0.1s window	HOC-Emotion Classifier (HOC-EC), and linear SVM, Mahalanobis Distance (MD), and a QDA. For HOC-EC authors separated the features for a single channel and combined channels features	Leave-out-n-out cross-validation. Randomly separated 100 groups of trials	Classification rate measure is reported: HOC-EC /QDA-> 62.03% single-channel, HOC-EC/SVM -> 83.33%
(Jenke, Peer, and Buss, 2014)	8 scene emotions 5-sec long trials for each 5 emotions per subject , taken from IAPS database. IAPS scenes are labeled using SAM and the emotion labels.	16 subjects (9 male, 7 female)	64-channel EEG cap with g.tec gUSBamp recording at 512Hz, with an initial filter between 0.1 and 100 Hz	5 emotions rated by the authors on the IAPS database: happy, curious, angry, sad, and quiet.	50Hz Notch Filter	None	A total of 22881 features were grouped per trial: 448 statistics, 128 Hjorth parameters, 64 Non-Stationary Index, 64 Fractal Dimension, 640 HOC, 3264 STFT, 4096 HOS, 320 HHS power, 192 DWT bior3.3 , 192 DWT db4, 12096 MSC estimate, 277 diff/derivative assymetry , and 232 radio assymetry . From this huge amount of features 4 Feature Extraction Methods were proposed mRMR, ReielfF, ES f2, ES γ, ES θ.	QDA evaluated per feature selection method	Leave-one-stimuli-out using a 8 fold over the training set for the feature extraction	Best Accuracies: ES γ -> 36.38%, ES θ -> 36.68%

(Frantzidis et al., 2010)	500ms long IAPS selected trials. 40 trials per subject and online annotation.	112 subjects healthy (56 male, 56 female)	Neurobehavioral Systems, Albany, CA Epoched time-locked 500 ms stimuli, EEG recorded at 500Hz	2-class dividing the Russels axis in a four different binary problems LAHV/HAHV, LAHV/LALV, LALV/HALV, HAHV/ HALV. And each axis separately for a 4-class problem.	Bandpass IIR butterworth filtering between 0.5 and 50 Hz.	EOG artifacts were removed using a LMS adaptive filter per trial	DWT , ERP oscillations were calculated using the five EEG bands δ, θ, α, β, and γ, and only the γ band included into 50 Hz. With SVMattribute value only 2 features are selected per trial one for amplitude and one for latency.	Mahalanobis Distance, linear SVM, a Polynomial SVM, Radial basis SVM	10 Fold cross-validation across all trials and all subjects	Accuracies for linear SVM best 2-class: LAHV/HAHV -> 89.3%, LAHV/LALV -> 92.9%, LALV/HALV -> 82.12%, and HAHV/ HALV ->100%. For the 4-class problem: Maha -> 79.46%, and linear SVM -> 81.25%
(Gao, Lee, and Mehmood, 2015)	Data collection from IAPS emotional scenes. 180 stimuli shown per subject. 45 trials x 4 emotions	21 subject, 9 male, 11 female	Emotiv - Epoc 1.5 s time-locked trials. LPP and P300 based	4 emotions (happy, calm, scared)	Filtering, and Hjorth parameter estimation	EEGlab plugging for EOG removal from (Gómez-Herrero et al., 2006)	Amplitude features from six critical channels Fpl, Fp2, C3, C4, F3, and F4	3 layers Semi-supervised Deep RBM, and SVM, KNN, and ANN baselines.	For RBM intra-subject 120 trials for training and 60 for test for each subject, and other two cross-vals 11 subjects for train, 10 for test, and channel selection	For Deep RBM Accuracies: intra-subject -> 68.4%, 11/10 -> 28.67%, and channel selection -> 57.2%
(Wang, Nie, and Lu, 2014)	The movie clips set includes six clips for each of two target emotional states: positive and negative emotions. Each movie clip duration is 4 minutes, with a 45 minutres for SAM	6 healthy volunteers (3 male, 3 female)	ESI-128, NeuroScan Labs, SCAN 4.2 software, and a modified 64-channel QuickCap. 4 minutes EEG trial length	2- class problem positive and negative video clip classification using features for the entire EEG trial.	Downsampling from 1000Hz to 200Hz. Trials with EOG/EMG components are removed manually	Features are smoothed using LDS, but nothing is applied to the signal or the entire signal distribution	PSD features extracted from δ, θ, α, β, and γ, and 27 asymmetry indexes from 27 pair of channels. Wavelet features grouped in 200 features for all the EEG bands. 200 features covers the decomposition and approximation coefficients, and the corresponding entropy level. And Non-linear dynamic feature (NLD) and a new entropy features estimation. LDS is used to calculate the entropy and the spectrum estimation cleaner	PCA, and LDA are used for feature reduction and a set of 3 SVM classifiers are used here: a linear, a polynomial, and a radiabasis kernel. The C value is estimated per fold.	10-Fold cross-validation across all trials and subjects	Linear SVM shows the best accuracies: PSD without LDS -> 87.53%, and PSD with LDS -> 87.53%. PSD asymmetry features -> 82.38%. Wavelet Features -> 78.41%. And NLD features -> 71.38%. Emotion trajectory estimated, in direction but not amplitude.
(Zheng et al., 2018)	168 movie clips separated tagged in 4 emotions. And 72 movies were selected between the subjects. Each EEG trial is 2 s and a 45 s after watch them is taking for self-assessments	44 participants (22 male, 22 female)	The Emotion-Meter hardware is composed of a SMI-ETG eye-tracking glasses, and 6 symetrical electrodes T7-T8, FT7-FT8, and TP7-TP8	Each video is labeled with one of 4 emotion: happy, sad, fear, and neutral	A bandpass filter betwen 1-75Hz. And EEG and eye-tracker data is resampled from 1KHz to 200Hz	Non-linear Dynamic Syste, but nothing is applied to the signal	PSD and DE Features are grouped for all the 5 EEG important bands and for the 6 channels δ, θ, α, β, and γ	A bi-modal Deep Autoencoder (BDAE) was used to encode EEG and eye-tracker features. A subsequent stacked RBM into the BDAE is used to reconstruct the features and classification.	1 session for train - and one session for test. Selected a-priori	Accuracies: Eye-tracker Features -> 67.82 ±18.04%, EEG -> 70.33 ±14.45%, and Feature Fusion ->75.88 ±16.44%

(Zhang et al., 2018)	4 min length 6 music videos selected from Youtube Chinese films. Divided in 12 sessions per subject. The SEED database and SEED-IV	15 subjects (7 male, 8 female) they filled Eysenck Personality Questionnaire (EPQ)	ESI Neuroscan - 4 min trials 45s rest between each video presented	Meanwhile the users are watching the video, each stimuli is labeled between 3-classes positive, neutral, and negative	A bandpass filter betwen 1-75Hz. And EEG and eye-tracker data is resampled from 1KHz to 200Hz	None	DE features taken from 5 EEG important bands and for the 6 channels δ, θ, α, β, and γ For all the 62 channels.	TCA, KPCA, TPT, and DANN optimization are implemented here, and Domain-Adversarial Neural Network is complemented using a final two layers with Maximum Mean Discrepancies (MK-MMD) blocks.	Leave-one-subject-out (LOSO) - Complementing the SEED dataset cross-val	Accuracies: DAN -> 83.81 ±8.56%, and DANN -> 79.19 ±13.34
(Soleymani et al., 2015)	20 videos excerpts chosen from MAHNOB-HCI database. 14 out of 21 were taking fro movies. Clips were taken between 34.9s to 117s . Data collection and labeling has a Cronbach alpha= 48.47%	28 healthy volunteers comprising 12 male and 16 female	EEG signals were acquired from 32 active electrodes on 10-20 international system using a Biosemi Active II device	Continuous labeling with the SAM scale -5 to 5 for Arousal and Valence levels across all the video length.	Band-pass Filter for Alpha, Beta and Gamma bands	The artifactual trials are rejected calculating the EEG bands correlation R-squared with the facial-movements. The Granger causality is calculated from the 49 facial points.	PSD features from 4 EEG bands such as θ, α, β, and γ. 32 electrodes x 4 bands features 128 features	A Long Short Term Memory Neural Network regressor, based Multilinear regression techniques applied to continuous labelling	10 Fold cross-validation across all trials and all subjects	Pearson Correlation Coefficient (PCC), and Root-Mean-Square Error (RMSE) metrics are reported: For EEG only PCC -> 0.24 ±0.34, and RMSE: 0.053 ±0.029. With feature fusion PCC-> 0.40 ±0.33, RMSE -> 0.047 ±0.025
(Liu, Zheng, and Lu, 2016)	DEAP - EEG signal 40 videos, and SEED Dataset videos 4 min length 6 music videos selected from Youtube Chinese films	all 32 subjects, 40 trials, 1 minute trial per subject for DEAP, and	ESI Neuroscan - 4 min trials 45s rest between each video presented for SEED, long-trials watching emotional videos, online tagging for DEAP	2-class high-low Arousal/Valence levels - For DEAP, and 3-classes positive, neutral, and negative for SEED	Filtered are applied as SEED and DEAP papers suggest. Additional filtering is not added in this study	None	PSD and DE Features are grouped for all the 5 EEG important bands and for all the channels δ, θ, α, β, and γ	A bi-modal Deep Autoencoder (BDAE) was used to encode EEG and eye-tracker features. A subsequent stacked RBM into the BDAE is used to reconstruct the features and classification. A DBN is used for DEAP, and BDAE for SEED dataset.	LOTO per subject is used for DEAP, and 9 sessions for train and 3 test is used for SEED	Accuracies: DE features: For SEED -> 91.01 ±8.91%. And for DEAP -> 85.2 Valence, 80.5 Arousal, 84.9 Dominance, 82.4 Liking
(Zhang, Ji, and Zhang, 2016)	DEAP - EEG signal	all 32 subjects, 40 trials, 1 minute trial per subjec	long-trials watching emotional videos, online tagging	2-class dividing the Russels axis in a four different binary problems LAHV/HAHV, LAHV/LALV, LALV/HALV, HAHV/ HALV.	Downsampling from 512Hz to 128Hz, Filter for 5 frequency bands theta, lower alpha, upper alpha, beta, and gamma PSD 128 samples	None	Empirical Mode Decomposition Features (EMD) - only channels are used per subject, and calculated across the LOSO cross-val. For each 2 channel Intrinsic Mode Functions (IMF) features	linear kernel SVMl	Leave-one-subject-out (LOSO) - Following DEAP original paper modality	Best Accuracy SVM -> 93.06%

(Li, Zhang, and He, 2016)	15 movie clips of 240s or 4 minutes from SEED dataset are shown in online self-assessment task	4 healthy participants	long EEG trials, separated in six emotions from SEED dataset	Each stimulus is labeled between 3-classes positive, neutral, and negative	Downsampling from 1000Hz to 200Hz, and a bandpass filter between 0.3 to 50 Hz	None	256 points STFT δ, θ, α, β, and γ, and DE maps calculated per trial using Shannon entropy models. The DE maps are organized in 2D using an scalp representation of 62 EEG channels. This organization is called as a Sparse map and it is one per frequency band as image channel	A Hierarchical 2-convolutional 2 max-pooling layers layer ConvNet and a final fully-connected with sigmoid activation function. The convnet is trained using a batch-size of 50 and a learning rate of 1.	Three valiation options : A. 3 Fold cross-validation 200 repetition intra-subject. B.Leave-one-subject out (LOSO). C. Using LOSO for pretraining, and intra subject for fine-tuning	The best accuracy is found using the γ. For A -> 88.2±6.66%, for B->37.10±1.91, and for C->80.01±9.09%

Table 3.1 The most relevant and recent studies including non-ASD participants where Computer Scientists and Engineers have evaluated EEG-based emotion recognition. Long and short time-locked trials are included here.

Study-Authors	Features	# Classes	N	Performace	Age-group
(Bosl, Tager-Flusberg, and Nelson, 2018)	DB4 Wavelet decomposition for 30 sec segments, 256 samples per second EEG resting state, Recurrence Quantitive Analysis (RQA), RR, DET, LAM, ENT, longest diagonal, mean diagonal length, trapping time (TT), First Poincare recurrence, Second Poincare recurrence, and Sample Entropy and Detrend Fluctuation Analysis	3 ASD, Low Risk Controls (LRC), and High Risk of ASD (HRA) classes based on family recruitment group sibling. ADOS was used to differentiate 2 no ASD-ASD binary classification, and Calibrated Severity Score 1-10 (CSS) was estimated	188 divided in LRC, HRA and ASD group. 89 LRC, 99 HRA, and the rest ASD. ADOS-CSS cut-off is 4 to identify ASD and non ASD participants	LRC and ASD are used for train, and HRA is group is used to test. And all the subjects together modalities. Leave one subject out cross-validation and SVM classifier for both ASD and non ASD problem, and CSS estimation. More than 95% sensitivity, specificity, and PPV were obtained for mostly all age groups from 3-36 months	150 participants received 36 months, 23 received 24 months, and 15 18 months age groups
(Heunis et al., 2018)	5 sec segments 10 RQA features RR, DET, LAM, and Poincare Recurrence. Another set of PCA multidimensional features were extracted, Kolmogorov–Smirnoff test feature selection, and Mutual Information features were also added in the feature set	Differentitating ASD and TD groups	46 TD, and 16 ASD subjects	Leave-one subject out classification, and 10-fold cross-validation modalities. 70-30 train-test split is used for the full range sample, and the second subsample. The 10-fold and leave-one out crossvalidations are used for the age-matched sample. 92.9% Accuracy 100% sensitivity, and 87.5% specificity for ASD-TD problem classification (leave-one-subject out). LDA, MLP, and SVM classifier comparisons. MLP shows the best accuracy for the full-range sample, and the SVM shows the best performances for 0-6 years subsample, and the age-matched subsample.	Analysis is subdivided in three age-groups 0-18 years, 0-6 years, and 2-6 years being this latter and age-matched sample
(Castelhano et al., 2018)	Time-frequency values from 10-90Hz and from 9 different categories of images and 45 images per each category. Photographic, Schematic, and Mooney, and three possibilities for each upright, inverted, and scrambled.	Classifying TD controls and ASD groups individuals	29 TD controls, and 10 high functioning ASD. 10 out of 29 controls were selected for classification analysis	85% Accuracy, and 94% AUC, for photographic face features the best ROC was obtained reaching an AUC of 98.6%. All these performances were evaluated for a linear hyperplane SVM. The best regularization parameter was calculated in a 3-fold cross-validation modality.	ADOS and Autism Diagnostic Interview-Revised (ADI-R) were used to screen ASD participants, and they performed the Welcher's Intellegence Scale assesment. All subjects were adults with a mean of 23.4 years for controls and 23.1 for ASD.

(Eldridge et al., 2014)	ERPs obtained from oddball task dae and daa phonemes elicited for 340 ms long. a) First features were extracted from sum of signed differences (SSD) computing median differences between deviant and standard stimuli, b) Second, time variance of the standard stimuli response, c) Third, calculation of modified Multiscale Entropy (mMSE)	Classifying TD controls and ASD groups individuals	19 ASD, 30 TD subjects same age peers	79% Accuracy obtained with mMSE+first and second feature sets using a Naive Bayes Classifier, Naive Bayes Classifier also shows a maximum for the first and second feature sets 66%. and linear SVM shows a maximum for mMSE 0.69. A threefold cross-validation modality was used for the evaluation.	ADOS-CSS and Childhood Autism Rating Scale (CARS) are obtained to separate ASD group. Adults were included in this evaluation
(Pistorius et al., 2013)	5s second resting state EEG 9 RQA features, and PCA dimension reduction features. A two-way ANOVA analysis was implemented to select features that are p<0.05 between TD/ASD groups in the training set	Classifying TD controls and ASD groups individuals	7 TD and 5 ASD	kNN k=3 and LDA are evaluated as possible claasifiers. kNN and LDA perform equally well with the EEG segments after artifact removal process. An accuracy of 83.3%, a sensitivity of 85.7%, and specificicity of 80.0% 10/12 subjects were correctly classified in a leave-one-subject-out crossvalidation modality.	ASD group was screening based on ADOS, ADI-R and Kaufman Assessment Battery for Children. Age was distributed between 8-17 years
(Bosl et al., 2011)	20 sec EEG resting-state segments were collected, subsequently a modifiefd multiscale entropy (mMSE) is calculated detecting longer-range correlation between multiple time ranges scales. Features for the time asymmetry are calculated for the EEG signal representing the number of irreversible time points for each signal excerpt	Classifying TD controls and High Risk ASD (HRA) groups individuals	79 subjects divided into 46 HRA, and 33 TD controls, 143 sessions and a age-group between 6-24 months were included	The mean mMSE shows an constant difference around all the channells between TD and HRA groups and across all the age groups. 192 values feature vector was used as input vector to feed three classifiers k-NN, SVM, and Naive Bayes. All infants are also evaluated for boys and girls only. A 10-fold crossvalidation modality is executed for each age group. A maximum accuracy 0.9 k-NN and 1.0 SVM is observed for boys in 9 months old and for 18 months 0.9 k-NN and 0.9 Naive Bayes.	TD and HRA participants were screened based on Infant Sibling Project study, Participants cover the age range between 9-24 months

(Jamal et al., 2014)	1000ms ERPs are extracted from the emotional and neutral faces elicitation, they feed the following pipeline: 1) a Continuous Wavelet Transform (CWT) for each channel is calculated and the instantaneous phase difference between each pair of channels is computed, as well as the frequency representation, 2) a k-means clustering is applied across all the channel combinations, to find phase connectivity patterns. 3) Depending on the number of clusters 3 brain synchrostates are calculated from the k-NN clustering, a phase synchronization index is calculated between all the channels and for each synchrostate. 4) A synchronization feature selection is processed using modularity, transitivity, characteristic path length, radius, and diameter per calulated graph connectivity between the channels. 5) 36 features were collected for 2 max-min states, 3 stimuli fear, neutral, happy, and the 6 parameters of the graph.	Classifying TD Controls and ASD groups individuals	24 participants, 12 TD and 12 ASD, processing 4 blocks of 10 happy, 10 neutral and 10 fearful faces	LDA, QDA, SVM linear to 4 polynomial order classifiers were evaluated using a and for three different cases of features, 1) the whole feature set for maxium and minimum synchrostates, 2) maximum, 3) minimum other 6 cases were evaluated using only a particular parameter from the 36 feature set parameters. A maximum performances is obtained in case 1 and 2 for linear SVM obtaining 94.7 accuracy, 85.7 sensitivity, and 100% specificity. The modality used for the evaluation is a leave-one-observation-out grouping all the observations per subject and excluding each them individually.	ASD group cover an age group between 6-17 years, with a mean of 10.2 years
(Ingalhalikar et al., 2014)	Magneto-encephalographic (MEG) and Diffusion Tensor Imaging (DTI) features were used. Two auditory experiments were executed: The first is a binaural sinousoidal wave at 45dB to detect the auditory latency called M100. The second is an oddball task between a deviant tone and two vowels tokens detecting a change in the Superior Temporal Gyrus (STG) called Magnetic Mismatch Field (MMF)	Classifying TD/ASD participants and classifying ASD/LI- / ASD/LI+ classes as separate problems	135 participants, were included in this study 42 TD controls, 57 ASD/without language impairment, 36 ASD/with language impairment. 55 out of 135 complete the MEG data	MMF+M100+DTI features were grouped together producing the highest accuracy feeding a LDA classifier. 5 Fold cross-validation is use for a classifier using fuision data, and single fusion data. For ASD/TD classifier the 5-fold cross-validation the authors obtained 83.3% accuracy, 72.9% sensitivity, and 86.1% sensitivity. Fusion performance on the new subjects gives 87% accuracy classifying well 20/23 subjects. For ASD/LI-/ASD/LI+ classifier the 5-Fold accuracy was 70.1%, 66.6% sensitivity, and 76.6% specificity.	ASD and TD were screened using ADOS and ASD/LI- and ASD/LI+ using Clinical Evaluation of Language Fundamentals(CELF-4). TD age mean is 10.4 years, and for ASD is 10.1 years.

Table 3.2 Recent and relevant studies focused on ASD diagnosis or Early ASD diagnosis using ML-based pipelines, and using neural outcome measures as inputs feature-set .

Study-Authors	Features	#Classes	#Subjects	Performance	Age-group
(Fan et al., 2017b)	EEG signals were filtered between (0.2-45 Hz). Each segment was cleansed with a 50% overlap. Artifacts were removed with EEGlab plugin (Gómez-Herrero et al., 2006) and slew-rate threshold removal. Features are extracted for a 60 mins long visit. Statistical, HOC, HHS, and HOS features were grouped per trial. Features were calibrated using this expression $F_c = \frac{f-f_{low}}{f_{high}-f_{low}}$.	Identify mental/affective states on driving skill training environment. Affective States and mental workload are quantifying on the experiment. Engagement, enjoyment, frustration, and boredom. Observer rate the mental/affective state and the workload with scale between 0-9. Workload was rated in binary way high-low intensity	20 ASD (19 males, 1 female) diagnosed using the ADOS-2. Each subject developed 6 visits comforming 120 sessions, and 111 were processed	A nested cross-validation was applied. In the outter loop it is a leave-one-subject out, and in the inner loop is a 10-fold were the F-value is calculated for each feature intra-subject. A kNN classfier is used to reduce the classification noise. A best average performance was obtained grouping all the features hyperparameters with macro F1 score of 0.90.	The age of the ASD group has a mean of 15.69±1.65, and a ADOS total raw score of 13.56±3.67
(Fan et al., 2017a)	EEG signals were filtered between (0.2-45 Hz). Each segment was cleansed with a 50% overlap. Artifacts were removed with EEGlab plugin Gomez-Herrero and slew-rate threshold removal. Features are extracted for a 60 mins long visit. Statistical, HOC, HHS, and HOS features were grouped per trial	Facial Affect Recognition task from a emotional avatar, 28 trials are presented to the participants. The social expression is shown for 3s after a social vignette. 4 emotions presented joy, sadness, surprise and neutral	8 high-functioning ASD male participants diagnosed using ADOS-2	5-fold stratified cross-validation, and 20 iteration used to formalize the random effect	The age of the ASD group was 13 to 18 years (M = 15.13, SD = 1.56) years

(Fan et al., 2015)	EEG signals were filtered between (0.2-45 Hz). Each segment was cleansed with a 50% overlap. Artifacts were removed with EEGlab plugin Gomez-Herrero and slew-rate threshold removal. PCA is applied for feature reduction	Identify mental/affective states on driving skill training environment. Engagement, enjoyment, frustration, and boredom. Workload is not included in this study, but difficult is included instead	16 ASD male participants diagnosed using ADOS-2	A 10-fold cross-validation is used across all the users and the authors include a naive Bayes, a radia-basis SVM, kNN with euclidean function estimation, and a random forest with a 100 trees. The best performances was registered for the kNN with 83.45% accuracy.	The age of the ASD group was 13 to 18 years (M = 15.24, SD = 1.63) years
(Simoes et al., 2018)	Epoched data is taken for 1.25s. The EEG signal was filtered between 1-100Hz. ICA infomax is used for EOG blinking artifacts. PSD features from θ, α, β, and δ bands was grouped. Features such as signal envelope (ENV), Teager energy operator (TEAG) and instantaneous power (POW) were taken from the bands, and Non-Linear Domain features. t-test comparison and evaluation is used for feature selection.	The experiment was divided in two task: 1) a Visual Stimulation Task, and 2) Mental Imagery Tasks. The identification of neutral faces, and emotion faces (happy/sad) in the MI task is the target of this study	17 male teenagers, were recruited and diagnosed using ADOS-2 and ADI-R. Other 17 TD male teenagers were recruited for comparison and effects analysis with ERP average analyses.	80-20% cross-validation with 50 repetitions is used for performance evaluation in MI tasks. SVM optimized kernel, and a WiSARD classifiers are used here. Best F1-scores were reported for WiSARD in emotion faces 76.2±3.3 %, and for neutral faces 69.7±3.4%	ASD: 16.4± 0.6 years; TD: 15.5±0.6 years

Table 3.3 Recent and relevant studies focused on lab-controlled ASD groups emotion or behavioral/outcome measures recognition using neural features, specifically EEG.

3.1.1 Pre-emphasis techniques used for EEG-based pipelines non-ASD

Most no-SOA ML pipelines regardless of ASD participants inclusion have given low importance to data pre-emphasis EEG signal treatment. In Table 3.1 most pre-emphasis methods are only including band-pass filtering. This pre-emphasis evaluation is vague because of there is no standardized for most no-SOA studies in the filter range included here.

A common gold-standard for filtering in EEG lab-controlled environments is a range between 0.1-30Hz. This is convenient for ERP oriented analysis (Lerner, McPartland, and Morris, 2013; Leventon, Stevens, and Bauer, 2014) and make the waveforms for a time-locked signal more distinguishable as neural activity. Nevertheless, these initial filtering ranges are therefore more conditioned by the requirements of the subsequent feature extraction and classification methodologies. Complete frequency ranges between θ and γ are more convenient in some examples such as (Gao, Lee, and Mehmood, 2015; Wang, Nie, and Lu, 2014). The strong contribution of these particular pipelines are in the classifiers and the training methodologies supervised/semi-supervised. Most of the features included in this pipeline have a low probability to be related to the stimuli elicitation and the corresponding neural synchrostate (Jamal et al., 2015), especially for video-elicited long EEG trials.

Other important aspects for pre-emphasis techniques in EEG-based is the downsampling (Jirayucharoensak, Pan-Ngum, and Israsena, 2014; Koelstra and Patras, 2013). The downsampling not only reduces the computational complexity of the subsequent processes but also enhances the Occam's razor effect for the subsequent feature extraction processes. In most cases with a sampling rate on 200 or 256Hz it will be possible to preserve the high-frequency differences observed in high γ bands (Maffei, Spironelli, and Angrilli, 2019).

Datasets Drawbacks

For most studies including non-ASD or healthy volunteers without screening processes the evaluations are broader in terms of the type of stimuli presented. For most cases stimuli videos are presented from two approaches 1) some authors explained data-collection where we found fewer examples, or 2) existing datasets such as DEAP (Koelstra et al., 2011), MANHOB-HCI (Soleymani et al., 2011), or SJTU Emotion EEG datase (SEED) (Zheng and Lu, 2015). Multiple other datasets such as DECAF (Abadi et al., 2015) or AVEC (Schuller et al., 2011) have been created as a result of previous implementations with DEAP or MANHOB-HCI, but the three first initially mentioned datasets have been created for EEG-based emotion decoding without stimuli synchronization.

From the generation of these multimodal datasets, computer scientists have focused their attention to enhance the baseline performances but the pre-emphasis techniques are not precisely important for the most relevant studies included in Table 3.1. The creation of these datasets are focused on making the multimodality easier on the data-collection methodologies, however despite the big amount of data collected, and the versatility of the video emotion stimuli, most neural data collected using this method are not always associated with the stimuli, and the continuous emotion tagging will add an elicitation lag and unmatching effect associated with expected emotion appraisal, and cognitive processing of the annotator (Baveye et al., 2017).

3.1.2 Artifact Removal Techniques

Artifactual trials contaminated with electro-occulogram (EOG) components such as blinking, and eye movement, and electro-myographic (EMG) high-frequency bursts generated by head muscles are very frequent in EEG-based studies.

In lab-controlled environments some companies such as Biosemi and Brain-Products have designed visual inspection tools to do manual artifact rejection, and neuroscientists invest multiple human resources for manual artifact rejection based on visual inspection (Delorme, Makeig, and Sejnowski, 2001). This process is taking an exponential amount of time in comparison with automatic approaches. One of the first semi-automatic for artifact removal was designed by (Gómez-Herrero et al., 2006), however, despite some no-SOA studies use this artifact removal method, it still requires the setting of a numerical threshold for a slew-rate and amplitude rejection per EEG trial.

The best automatic artifact removal techniques have been included as plugins on EEGlab (Delorme and Makeig, 2004) a broadly used Matlab based toolbox for EEG signal processing. EEGlab plugins such as ADJUST (Mognon et al., 2011), Artifact Subspace Reconstruction (ASR) (Mullen et al., 2013), ICA-based artifact removal methods (Winkler, Haufe, and Tangermann, 2011), and other Supervised methods such as MARA (Winkler et al., 2015), or the subordinated methods included in the Prep pipeline (Bigdely-Shamlo et al., 2015). We will describe the usage of some of these methods in the chapters below where we are applying them in our proposed pipeline.

Ironically, these automatic methods are not precisely used for the methods included in the Table 3.1. For no-SOA EEG-based pipelines the eye movements, blinking, and head movements artifacts are more recurrent and affect the quality of the neural activity perceived from the EEG signal. For some studies (Li et al., 2018; Liu, Zheng, and Lu, 2016; Zhang et al., 2018; Zheng and Lu, 2015) there is not any artifact removal pipeline or plugin reported given almost the entire pipeline robustness and statistical contribution to the classifier. This is risky statistically-talking because of the low-quality of the neural representation implied on the artifactual EEG signal recorded, and the unreliability of these signal's features for using them in the subsequent classifier (Bos, 2006).

This previous assertion does not only give a low statistical power to the results obtained on results without artifact removal pre-processing, but it is conditioning generative and discriminative classifiers to learn from non-realistic features (Soleymani et al., 2015).

3.1.3 Features used for EEG-based Emotion Recognition

Some feature selection and extraction methodologies were mentioned in the chapter above. But as well as filtering ranges on pre-emphasis methodologies the variety found in feature selection and extraction methodologies is very vast for EEG-based emotion recognition pipelines (Nakisa et al., 2018).

A very common feature selected for emotion decoding in long EEG trials is the PSD of five important EEG bands such as δ [1-3] Hz, θ [4-7] Hz, α [8-13] Hz, β [15-40]Hz, and γ [> 40Hz] bands in average, and/or using Differential Entropy (DE) measures from the single-trial spectrogram (Frantzidis et al., 2010; Koelstra et al., 2010). These frequency-domain features are commonly used in long EEG trials involving video and a posterior annotation, but frequency measures are also used for time-locked trials such as ERSD, ERD, and ERS as we mentioned above.

More recent studies evolve from entropy and PSD features to High-Order-Crossing (HOC) features, and High-Order-Statistics (HOS). These two models extend the variability of simple frequency-domain features constructing the frequency domain representations to high-order and statistics distributions across the important frequency bands (Petrantonakis and Hadjileontiadis, 2011). These model extensions are commonly supported with Discrete/Continuous Wavelet Decomposition methods (D/CWT) that can model spectrum, phase, and amplitude domains with multiple quadrature family functions (Sharma et al., 2017). Some studies use an even symmetry function such as Daubechies4 or Bi-orthogonal functions (Jenke, Peer, and Buss, 2014) to model more accurately EEG signal slopes and ERPs.

Other common feature extracted from EEG single-trials is the Common-Spatial-Filters (CSP) weights. CSP metrics are positively correlated with spatial symmetry indexes across EEG channels (Kothe and Makeig, 2011). For binary-class problems in motor imagery CSP is used to find the most separable features between two movement (Sturm et al., 2016), this effect is also observed for other multi-class problem in EEG-based emotion decoding (Koelstra et al., 2010).

An extra spatial-based signal process for an EEG signal modelling is the Hjorth filter parameters. The Hjorth model propagate the variance of the spectrogram across the channels in the scalp with three critical features such $Activity = var(y(t))$, $Mobility = \sqrt{\frac{var(y'(t))}{var(y(t))}}$, and

$Complexity = \frac{Mobility(y'(t))}{Mobility(y(t))}$. These features are indicators of slow and high frequency patterns in long EEG trials (Soleymani et al., 2011).

Some researchers have focused on include system dynamics and linear and non-linear descriptors to explain how the EEG signal is describing patterns, especially for long EEG trials. Initially, computer scientists proposed a more descriptive but noise set of features based on DE denoted as $h(X) = -\frac{1}{2}\log(2\pi e\sigma^2)$. This cross-entropy approximation assumes the EEG signal as a Gaussian representation and following this assumption we can define the Differential Asymmetry (DASM) $DASM = DE(X_{left}) - DE(X_{right})$, the Rational Asymmetry (RASM) $RASM = \frac{DE(X_{left})}{DE(X_{right})}$, and the Differential Caudality (DCAU) $DCAU = DE(X_{frontal}) - DE(X_{posterior})$.

All these features are very descriptive in terms of the overall neural activity propagated through the scalp in long term (Duan, Zhu, and Lu, 2013). However, these features need a proper smoothing process such as Linear Dynamical System (LDS). The LDS consist of a feature reconstruction based on a previous measured feature per channel and per frequency band using a linear model function to link previous and posterior samples as F_i. The complete model for the LDS is defined as follows $F_i(X_v) = AF_i(X_{v-1}) + \omega$, where X_v is the EEG signal and A the class transition matrix and ω the modelled noise per trial (Zheng et al., 2015).

Another dynamic EEG signal descriptor is the Non-linear Dynamic Feature (NLD) as a complex marker to measure the EEG signal regularity in time for long-trials. The NLD model describe the entropy as a complex entropy based on Lyapunov series (Aftanas et al., 1997) and the non-stationary properties of the EEG signal using the Hurst Exponent (Wang et al., 2015).

In summary, using all the features set described above for decoding emotion labels in continuous or non-continuous way is not only a complicated task. However, these model are not easy to model EEG signals as a deterministic interaction between the stimulus and the neural activity, and for no-SOA modalities it is not easy to identify real neural activity and artifactual responses (Jenke, Peer, and Buss, 2014). In the next subsection we will describe the most recent and relevant classifiers used for EEG-based emotion decoding for non-ASD population.

3.1.4 Classifiers used for EEG-based Emotion Recognition Pipelines

From Table 3.1 summary we did not find a considerable variety of classifiers to evaluate EEG-based emotion recognition for non-ASD participants. However, for recent studies computer science researchers are trying to include Deep Classifiers to encode any neural feature/outcome measure (Spampinato et al., 2017).

Historically, the Support Vector Machine (SVM) is a flexible classifier to evaluate multiple types of kernels such as linear, polynomial, and radial-basis (Mehmood and Lee, 2015). The usage of SVM has been employed as a first baseline not only for EEG-based emotion recognition but as a baseline for motor imagery, and multiple disorders diagnosis (Subasi and Gursoy, 2010).

As we described above the NLD, DE, and LDS feature smoothing process assume a Gaussian cross-entropy calculation. We can check this in Table 3.1 where the best SVM performances are associated with the linear-kernel SVM (Petrantonakis and Hadjileontiadis, 2011; Zhang et al., 2018).

Other classifiers such as k-Nearest Neighbors (kNN) with the number of neighbours parameter $k = 5$ or $k = 4$ has been used but due to the large variety of feature selection/extraction, this parameter is not easy to calculate or infer for a better performance (Mehmood and Lee, 2016). A more simplified classification approach is the linear classifier and in some implementations such as Naive Bayes, Linear Discriminant Analysis (LDA), or Gaussian Mixture Model (GMM)(Lotte et al., 2007) contributing with good performances in some pipelines (Mehmood, Du, and Lee, 2017; Wang, Nie, and Lu, 2014). A Quadratic Discriminant Analysis (QDA) classifier is also applied with good performances (Jenke, Peer, and Buss, 2014), and this unexpected results are justified when 2-class problem represented with PCA and ICA-based features are solved based on linear hyper-planes separability (Blankertz et al., 2011).

Other classification approaches use Transfer Components Analysis (TCA), Kernel Principal Component Analysis (KPCA), and Transductive Parameter Transfer (TPT). All these models perform parameter learning and subsequent transfer learning from linear hyperplanes. For instance TCA use a calculation of a modified feature map ϕ that minimizes the cost between the input features hyperplane X_s and the predicted hyperplane X_t expressed in Equation 3.1

$$Dist(X_s, X_t) = \left\| \frac{1}{n_1} \sum_{i=1}^{n1} \phi(x_s) - \frac{1}{n_2} \sum_{i=1}^{n2} \phi(x_t) \right\|_{\nabla} \tag{3.1}$$

For these methods x_s is represented as a the complete 2D feature-set composed of channels × time per trial, and similar to TCA, KPCA classifier uses a kernel function to optimize the ϕ calculation. Following (Pan et al., 2010) the Equation 3.1 can be re-written with as a eigenvalue decomposition for a kernel function approximation K as in Equation 3.2.

$$Dist(X_s, X_t) = tr(W^T KLKW) \tag{3.2}$$

These two latter models are not transferring learnt weights and parameters from an specific type of classifier to another. However, TPT is a linear model that transfer the learning weights from a intra-subject SVM classifier to another classifier to leave-one-subject-out subsequent SVM classifier. In (Sangineto et al., 2014) TPT is explained as aa first phase where the initial SVM parameters are learnt from multiple intra-subject SVM classifiers $\Theta = [\omega_i^l, b_i]$. These parameter set is transfer to a subsequent subject independent Regressor SVM-R. The full learning model expression is described in Equation 3.3

$$min_{\pi}\frac{1}{2}\sum_{i=1}^{M+1}||\beta_i||^2 + \lambda_E \sum_{i=1}^{N} E\left(||\Theta_i - f_{\pi}(X_i)||\right) \tag{3.3}$$

In Equation 3.3 the β_s represent the slope parameters of the SVM regressor. The specific application of TPT in EEG-based pipelines was found in (Li et al., 2018) to enhance the subject independent neural emotion decoding.

With the exhaustive search reported in Table 3.1 we found additional Deep classifiers joining supervised and unsupervised training methods. In the next subsection we will describe this models in detail.

3.1.5 Deep Classifiers used for EEG-based Emotion Recognition Pipelines

Deep learning is a hot topic in computer science and in interdisciplinary research constituting a high implementability trend. However, the inclusion of Deep Learning pipelines is missed in EEG decoding pipelines (Martinez, Bengio, and Yannakakis, 2013).

The last Turing award winners prof. Geoffrey Hinton, Yann LeCun, and Yoshua Bengio have contributed incredibly in the ML state-of-the-art (LeCun, Bengio, and Hinton, 2015), and thus have opened the door for exploring new feature-set conformations in 2D, 3D, or 4D arrangements with multiple sensors such as a multi-channel image.

From this point of view, a recent study (Andreotti et al., 2018) has changed the paradigm of Deep Learning and have applied Deep Learning classifiers successfully into the clinical trials. Nevertheless, despite the immaturity of these models in lab-controlled clinical trials multiple researchers have included Deep classifiers in multiple clinical modalities as we described in Table 3.1.

The first Deep classifier which gains attention in lab-controlled environment was the Deep Belief Network (DBN). In previous implementation DBN has been used to extract the most separable feature-space from a initial clinical feature set (O'Leary et al., 2017a,b). For more recent implementation the DBN is modelled as composition of Restricted Boltzmann Machines (RBM) units with the possibility to be trained it in semi-supervised way (Bengio, 2009). The new revolutionary idea with these multi/greedy layer schemes is to avoid overfitting presented in a typical backpropagation process when the number of hidden-layers is high (Asthana, Goyal, and Pandit, 2017).

The DBN can be constituted as a greedy-layer wise model. This model is a biderectional neural structure described in (Bengio, 2009; Goodfellow, Bengio, and Courville, 2016) trained using a semi-supervised process denoted as a pre-training. However, in recent computer vision pipelines there is not any more pre-training associated with unsupervised method but associated with a previous self-supervised or transfer learning process (Iyer et al., 2018), and not necessarily associated with clinical lab-controlled environments.

This unsupervised process is also described as a Gibbs sampling for a unitary layer assuming the DBN as a set of RBMs. This process is referred also as a Constrastive Divergence (CD) process for the complete DBN structure. The initial process give the neural network the name RBM propagating the parameter learning based on the feature-set distribution X_i, and it will modulate the unsupervised parameters θ. Equation 3.4 expresses how the parameters learnt from CD process are proportional to a gradient calculation across layers.

$$\Delta\theta \propto \frac{\partial FreeEnergy(X_i)}{\partial\theta} - \frac{\partial FreeEnergy(\hat{X}_i)}{\partial\theta} \tag{3.4}$$

The propagation is modelled by Equation 3.5 and it propagates the learning parameters from bottom to top in the network structure. Equation 3.5 shows the general model for parameter learning in the previous mentioned CD process.

$$\frac{\partial \log P(X_i)}{\partial\theta} = E\left[\frac{\partial \log P(X_i|h_1)}{\partial\theta}\right] - E\left[\frac{\partial \log P(h_1)}{\partial\theta}\right] \tag{3.5}$$

After the unsupervised pre-training the new results for $\hat{X}_i$, thus the training process for the DBN is complemented with a *fine-tuning* supervised process to adjust the new parameters in the network and avoid overfitting. Most cases implementing DBN use different learning rates for the pre-training, and fine-tuning process. Equation 3.6 shows the general model for the supervised greedy-layer wise training DBN system.

$$P(X_i, h^1, h_2 ..., h^l) = \left(\prod_{k=0}^{l-2} P(h^k|h^{k+1})\right) P(h^{l-1}|h^l) \tag{3.6}$$

The expression on Equation 3.6 shows how the posteriors $P(X_i, h)$ are calculated based on the hidden values parameters h from the layer k to the layer $k+1$. This deep modelling is applied in some EEG-based emotion recognition studies (Gao, Lee, and Mehmood, 2015; Zheng and Lu, 2015) outperforming in the most cases the SVM baselines.

On the other hand, some researchers use a variation of a bimodal Autoencoder named as Bi-Modal Deep Autoencoder (Bi-DAE). This structure has been used when eye tracking features are included in the study. This structure is composed of a Deep RBM on top and a Deep RBM on the bottom in other to find a shared representation for long-trial inputs. This new model and posterior probabilities calculation was defined by (Ngiam et al., 2011) and the new model s process following this $P(h^j|x_i) = \frac{1}{1+exp\left(\frac{b_j+\omega_j^T x_i}{\sigma^2}\right)}$. In our review Table 3.1 we report two pipelines including Bi-DAE with eye tracking and EEG pipelines with each modality as a uni-dimensional feature vector (Liu, Zheng, and Lu, 2016; Zheng et al., 2018). Other Deep Classifier option is the adaptive Deep Neural Network scheme or for other researchers Deep Adaptive Neural Network (DANN). This network is trained with special method denoted as Multiple-Kernel Maximum-Mean-Discrepancies (MK-MMD). This new model consist of a Deep ConvNet (Dumoulin and Visin, 2016; Krizhevsky, Sutskever, and Hinton, 2012) with multifunctional layers, with two or more tranferable convolutional layers. To complement the transfer model an EEG-based implementation (Li et al., 2018) used the model on Equation 3.7 similar to the transfer model of TPT in the previous subsection adding a penalty constant λ to the parameters of the layer $l-1$ propagating the transfer to the full-connected (FC) layers (Long et al., 2015). The kernel function d_k is modulated by the input and target values and for the EEG-based recognition pipeline the authors used this for a better performance using session transfer-learning.

$$\underbrace{min_{\Theta}\frac{1}{n_l}\sum_{i=1}^{n_l} J\left(\Theta(X_i^l, y_i)\right)}_{\text{ConvNet Output Parameters}} + \underbrace{\lambda \sum_{l=1_1}^{l_2} d_k^2(X_i^l, X_t^l)}_{\text{MK-MMD parameters}} \tag{3.7}$$

In this dissertation we will base our ML approach in a Deep ConvNet but with a traditional discriminative training based 2D functional layers as we will explain in the next chapters and sections. Unfortunately, we did not find an recent and relevant implementation of EEG-based emotion recognition using a simple function Deep ConvNet with a exception of (Li, Zhang, and He, 2016) where a Hierarchical Deep ConvNet (HCNN) is implemented with learning transfer as we can see in Figure 3.1. However, in general reviewing EEG-based emotion recognition is going through a multimodal training methodologies (Andreotti et al., 2018) more than using the EEG channel features only. In this study we introduce this critical aspect and a novelty, and a crucial part of our pipeline constructing a new 2D EEG feature set dedicated to our Deep ConvNet classifier.

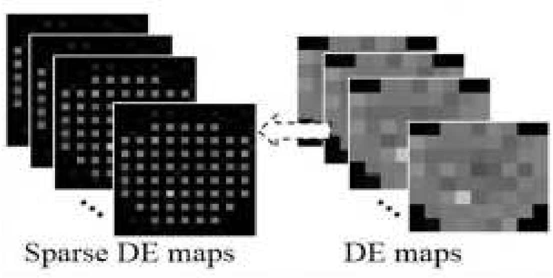


Fig. 3.1 Sparse 2D DE features across channels and frequency bands arrangement for ConvNet on (Li, Zhang, and He, 2016)

3.1.6 Cross-validation Modalities

The summary is reported in the Table 3.1. We found a considerable different amount of cross-validation modalities for EEG-based emotion recognition pipelines.

The first level of cross-validation evaluated in the EEG-based is for instance 5-Fold and 10-Fold cross-validation across subjects. For a single-trial EEG classification use a folding with repetition across all trials is not precisely the most realistic possible scenario.

The next level of formality is the Leave-one-Subject-Out (LOSO) cross-validation that assures a subject independence isolating the trials for a unique subjects for test and the rest for training. For this particular case EEG features and the signal itself it is not easy to find separability between biological, and behavioral features (Blankertz et al., 2004). The EEG trials variability construct a personal neural model per participant. Therefore, an intra-subject cross-validation such as Leave-One-Trial-Out (LOTO) for K-Fold per subject will provide a more significant feature set for classifier evaluation.

In the summary Table 3.1 we reported multiple cross-validations for a single study, where the LOSO cross-validations always show lower performances, in comparison with general K-Fold cross-validations, and other intra-subject LOTO cross-validations. In this dissertation we have multiple intra-subject cross-validations for our preliminary no-SOA/SOA studies, and for our lab-controlled Autism experiment we use a LOTO cross-validation per subject to compare the machine and behavioral performances in a more personal model (Blankertz et al., 2011). We will detail this methodology in the next chapters.

3.2 EEG-based classifiers Emotion Recognition - ASD

Table 3.2 and Table 3.3 shows the more relevant studies for studies applied for ASD early or non-early diagnosis, and EEG-based emotion state decoding. In comparison with studies

including only neurotypical participants, we found fewer studies ML pipelines for diagnosis, and emotion-state decoding including ASD participants.
As we mentioned in previous sections the difficulty to find ASD participants for a ML study including screening process is very considerable. In spite of these drawbacks and the additional care to take into with a no-SOA signal acquisition. In the following subsections we will discuss in detail the results and methodologies use for ML studies including ASD participants, and how these metrics will persuade us for the construction of our pipelines in the next chapters.

3.2.1 Classifiers applied to ASD diagnosis

In the studies reported in Table 3.2 we did not find a big variety of classifiers, and recently there is not a Deep classifier included for ASD diagnosis in the current literature.
For an EEG-based ASD Early diagnosis in (Bosl, Tager-Flusberg, and Nelson, 2018; Bosl et al., 2011) more than 188 subjects diagnosed with Low-Risk Controls (LRC), basal ASD, and High-Risk of ASD (HRA) using Infant Sibling Project assessments (Stone, McMahon, and Henderson, 2008) considering screening methods. In the first study the features were evaluated across groups to define a biomarker separability, and in the subsequent study a set of classifiers such as k-NN, SVM, and linear Naive Bayes classifiers. The main novelty of this set of studies is modelling the EEG resting-state signals using the Recurrence Quantitative Analysis (RQA) features (Acharya et al., 2011).
First the recurrence plots $R_{i,j}$ is defined as a time displacement 2D quantitative metric. This plot is defined as a gray-scale matrix showing amplitude difference in time through long-term trials such as resting-state $R_{i,j} = \Phi\left(\varepsilon_i - \left|\left|x_i - x_j\right|\right|\right)$. The x_i is a no-displaced EEG signal and x_j is the displaced signal modelled with a different z-representation. The final recurrence plot is normalized using a Heaviside function Φ , and the threshold distance ε_i.
The studies using RQA for ASD diagnosis use a particular set of features from the 2D recurrence plot. Features such as the Recurrence Rate $RR = \frac{1}{N^2}\sum_{i,j=0} R_{i,j}$, the Determinism $DET = \frac{\sum_{l=l_{min}} lP(l)}{\sum_{i,j} R_{i,j}}$ being the $P(l)$ the frequency distribution, and l_{min} as the minimum length of the recurrence plot diagonal, Mean Diagonal Line Length $< L >= \frac{\sum_{l=l_{min}} lP(l)}{\sum_{l=l_{min}} P(l)}$, the recurrence plot Entropy estimation $ENTR = \sum_{l=l_{min}} P(l)\ln(P(l))$, the Laminarity (LAM) or the fractions of points in the recurrence plot that forms vertical lines $LAM = \frac{\sum_{v=v_{min}} vP(v)}{\sum_{v=1} P(v)}$ being v the length of the vertical line, the Trapping Time (TT) or the mean length where the EEG representation is trapped in vertical lines $TT = \frac{\sum_{v=v_{min}} vP(v)}{\sum_{v=v_{min}} P(v)}$, longitudinal features such as Longest Vertical Line $V_{max} = \max\left(v_i; i = 1..N_v\right)$, Longest Diagonal Line $L_{max} = \max\left(l_i; i = 1..N_l\right)$.
From the first study in the summary, the authors correlate the modifier Multiscale Entropy

(mMSE) EEG features with the LRC, and HRA prediction. In this study and in the subsequent second study three classifiers were evaluated k-NN, SVM, and Naive Bayes using LRC, and ASD basal diagnosis subjects' RQA and DWT features for training and HRA for test. In some analysis they used only the channels T7 and T8 features and achieving good performances.

Other studies (Eldridge et al., 2014; Pistorius et al., 2013) follow the biomarker analysis initially proposed by (Bosl et al., 2011). RQA, and mMSE features are included again in this case to diagnosis early ASD, but with an alternative screening methodology such as the ADOS-2 and the ADI-R.

The most popular classifier in Table 3.2 is the linear kernel SVM reporting the better performances. Multiple studies (Castelhano et al., 2018; Eldridge et al., 2014; Heunis et al., 2018; Jamal et al., 2014) use the SVM as baseline for diagnosis 2-class prediction. Some of these studies extend the statistical an incidence of RQA and DE features across frequency. However, (Heunis et al., 2018) has been the first to include multi-variable classifier such as 1 hidden-layer neural network as the Multilayer-Perceptron (MLP) obtaining good performances.

A more specific study (Jamal et al., 2014) is the first study including CWT, DE, synchronization in time and frequency domain features, and k-NN clustering features from emotional face stimuli elicitation. In this study a leave-one-subject cross-validation is evaluated, thus obtaining a good performances from the EEG single-trial's features. This study is the first to include adult participants for ASD diagnosis prediction.

In the table we report only one study including MEG signals (Ingalhalikar et al., 2014). This study use an alternative assessment for ASD screening based on the Clinical Evaluation of Language Fundamentals (CELF-4) (Paslawski, 2005), and the classes for the automatic diagnosis is the diagnosis of ASD with low language impairments ASD-LI-, and ASD with high language impairments ASD-LI+. The authors used the Diffusion Tensor Imaging features from posterior, temporal, and central channels related to auditory stimuli.

3.2.2 Classifiers applied EEG-emotion recognition including ASD individuals

Table 3.3 is showing the studies including ML pipelines for emotion-state decoding using EEG features. Checking the state-of-the-art and literature it was not easy to find studies for EEG-based emotion recognition with previously diagnosed ASD participants.

For the studies we found, we concatenate three studies in time evaluating complex emotion states on a driving Virtual Reality (VR)environment. In (Fan et al., 2015, 2017b) the authors elevate the level of complexity of cross-validations trying to decode first 4 or in the more recent studies 5 different complex emotion states such as engagement, enjoyment, frustration, and boredom, or an additional measure of the level of workload from a 5-fold stratified cross-validation to a 10-fold nested cross-validation with a LOSO cross-validation outer loop.

In the earlier study (Fan et al., 2017a) the authors predict 4 more basic emotions such as joy, sadness, surprise, and neutral following the same VR driving environment of the studies mentioned above. For all these studies the authors use HOC, HOS, and Hilbert-Huang Spectrum (HHS) features. As we explained above these features are essential for long EEG trials decoding (Jenke, Peer, and Buss, 2014).

The more recent study for EEG-based emotion decoding including ASD groups (Simoes et al., 2018) decode emotion from PSD features including θ, δ, α, and β bands, and other morphological features such as Teager Energy operator as a discrete time energy operators $\omega^2 x(\omega) * x(\omega)$ where the $*$ is the convolution of the two frequency representations. The envelope-differential operator (ENV) $|x(\omega) + j\mathscr{H}(x(\omega))|$ being $x(\omega)$ the frequency response of the EEG signal, and $\mathscr{H}$ the Hilbert Transform, and the Instantaneous Power (POW) $X\left(\frac{dx(t)}{dt}, \omega\right)$. The big difference between this study and the rest including ASD groups is: 1) a face emotion recognition task used for training a SVM and a Wilkes, Stonham and Aleksander Recognition Device (WiSaRD) classifier (França et al., 2014) using EEG and image features, and 2) a subsequent mental imagery task remembering the previous faces presented in the previous stage.

Chapter 4

Machine Learning preliminary Evaluation on DEAP, Object Categories, and TROIKA datasets

In this chapter we present the results of the preliminary ML pipelines evaluation for 1) DEAP EEG-based emotion arousal/valence level decoding, 2) An object category prediction using EEG features subspace transformation, and 3) a HR prediction and IBI signal calculation using regression light complexity implementations, applied for treadmill exercise, and real-life scenarios.

The methodologies described here are well documented in previous pipeline implementations, we will focus on the results, research implications, and the new methodologies included in these implementations. We will cite the papers related to these implementations.

4.1 DEAP EEG-based Emotion Recogniton

In this section we will describe four different items of our DEAP dataset EEG-based emotion recognition preliminary pipeline: A) the DEAP dataset pipeline and its corresponding baseline replication, B) our signal processing pre-emphasis pipeline part composed of the Hilbert Transform representation, C) the Bhattacharyya feature-selection phase explanation, and D) the DNN, K-NN, and GMM classifiers used in this pipeline.

DEAP (Koelstra et al., 2011) dataset is a multimodal dataset where 40 1 min length music videos clips trials are shown to 32 subject who was instructed to recognize high and low arousal-valence levels in a continuous annotation modality. DEAP authors used a Biosemi ActiveTwo EEG device amplifier.

DEAP paper proposes to analyze the feature representation in a personalized way as we explained above. Then, the DEAP baseline should be evaluated using a LOTO cross-validation per subject using the symmetrical features from 28 out of the 32 channels except Cz, Fz, Oz and Pz channels. Features such as statistical features mean, standard deviation, maximum, minimum and the Power Asymmetries Indexes (PAI)-Equation 4.1 show the difference between the absolute powers between the left and right channels- for four important EEG bands such as α , β, θ, and γ. A Gaussian Mixture Model (GMM) classifier calculating two mixtures for arousal-valence (high/low) classes

$$PAI_{left-right} = 10(\log_{10}(X(\omega)_{left}) - \log_{10}(X(\omega)_{right})) \tag{4.1}$$

On the other hand, for our specific approach each EEG single-trial is filtered using two 150 order Blackman-Harris FIR filters to preserve α and γ rhythms. For each EEG single-trial $f(u)$ and for each k channel the EEG signal in time-domain is transformed using the absolute value of the Hilbert transform following Equation 4.2 Torres, 2013. On Figure 4.1 we describe the whole 2-channels EEG-based emotion decoding pipeline on DEAP pipeline.

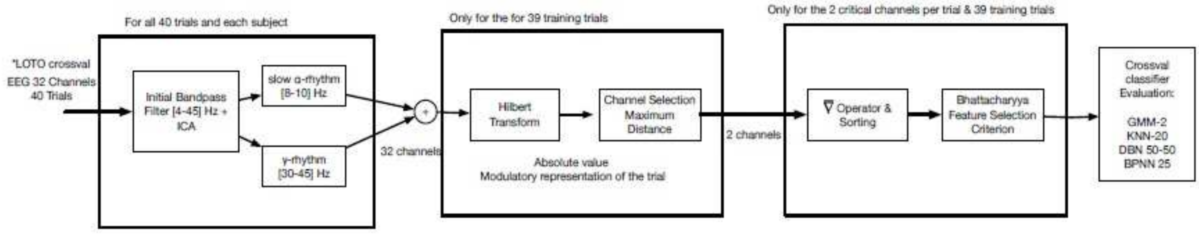

For all 40 trials and each subject
Only for the for 39 training trials
Only for the 2 critical channels per trial & 39 training trials
*LOTO crossval
EEG 32 Channels
40 Trials
Initial Bandpass Filter [4-45] Hz + ICA
slow α-rhythm [8-10] Hz
γ-rhythm [30-45] Hz
+
32 channels
Hilbert Transform
Channel Selection Maximum Distance
Absolute value
Modulatory representation of the trial
2 channels
∇ Operator & Sorting
Bhattacharyya Feature Selection Criterion
Crossval classifier Evaluation:
GMM-2
KNN-20
DBN 50-50
BPNN 25

Fig. 4.1 EEG-based emotion decoding pipeline, thus showing per subject cross-validation for a single-trial classification.

$$\begin{gathered}|\mathscr{H}(x)| = \left|\frac{1}{\pi}\int_{-\infty}^{\infty} f^k(u)\frac{1}{(x-u)}du\right| = \\ \left|\lim_{x\to 0}\left(\int_{-\infty}^{x-\varepsilon} + \int_{x-\varepsilon}^{\infty}\right) f^k(u)h(x-u)du\right|\end{gathered} \tag{4.2}$$

The absolute value of the Hilbert Transform $|\mathscr{H}(x)|$ provides a positive time-domain and a modulatory signal representation related mainly to α low-frequency formers. These positive values describe better spectrum asymmetries in α rhythms. The pair of channels with the higher difference between the $|\mathscr{H}(x)|$ means are selected only using the training-set trials -39 trials for training-. Figure 4.2 shows the histogram for this 2-channels selection where we can observe some high occurrences in left-fronto-central and parietal regions (Samaha et al., 2015).

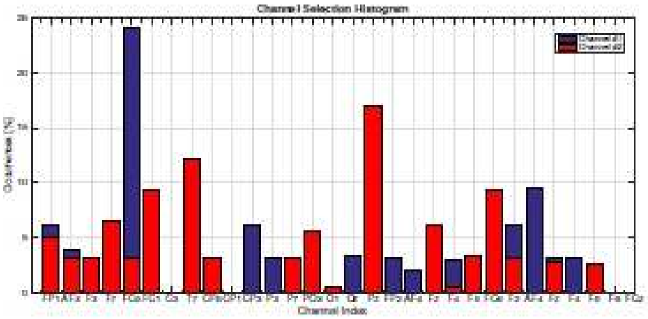


Fig. 4.2 2-channel selection histogram based on the highest Hilbert Transform difference between both channels for DEAP dataset.

The top selected channels using the Hilbert transform differences were FC5 : 23.71 %, Pz : 17.32 % and T7: 13.34 %. These percentages are calculated from the total occurrences, and validating $\alpha + \gamma$ spiking in fronto-central region during long audiovisual stimuli presentation (Kang et al., 2015). However for a non-SOA study including continuous annotation and emotion elicitation, the synchronization between the stimulus and the neural activity is not locked and instantaneous.

With the pair of electrodes selected, we concatenate and sort the Hilbert signal representations. Subsequently, we use the gradient operator ∇ to obtain the most variable positive ranges across the Hilbert representation, and sort the resulting feature-vector only using the 1000 more variable peaks per trial.

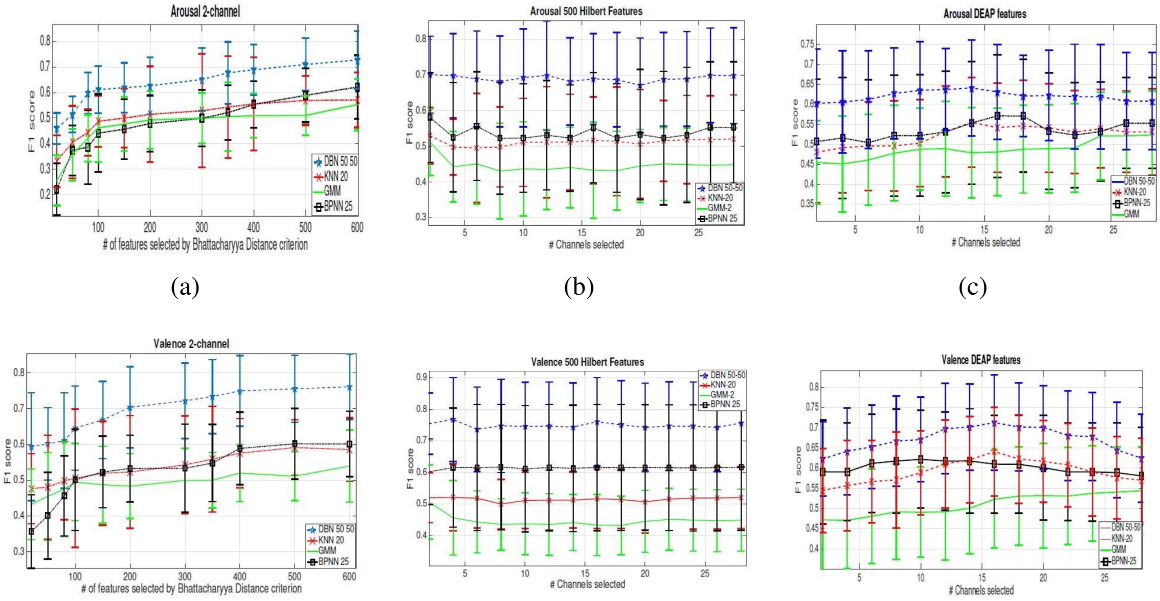
With the sorted Hilbert peaks we use the Bhattacharyya distance criterion (Obermaier et al., 2001) described by Equation 4.3. This process assumes a non-linear Bayes-Error's upper bound representation (E_{min}) with the high-low arousal-valence levels' priors denoted by $p(\omega_1)$ and $p(\omega_2)$, and the corresponding posteriors denoted by $p(z|\omega_1)$ and $p(z|\omega_2)$.

$$J_{bhatt} = -\ln \left[\int_z \sqrt{p(z|\omega_1)p(z|\omega_2)} \right] \tag{4.3}$$

We observe each trial from training-set obtaining a subset of features after the *sequential search* from the Bhattacharyya feature map computing the Equation 4.4 (Somol, Novovičová, and Pudil, 2006) and we picked the best 500 ranked features out of 1000 maximum Hilbert peaks per trial.

$$E_{min} \leq p(\omega_1)p(\omega_2) \left[\int_z \sqrt{p(z|\omega_1)p(z|\omega_2)} \right] \tag{4.4}$$

After the feature selection process the resulting feature-set for training set is 39 trials × 500 features and the corresponding test-set 1 trial × 500 features per subject.
The feature-set is normalized using $X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$. A GMM parametric classifier -DEAP baseline- and a set of non-parametric classifiers such as KNN-20, and BPNN-25 with a sigmoidal activation function are trained using PRtools Duin, 2000 Matlab Toolbox. We evaluate a DBN 50-50 units classifier using *DeepLearn* Matlab toolbox (Palm, 2012) with a sigmoidal activation function, 100 unsupervised Contrastive Divergence iterations, and 100 fine-tuning supervised iterations. We set learning rates of unsupervised and supervised processes in 0.01 and 0.1 respectively.
Using the 500 more significant features selected by the Bhattacharyya distance criterion to train the DBN 50-50 classifier. As a results we found a F1 score of 0.755 for valence, and 0.711 for arousal 2-class low high level problem. Tables 4.1 and 4.2 show the results for valence and arousal classification levels, and we found the maximum performance obtained by 2-channels+DBN setting. The DEAP baseline is symbolized by $*$.
To complement our results analysis we vary the number of features selected by the Bhattacharyya distance criterion starting from 25 to 600 features. Figures 4.3a and 4.3b show the F1-score variation in terms of the number of features selected for arousal and valence. Comparing these results with the DEAP baseline on Tables 4.1 and 4.2 we can observe for 2-channels and more than 100 features selected it is possible to obtain a better performance in comparison with the DEAP baseline.
An extra analysis in this evaluation was to change the Hilbert differences features. For this evaluation we used all the possible symmetrical pair of channels in both hemispheres, we extend the channel selection per pairs from 2 to 28 selecting from the higher to the lower best pair of channels inferred by the Hilbert Transform phase. Figures 4.3c and 4.3d describe the variability in terms of number of channels for the Hilbert Transform differences, and Figures 4.3e and 4.3f show the same variability using the DEAP features. Comparing approaches we observe a little performance increasing $\sim$ 15-18 channels for DEAP baseline in comparison with our features. However, our features represent a better performance for larger number of channels.
We can conclude from this initial evaluation on a broadly known dataset that the combination of good signal envelope representation such as Hilbert transform with a robust classifier such as DBN can contribute on good EEG-based arousal valence levels decoding performances, especially with long EEG trials.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4.3 Features and channels variation, using DEAP and our methodology features results. Figures 4.3a and 4.3b refers the arousal and valence results using the Bhatacharyya features variation from 25 to 600. Figures 4.3c and 4.3d are associated with the channel variations using our features. Figures 4.3e and 4.3f that are related to the same variation but for the DEAP features

Table 4.1 Valence results in a LOTO per subject modality, comparing DEAP (28 channels) PSD features and our pipeline (2-channels + 500 features),*DEAP baseline $p < 0.05$

Valence - Methods		**DBN 50-50**			**GMM ***			**KNN-20**			**BPNN-25**		
		Pr	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
DEAP (28 chan) (Koelstra et al., 2011)	**AVG**	0.641	0.592	***0.6123***	0.555	0.542	***0.546****	0.591	0.556	0.571	0.594	0.581	0.587
	STD	0.151	0.096	0.112	0.087	0.069	0.074	0.111	0.074	0.079	0.152	0.102	0.112
Hilbert + Bhattacharyya (2 channels)	**AVG**	0.790	0.728	***0.755***	0.511	0.503	0.502	0.598	0.586	0.591	0.612	0.582	0.596
	STD	0.086	0.119	0.097	0.108	0.073	0.077	0.077	0.062	0.062	0.122	0.101	0.114

Table 4.2 Arousal results in a LOTO per subject modality, comparing DEAP (28 channels) PSD features incidence and our pipeline (2-channels + 500 features),*DEAP baseline $p < 0.05$

Arousal - Methods		**DBN 50-50**			**GMM***			**KNN-20**			**BPNN-25**		
		Pr	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
DEAP (28 chan) (Koelstra et al., 2011)	**AVG**	0.631	0.591	***0.606***	0.539	0.537	***0.538****	0.538	0.529	0.532	0.545	0.535	0.541
	STD	0.157	0.132	0.133	0.079	0.072	0.075	0.113	0.077	0.086	0.125	0.113	0.121
Hilbert+ Bhattacharyya (2 channels)	**AVG**	0.762	0.671	***0.710***	0.504	0.512	0.510	0.569	0.566	0.569	0.587	0.563	0.576
	STD	0.127	0.105	0.103	0.086	0.081	0.085	0.098	0.091	0.093	0.113	0.102	0.111

4.2 EEG-based Object Category Decoding

For EEG-based object category decoding implementation we used the parameters set in (Torres, Stepanov, and Riccardi, 2016). In this study we propose a combination of Bhattacharyya distance Criterion as mapping as we explained in the previous section, and a Deep Neural Network (DNN) classifier to improve the performance of stimulus concept decoding for within-subject analysis as the 5-fold cross-validated classification, and a cross-subject analysis as leave-one-subject-out (LOSO) cross-validation.

We refine the previous analyses of DEAP baseline selecting features from seven Regions-Of-Interests (ROIs) such as Middle-Frontal, Left-Anterior, Right-Anterior, Middle-Anterior, Left-Posterior, Right-Posterior, and Middle-Posterior finding significant differences for grand-average ERPs comparing the two concept classes in the N2 component range [150-200]ms. EEG data was obtained from (Murphy et al., 2011) and it has been recorded using a 64 electrode Brain-Vision-Brain-Amp system with a sampling frequency $f_s = 500$Hz and a right-ear lobe channel reference. Each trial was filtered using a Butterworth bandpass filter between 1-50 Hz, and down-sampled to 120 Hz with purpose of removing high and low frequency noise, and eye-artifacts were removed by hand from an ICA-infomax decomposition.

We take data from seven healthy Italian speakers (5 male and 2 female,$\mu = 29$). They were asked to silently name animal and tool objects classes presented in normalised grey-scale photographs. The study presented 30 land-mammals, and 30 work-tools photographs each presented in random order composing a total of 180 trials for each class.

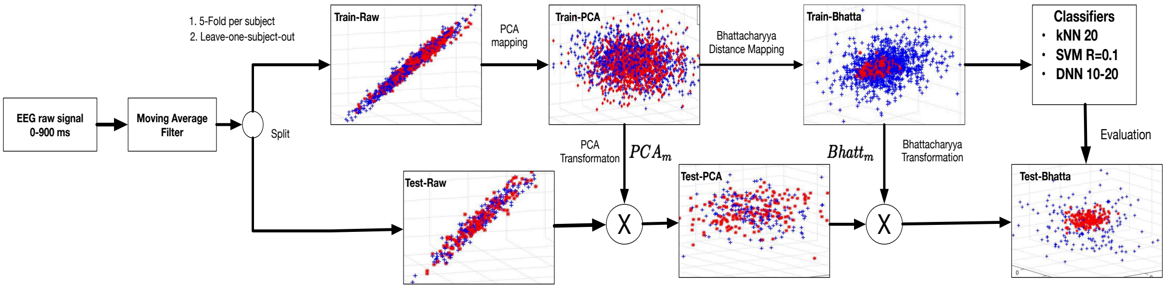


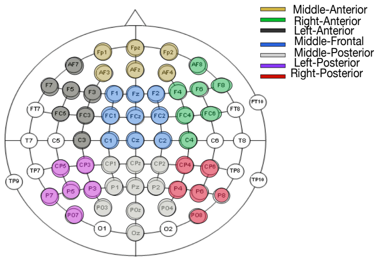
Fig. 4.4 Object decoding pipeline with the corresponding subcomponents. The training and test distribution pass through PCA and Bhattacharyya feature extraction and selection analysis, thus obtaining a concentric separation regions.

The EEG-based object decoding pipeline methodology in Figure 4.4 is composed of 3 stages: 1) a pre-processing stage consisting in a moving average filter through all the 64 channels, and per the trial, 2) we split the complete dataset for a leave-one-subject-out 6-to-1 cross-validation, and a 5-fold cross-validation per subject following the number of iterations proposed by (Murphy et al., 2011) to generalize unseen exemplars per class.

From the stimuli neural response we calculate PCA (PCA_m) and Bhattacharyya distance criterion ($Bhatt_m$) maps using the training set only. Multiplying the corresponding transformation matrices with the corresponding test set feature matrix we obtained the distributions defined by PCA and Bhattacharyya mappings across the pipeline execution.

Each channel out of 64 are filtered using a moving average filter with order $M = 5$ to preserve the frequencies between $[0-32]$ Hz as low-pass filter. The resulting signal per channel is smoothed to increase the SNR and include α and γ rhythms, and preserving the critical negativities in N2 ranges for a grand-average evaluation.

We use 72 samples from 0 to 500 ms range from each channel for the ERP analysis. The initial feature space is composed of 64 channels $\times$ 72 samples $= 4608$ features across the 180 trials for each class, and per subject. Subsequently, we group these features according to the set of channels, and ROIs described in Table 4.5b and Figure 4.5a (Meulman et al., 2014).



(a) scalp ROI Distribution

ROIs	Channels
Middle-frontal	Fz, FC1, FC2, Cz, FCz
Left-Anterior	F7, F5, AF7, FC7, FC5
Right-Anterior	F8, F4, AF8, FC8, FC4
Middle-Anterior	Fp1, Fp2, AF3, AF4, F1,F2
Middle-Posterior	CP1, CP2, P1, Pz, P2, POz, PO4, PO3, O1, O2
Left-Posterior	CP5,CP3,P7,P5,PO7,P9
Right-Posterior	CP6,CP4,P4,P6,PO8,P8

(b) ROI distribution Table

Fig. 4.5 Channels distribution per ROI in the scalp Figure, and Table showing the channels distribution

PCA is used as parametric model in the training set side to calculate the covariance matrix Σ from both semantic classes. This matrix is diagonalized to extract the most representative eigenvalues λ_n as well as the related eigenvectors ϕ_n.

$$\sum_{n=0}^{M} \phi_n \Sigma \phi_n^T = \sum_{n=0}^{M} \lambda_n (\phi_n \phi_n^T - 1) \tag{4.5}$$

The PCA mapping is applied following Equation 4.5 thus reducing the number of features per trial to $M = 180$ and using the eigenvalues including the 95 % of the variance in the resulting distribution.
Using the Bhattacharyya distance criterion to complement the PCA feature selection we first define Φ as an orthogonal matrix for reducing the PCA features $N \times M$ to $N \times M_b$ with $M_b < M$. We define $\Phi = [\Phi^1, \Phi^2, \Phi^3, \ldots, \Phi^{M_b}]$, M_b as the number of features, and N the subspace dimensionality, thus updating each column vector following the *sequential search* (Somol, Novovičová, and Pudil, 2006). Equation 4.6 shows how each vector should be updated from its own direction based on the *sequential search*, being δ a constant transformation step.

$$\Phi^i_{new} = \Phi^1_{old} + \delta \Phi^i_{old} \tag{4.6}$$

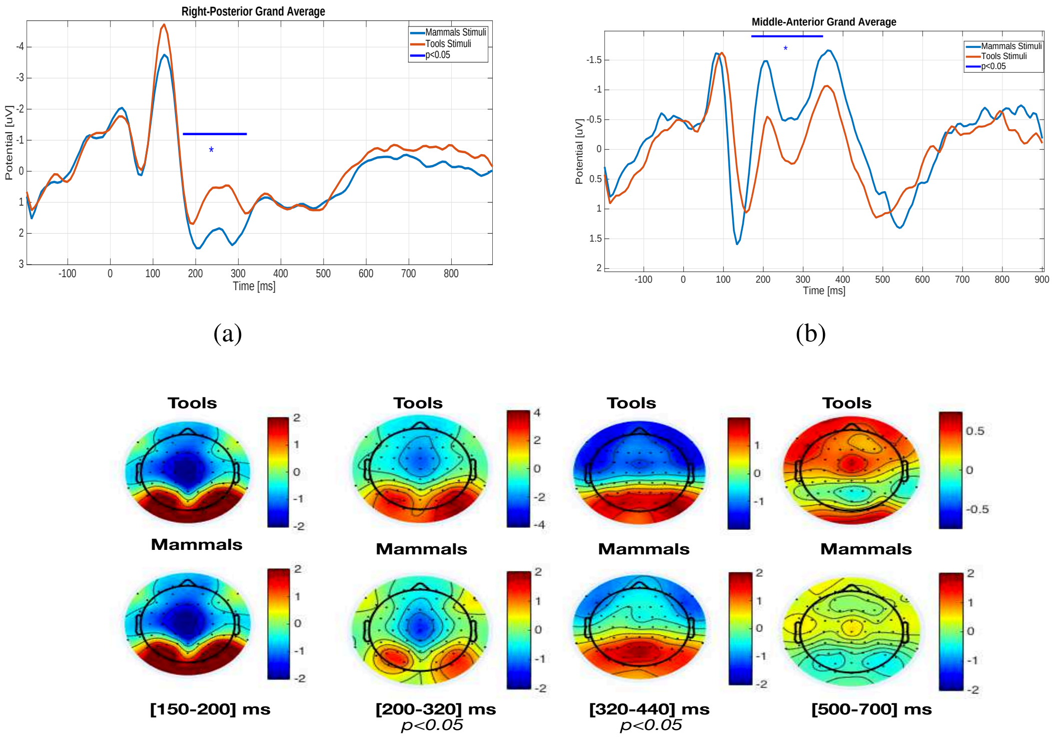
The Φ matrix is randomly initialized from the PCA distribution taking a subset from both concept classes. We use $\delta = 0.1$ as (Choi and Lee, 2003) iterating 500 times along the feature-set thus reducing the features to $M_b = 180$. The resulting distribution after the sequential search is presented in Figure 4.4. The DBN used here follows a similar proces we used in the DEAP approach in the section above. The training process is composed of two stages: 1) the unsupervised pre-training process defined as CD, and 2) a fine-tuning supervised process defined as a back-propagation step. Now, we will report the results for the grand-average ERP, and the classification performances.

4.2.1 Grand Average Analysis

We evaluate the statistical difference between Tools and Mammals class neural responses with a non-parametric test based on a Montecarlo permutation using a 5000 iterations, and a Bonferroni-Holm correction after the permutation for each subject, and ROI based on (Maris and Oostenveld, 2007).

This analysis shows that some specific ROIs in early time ranges exhibit significant difference between Tools and Mammals concept classes on: the Medial-Frontal region in $[150-200]$ ms $p = 0.0321$, Middle-Posterior in $[150-200]$ ms $p = 0.0381$, and Left-Posterior in $[200-320]$ ms $p = 0.0482$. This is consistent with propagation of semantic information from the visual-cortex, the posterior and occipital regions to the Prefrontal-Cortex (PC) (Jeon and Friederici, 2015).

Analyzing other ROIs we found significant differences in Right-Posterior $p = 0.0282$ and Middle-Posterior $p = 0.0184$ in $[200-320]$ ms. Figure 4.6a shows the ERP plots for these neural response ranges being consistent with the good performances obtained using the features on this time ranges. We also found significant differences in Middle-Anterior regions in the same time ranges for the Right-Posterior ROI $p = 0.0342$, in Figure 4.6b we can see the waveform for the Middle-Anterior region. All these p-values support the appearance of early and posterior positivities correlated with semantic decoding (Simanova et al., 2010). Figure 4.6c shows the averaged scalp topoplots for the time regions mentioned above.



(a)

(b)

(c)

Fig. 4.6 ERP plots for Right-Posterior and Middle-Anterior in Figures 4.6a and 4.6b responses show a significant region between $[150-320]$ ms, consistent with memory processes across early neural responses (Friederici and Singer, 2015). The scalp plots (Figure 4.6c) show significant activations around posterior and middle-posterior regions, especially for N2 ranges. The colorbar scale show the variation of neural activity between the maximum and minimum of the average signal.

4.2.2 Classification Results

To evaluate DBN classifier performances we use other two extra common baseline classifiers such as SVM, and kNN. For all the classifiers we use the 180 best ranked Bhattacharyya features. We use a LOSO, and a 5-fold per subject cross-validations introducing initial EEG signal amplitude feature in the range $[0-500]$ ms for each channel and for each trial.

After the application of the Bhattacharyya mapping the dataset is composed of 180 examples for each class, and for each subject. For the training process we set up the classifiers as follows: 1) The SVM classifier using a radial-basis type with $R = \frac{1}{N}$ as the baseline paper described, 2) the kNN classifier using a $k = 20$ number of neighbors, and the DBN with 2 hidden-layers composed of 10-20 units, and with learning rates $\varepsilon_{pre-training} = 0.01$, and $\varepsilon_{fine-tuning} = 0.1$, iterating 120 times for the unsupervised pre-training, and 520 times for

fine-tuning supervised process encoding 10 mini-batches per iteration. This DBN array was implemented with the Deep-learning Matlab toolbox (Palm, 2012), and the SVM and kNN classifiers using PRTools v4.0 (Duin, 2000).
To validate our proposed pipeline we replicate the process described in (Murphy et al., 2011) where for each trial we use a CSP for adjust feature separability using only the features between $[0-500]$ ms. Subsequently, we train a SVM classifier with these same features, using the same 5-fold cross-validation explained in (Murphy et al., 2011). All the baseline results are reported on Table 4.3.

Table 4.3 Baseline classification results obtained after we replicate the process of (Murphy et al., 2011), using a radial basis SVM $R = \frac{1}{N}$ in which N is the number of training exemplars for each cross-validation modality.

LOSO	**Sub1**	**Sub2**	**Sub3**	**Sub4**	**Sub5**	**Sub6**	**Sub7**	**Avg**
Acc	0.699	0.655	0.712	0.592	0.661	0.585	0.671	**0.653**
5-Fold	**Sub1**	**Sub2**	**Sub3**	**Sub4**	**Sub5**	**Sub6**	**Sub7**	**Avg**
Acc	0.722	0.712	0.721	0.682	0.721	0.713	0.698	**0.709**

The results using our proposed feature-set are reported on Table 4.4. We achieved high classification performance in Left and Right Posterior, and Middle-anterior ROIs, thus showing significant difference comparing the performances with a multi-variable ANOVA analysis, and thus being consistent with the results obtained in the grand average analysis. The DBN was the best classifier across ROIs, and also for All-ROIs feature combination. The Deep model advantages allow to identify better the entangled separation between Tools and Mammals classes feature-space.

Table 4.4 Accuracy average results for LOSO and 5-Fold modalities specifying the ROIs. Bold+italics values are $p < 0.05$ using t-test and inter-classifier comparison. We achieve high performance using features from $[0-500]$ ms ranges. This suggests that semantic information is contained in a higher level of entropy ranges such as $[0-500]$ ms used in (Murphy et al., 2011).

Modalities	**LOSO**			**5-Fold per subject**		
ROIs	**kNN-20**	**SVM R=0.1**	**DNN 10-20**	**kNN-20**	**SVM R=0.1**	**DNN 10-20**
Middle-frontal	0.535	0.567	0.731	0.401	0.699	0.782
Left-Anterior	0.532	0.651	0.746	0.443	0.631	0.771
Right-Anterior	0.510	0.583	0.730	0.467	0.681	0.789
Middle-Anterior	0.545	0.657	0.721	0.378	0.674	0.797
Middle-Posterior	0.578	0.527	***0.752***	0.452	0.731	***0.824***
Left-Posterior	0.612	0.611	0.744	0.534	0.743	***0.834***
Right-Posterior	0.624	0.626	***0.758***	0.452	0.761	***0.853***
Average	0.562	0.603	**0.740**	0.447	0.703	**0.807**
All-ROI$_s$	0.601	0.633	**0.751**	0.527	0.718	**0.838**

4.3 HR prediction using HR spectrum and Adaptive Filtering - TROIKA

In this section we will describe the details of the paper (Torres et al., 2016) where we explain the TROIKA dataset baselines, and our proposed pipeline for a low-complexity HR and IBI signals calculation from noisy, and distorted BVP signal under an environment affected by movement artifacts (MA) such as physical exercise, and daily life activities.

We use the publicly available TROIKA dataset (Zhang, Pi, and Liu, 2014) used in the 2015 for the *IEEE Signal Processing Cup* [1]. This dataset recreates closely the real-life motion activities in our experiments. TROIKA is composed of 5 minute long treadmill trials performed by 12 different subjects. The biosignals are recorded from PPG, and accelerometer sensors including an extra ECG ground-truth. Each 5 min trial is divided into 6 different tasks such as: 30 seconds - rest (1-2 km/h), 1 min - Walking (6-8 km/h), 1 min - Running (10-12 km/h), 1 min - Walking (6-8 km/h), 1 min - Running (10-12 km/h) and finally 30 seconds - rest (1-2 km/h). The PPG and the accelerometer sensor were sampled at 250Hz.

For these TROIKA experiments the authors stablish an sliding window of 8 seconds with a 6 second overlap between windows following that the HR and IBI calculation should be done per window, or per each pair of ECG R-peaks.

The HEAL-T pipeline is shown in Figure 4.7 and consists of 1) a initial filtering stage which consists of a fast-ICA decomposition and a moving-average-filter application, 2) a Recursive Leave-Squares (RLS) filter scheme for intrincated MA removal, 3) a Blackman-Harris Window (BHW) FIR filter bandwidth adjustment, 4) a spectral peak tracking process added as a novelty, and 5) a final IBI estimation task.

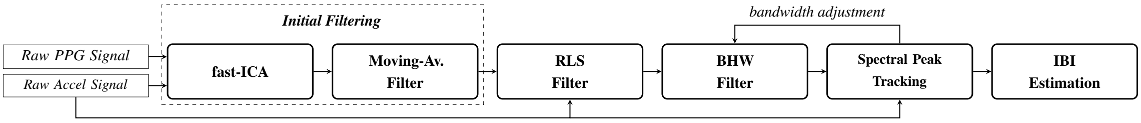


Fig. 4.7 The proposed HEAL-T for HR estimation method applied per signal-window.

For the filtering stage each window is assumed to be statistical independent modelling and we can stablish a linear phase on the signal to preserve the BVP important frequency ranges with an adequate spectral resolution (Cui et al., 2015).

The fast-ICA decomposition or unmixing models the negentropy gradient between PPG and Accelerometer channels in order to calculate the unmixing matrix W_s (Hyvarinen, 1999). This process returns 5 independent-components (ICs) one for each PPG channels, and one

[1] http://archive.signalprocessingsociety.org/community/sp-cup/ieee-sp-cup-2015/

for each x, y and z axes from the Accelerometer signal.

The subsequent fast-ICA mixing process use the two PPG channels and Accelerometer ICs are ignored. To remove the residual MAs after applying the fast-ICA we apply a moving average filter. For window X^i a linear compositional model can be defined as $X^i = PPG^i + Accel^i + N^i$. PPG^i and $Accel^i$ are PPG and Accelerometer signals, and N^i is a high-frequency additive noise.

The filter is a computational inexpensive convolution between X^i signal and the constant impulse response $\frac{1}{M}$. Equation 4.7 shows the moving average filter that is proposed for increasing the SNR and reduces the signal additive noise (Lee, 2014).

$$\hat{X}^i(k) = \frac{1}{M}\sum_{n=0}^{M-1} X^i(k-n) \tag{4.7}$$

4.3.1 RLS Filter

We set our moving average filter with an order $M = 20$ as a low-pass bandwidth to preserve the critical PPG channel frequency ranges $[0.9 - 2.5]$Hz (Singh et al., 2015).

Subsequently, we iteratively apply a RLS filter to remove the incidence of MA in the BVP signals. RLS provides a non-misadjusted solution for least-squares cost function and a resulting smoothed spectrum. The filter uses an optimal adaptive noise-cancellation algorithm for the critical PPG channels low-frequency bands (Shimazaki et al., 2014).

Equations 4.8 and 4.9 show evolution of the RLS algorithm where we set a forgetting factor λ in order to modulate the filter weights $\omega(n)$ as a function of the previous weights $\omega(n-1)$.

$$\omega(n) = \omega(n-1) - \kappa(n-1)\,[d(n) - \omega(n-1)\upsilon(n)] \tag{4.8}$$

$$P(n) = \lambda^{-1}P(n-1) - \lambda^{-1}\kappa(n-1)\upsilon^T(n)P(n-1) \tag{4.9}$$

RLS parameters such as $\upsilon(n)$ the accelerometer signal, the factor $\kappa(n) = (\lambda^{-1}P(n-1)\upsilon(n))/(1+\lambda^{-1}\upsilon^T(n)P(n-1)\upsilon(n))$, and the desired response $d(n) = \hat{X}^i(n)$ should be synchronized in a sequence to generate the corrected output $\hat{X}_r$. We set the RLS parameters as $\lambda = 0.99$ and $P(0) = 10^{-3}I$ following (Shimazaki and Hara, 2015), and a filter order of $N = 32$ (Han and Kim, 2012).

4.3.2 BHW bandwidth adjustment

The corrected signal $\hat{X}_r(k)$ or the out of the RLS method is then adjusted between $[0.9 - 2.5]Hz$ reducing the phase non-linearities and the stopband ripple (McDuff, Gontarek, and Picard, 2014). Subsequently, a 4-term model BHW filter is applied to achieve desired stopband attenuation of $A_s = -60dB$. Equation 4.10) described the BHW bandwidth adjustment using the BHW impulse response $\dot{h}(n)$ truncated in $N = 150$ order, and a stop-band attenuation $\hat{A}_s = -52.66dB$ with a maximum stopband ripple of $|\delta_s| = 2.32dB$. When we compared with other FIR filters such as Rectangular or Bartlett, the BHW yields superior stopband attenuation and lower ripple variance.

$$\begin{aligned} \dot{X}(k)^i = \sum_{n=0}^{N-1} \hat{X}_r(k-n)^i \dot{h}(n) \left[0.3587 - 0.4883cos\left(\frac{2\pi n}{N-1}\right) \right. \\ \left. +0.1413sin\left(\frac{4\pi n}{N-1}\right) - 0.0116cos\left(\frac{6\pi n}{N-1}\right) \right] \end{aligned} \tag{4.10}$$

4.3.3 Spectral Peak Tracking

(Zhang, Pi, and Liu, 2014) support the usage such as Sparse Signal Reconstruction (SSR) techniques specially the Focal-Underdetermined-System-Solver extension (M-FOCUSS) (Cotter et al., 2005) to increase the level of numerical sparsity, and increase the spectral peak identifiability between no signal spectrum peak and the real HR peaks (Gorodnitsky and Rao, 1997). Multiple MA can increase the error probability in a HR peak detection process. The performance of HR peak detection can be substantially improved using a *peak tracking* process Sun and Zhang, 2015. Our proposed spectra peak tracking method is divided in two stages: 1) A Peak Selection process described in Algorithm 1, and 2) a Peak Verification process described in Algorithm 2.

All possible HR peak spectrum candidates are selected using the Algorithm 1. This process starts with an FFT (*FFTfunc*) for PPG and Accelerometer channels separately. In the Algorithm 1 the *GetNpeaks* subprocess find the possible HR peak candidates (N_{peaks}) on the BVP signal $\dot{X}(k)^i$ spectrum. Each peak selected is a local maximum above 30% of the normalized spectrum amplitude.

For each selected peaks the corresponding accelerometer spectrum peaks $Accel(n)$ are subtracted from the BVP signal spectrum amplitudes $HR(n)$. If the difference between these spectrum amplitudes is lower than 0.10 the subprocess *searchHRpeak* is executed. The process *searchHRpeak* is a function for inner and more specific peak selection associating a search direction depending on flag activation given by the Peak Verification process. Depending of the verification flag calculated by the Algorithm 2, the subprocess *searchHRpeak* selects the first peak from left-to-right using the function $increase_{peak}$, or in the opposite direction using the function $decrease_{peak}$. These functions return the variance associated with new truncated peak candidates in the HR vector.

The HR Peak Verification is a subordinate role returning every possible HR peak candidate

Algorithm 1 HR Peak Selection Process denoted as *PeakSelection*

```
$Input : PPG, ACCEL, NFFT$
$HR \leftarrow FFTfunc(PPG,NFFT)$
$Accel \leftarrow FFTfunc(ACCEL,NFFT)$
$N_{peaks} \leftarrow GetNpeaks(HR)$
while $n < N_{peaks}$ do
    if $(abs(HR(n)) - abs(Accel(n)) > 0.10)$ then
        $[HR, HR_{var}] \leftarrow searchHRpeak(HR)$
        $return([HR, HR_{var}])$
    else
        $n \leftarrow n+1$
    end if
end while
```

($HR(k)$) and their corresponding variance (HR_{var}) associated with the current and truncated BVP normalized frequency segment $X\hat{(k)}$.

Each HR peak found by the *PeakSelection* function is evaluated on the Algorithm 2. The Algorithm 2 start with the four initial HR selected peaks assigned by default as the maximum HR peak candidate inside $[0.9 - 2.5]$ Hz range, and depending on the HR selected peaks the $increase_{peak}$ or $decrease_{peak}$ functions activate a flag that will be received by the *searchHRpeak* process, and thus proceeding as we explained above in Algorithm 1. The BHW bandwidth adjustment process $change_{BHW}$ function is applied iteratively narrowing the

Algorithm 2 HR Peak Verification process

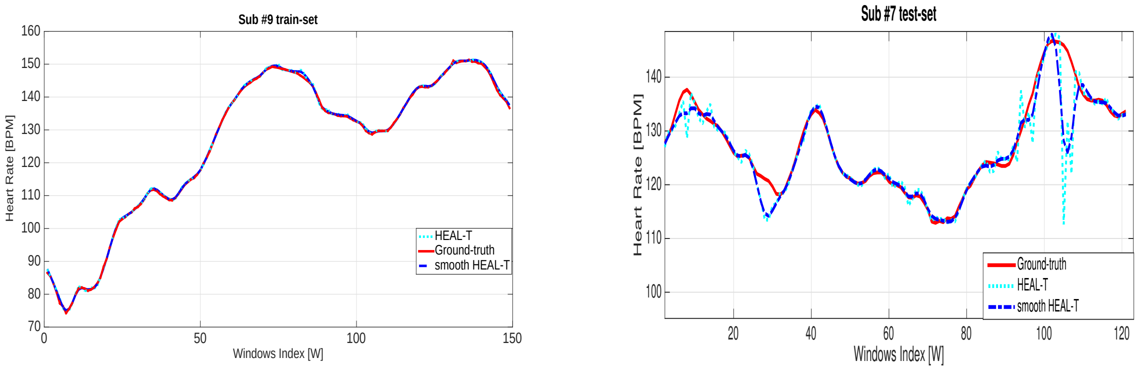
```
$Input : PPG, ACCEL, n_{window}, NFFT = 65,536$
$k \leftarrow 1$
while $k < n_{window}$ do
    $[HR(k), HR_{var}] \leftarrow PeakSelection(PPG(k),ACCEL(k),NFFT)$
    if $k > 4$ then
        if $(HR_{var} < 0.10)$ then
            $threshold \leftarrow 0.05$
        else
            $threshold \leftarrow 0.10$
        end if
        if $HR(k) <= mean(HR(k-4:k-1)) - threshold * mean(HR(k-4:k-1))$ then
            $HR(k) \leftarrow increase_{peak}(HR(k))$
            $change_{BHW}(HR, \omega_p + 0.30, \omega_s)$
        end if
        if $HR(k) >= mean(HR(k-4:k-1)) + threshold * mean(HR(k-4:k-1))$ then
            $HR(k) \leftarrow decrease_{peak}(HR(k))$
            $change_{BHW}(HR, \omega_p, \omega_s - 0.30)$
        else
            $k \leftarrow k+1$
        end if
    end if
end while
```

BHW bandwidth adding or subtracting $0.30Hz$ to ω_p and ω_s being this HR peaks initialized using the Equation 4.11 also depending on the four previous HR peak amplitude.

$$[\omega_p, \omega_s] = \begin{cases} [0.9, 2.5]\text{Hz} & \text{if } HR(1) \leq 120BPM \\ [1.7, 3.5]\text{Hz} & \text{if } HR(1) > 120BPM \end{cases} \tag{4.11}$$

4.3.4 IBI and HR estimation Results

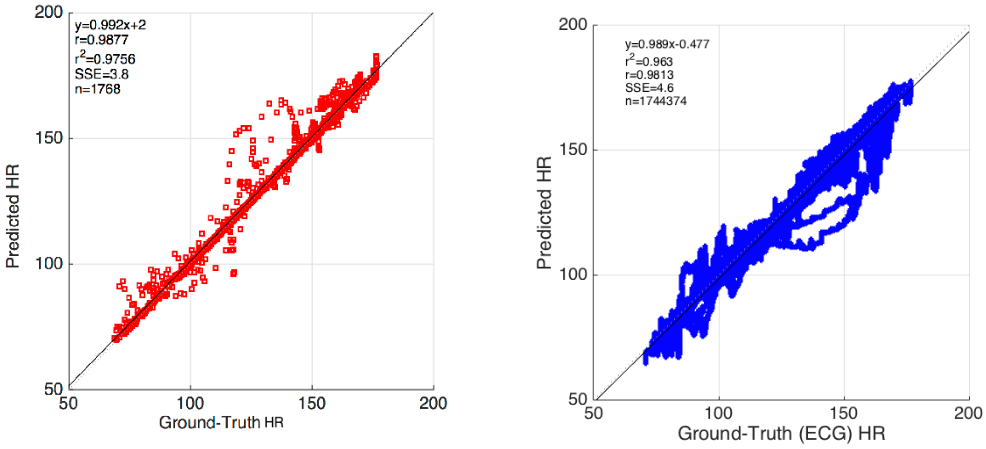
The variance of the HR peak candidates HR_{var} is used to set a new $threshold$ value as follows: If the HR_{var} value returned by the Peak Search process is lower than 0.10 the $threshold$ will be equal to 5%, otherwise, the $threshold$ will be equal to 10%. Thus, any peak candidate which overcomes all these conditions are accepted as definitive HR values per signal window.



(a)

(b)

Fig. 4.8 HR estimation results for subjects #9 from the TROIKA training set in Figure 4.8a, and #7 from the TROIKA test set in Figure 4.8b.



(a) (b)

Fig. 4.9 Bland-Altman plots for HR estimation using the HR reference in Figure 4.9a, and using the ECG groundtruth for IBI estimation in Figure 4.9b, this latter having more points

For the IBI signal estimation we grouped the dichrotic notches from the BVP signal, and the corresponding R-peaks of the ECG signal groun-truth. The selected notches may be above 50 % in amplitude for each given window. Subsequently, for this IBI estimation process we calculate the time-difference between the adjacent/nearest dichrotic notches.
To estimate the IBI from the BVP signal we first compensate the window overlaps, and calculate the IBIs between the BVP notches and on the overlap IBIs we calculate the averaged between both overlapped windows while we group the values from the non-overlapping segments. Subsequently, we use smoothing Cubic spline to calculate the Interpolated-IBI (IIBI) signal, and thus reduce undesired IBI spectral harmonics and discontinuities.
For the HR estimation performance, and the computational efficiency measurement, we use the LMS using the same parameters in (Han and Kim, 2012), and the TROIKA framework (Zhang, Pi, and Liu, 2014) as baselines.

Table 4.5 Absolute error for each subject for our approach, LMS baseline (Han and Kim, 2012) and the TROIKA framework (Zhang, Pi, and Liu, 2014) baselines. The results report significances $p < 0.01$ in bold italics values.

Subjects	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**Avg**	**SD**
(Torres et al., 2016)	3.96	**1.73**	***0.91***	**2.21**	***0.32***	***1.19***	***0.32***	***0.47***	***0.26***	4.22	**0.87**	**1.41**	**1.49**	1.36
(Han and Kim, 2012)	5.21	2.22	1.45	3.44	0.88	3.42	0.58	1.33	2.45	4.55	1.21	4.33	2.59	1.57
(Zhang, Pi, and Liu, 2014)	2.87	2.75	1.91	2.25	1.69	3.16	1.72	1.83	1.58	4.00	1.66	3.33	2.40	0.80

Table 4.6 IBI-based HR estimation for our approach, the LMS in (Han and Kim, 2012), and the TROIKA framework (Zhang, Pi, and Liu, 2014) base. The results report significances $p < 0.01$ in bold italics values.

Subjects	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**Avg**	**SD**
(Torres et al., 2016)	**5.42**	**4.54**	**2.53**	**3.39**	**2.55**	**3.04**	**2.26**	**2.58**	**2.75**	**5.04**	**3.46**	**3.21**	**3.40**	1.05
(Han and Kim, 2012)	6.41	5.64	6.01	5.43	2.88	4.12	4.08	3.45	3.88	6.46	5.55	5.75	4.97	1.22
(Zhang, Pi, and Liu, 2014)	6.55	5.43	5.12	4.45	2.81	3.78	2.78	3.33	3.42	6.74	4.52	4.64	4.47	1.32

Table 4.7 Absolute error for each subject for our approach using the TROIKA test-set, LMS in (Han and Kim, 2012), and the TROIKA framework (Zhang, Pi, and Liu, 2014) baselines. The values significantly different for $p < 0.01$ are in bold italics.

Subjects	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**Avg**	**SD**
(Torres et al., 2016)	**5.43**	**4.54**	**6.71**	**3.01**	**2.71**	**5.37**	**1.39**	**0.92**	**3.21**	2.10
(Han and Kim, 2012)	7.42	4.54	15.34	5.47	5.35	10.35	2.56	3.45	6.72	4.54
(Zhang, Pi, and Liu, 2014)	5.78	4.33	12.45	3.79	3.09	7.74	4.56	2.42	5.47	3.53

The HR estimation performance is evaluated in two modalities: 1) comparing obtained HR spectrum peaks with the HR groundruth per window, and 2) comparing the extracted IIBI defined as $HR_{IIBI} = 60/IIBI$ to compare these calculations with the ECG RR-peak distances. To calculate a proper RR-peak using the ECG ground-truth we filtered the ECG signal wth a Daubechies Wavelet filter with level 3, and order 3. The HR estimation performance evaluation are reported in Tables 4.5 for the TROIKA training-set subjects, and 4.7 for the TROIKA test-set subjects. The IBI-based evaluation modality results are reported in Table 4.6. The ECG signal is not available for the test set, therefore, we report the IBI-based evaluation only for the training set.

$$Error_k = |HR_{predicted_k} - HR_{groundtruth_k}| \tag{4.12}$$

For the HR estimation per window we obtain a BPM absolute error of 1.49 ± 1.36 for the TROIKA training-set , and 3.21 ± 2.10 for the TROIKA test-set. The absolute errores were calculated following Equation 4.12 per window k. Evaluating our approach we obtained combined absolute error value of the train-test set average 2.25 ± 1.93, For the LMS ,and the M-FOCUSS baselines the combined absoluted errors are 4.08 ± 4.13 and 3.64 ± 2.59. For the IBI evaluation we obtain a BPM absolute error of 3.40 ± 1.05 reported on Table 4.6.
Figures 4.8a and 4.8b show the performances of the proposed algorithm in time-domian for subjects # 9 on the TROIKA training-set, and subject # 7 on TROIKA test-set -in red- comparing them with the HR ground-truth in blue. Figures 4.9a and 4.9b present the Bland-Altman plots (Euser, Dekker, and Cessie, 2008) for the two modalities evaluated here such

as for the HR prediction obtaining a Pearson $\rho = 0.9877$, and for the IIBI v.s ECG modality a $\rho = 0.9813$.

4.3.5 Computational Load

In comparison with the SSR techniques this pipeline is less computationally demanding. In this study we evaluate a preliminary estimation of the computational load using the Matlab R2015a profiler to compute the execution time of this approach thus varying the BVP window size. The execution times reported in Figure 4.10 are averages of 20 different executions per window (Altman, 2014). In Figure 4.10 the LMS adaptive filtering, and the M-FOCUSS based pipelines were set with the learning parameters as $\lambda = 0.1$, $\gamma = 10^{-4}$, and $max_{iter} = 500$ (Cotter et al., 2005).

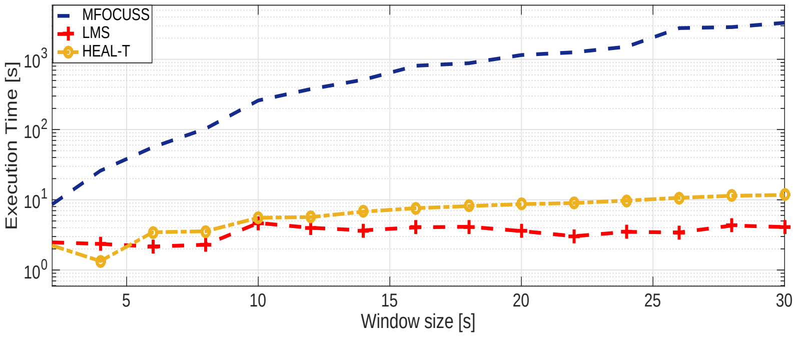


Fig. 4.10 Comparison between our approach, the LMS (Han and Kim, 2012), and the M-FOCUSS (Cotter et al., 2005; Zhang, Pi, and Liu, 2014) baselines in terms of execution time in the log Y-axis, and the signal window size from 2 up to 32s in the X-axis.

For this computational load comparison we pick the subject # 10 from the TROIKA training-set due to it is the worst case measurement scenario in terms of the number of calls from the Matlab profiler evaluation our approach. The BVP window size is varied from 2 to 32 seconds with the 33% overlap in the x-axis of Figure 4.10. We can observe that our approach execution time is 2 orders of magnitude lower than M-FOCUSS, and closer to the LMS computational load minimum quota (Choudhury et al., 2014). These results support the consideration of our HR+IBI estimation pipeline as a more suitable and accurate solution for real-time HR wearable devices.

4.4 HR and IBI calculation for Empatica Device

For this approach we evaluated our previous explained approach (Torres et al., 2016), denoted as HEAL-T, thus applying the signal cleansed processes explained above on a real-life study to characterize and classify Hypertension patients (Ghosh et al., 2015). The same pipeline described in HEAL-T was applied to BVP and Accelerometer signals collected from an Empatica E3 device during multiple days recording trials.

Our study was considered a Pilot Study including 10 Hypertensive and 10 Normotensive (Healthy Control) adults recording 10 days each. The adults included in the Pilot study were between the age of 30 and 65, and with the purpose of the clinical analysis we only include patients with diagnosed essential (EH, n=8) who received treatment at the *Centro Ipertensione Ospedale Molinette*, in Turin, Italy. The Normotensive (HC) and the Hypertension patients were diagnosed by a psychologist to rule out hidden hypertension, and the hypertension severity or the evidence of other disorder comorbidities which can affect the study. The institutional ethics committee of the *Azienda Ospedaliera Citta della Salute e della Scienza di Torino* and the ethics committee of the *Universita degli Studi di Trento* approved this research study.

For 10 days the BVP and Accelerometer data were recorded from the Empatica E3 wristband continuously. The main idea of this study was to monitor the participants during their work day where the level of stress is higher. The participants have recording oscillating the complete 24 hours recording, starting at morning before going work and until they went to sleep. Along the experiment the participants answered questions and took notes regarding their mental state, and the current activity using a mobile-agent application on the cellphone.

To balance the dataset being modified to include only eight participants with essential hypertension we selected other eight normotensive participants sorting them in terms of the quality of the signals collected. A total of 756 hours of data from the hypertensive patients and 780 hours from normotensive subjects have been included in this study.

With a purpose to enhance the classification accuracy our study includes other signals from the Empatica E3 device synchronizing them due to the sampling frequency differences. Signals such as Galvanic Skin Response (GSR), and Skin Temperature (ST) sampled at 4Hz were processed using low-pass Butterworth filter between [0-16] Hz, and linear detrending operator.

We apply a modification of the HEAL-T pipeline to optimize the quality of BVP signal from the PPG sensor on the Empatica E3 device following the pipeline on Figure 4.11. Using all the recordings from the EH and HC participants we evaluate the TROIKA framework using M-FOCUSS (Zhang, Pi, and Liu, 2014), and our pipeline with the impossibility to compare our performances with a ECG or HR ground-truth reference, however, the performance for hypertension detection is given better metrics using our pipeline in comparison with the TROIKA framework.

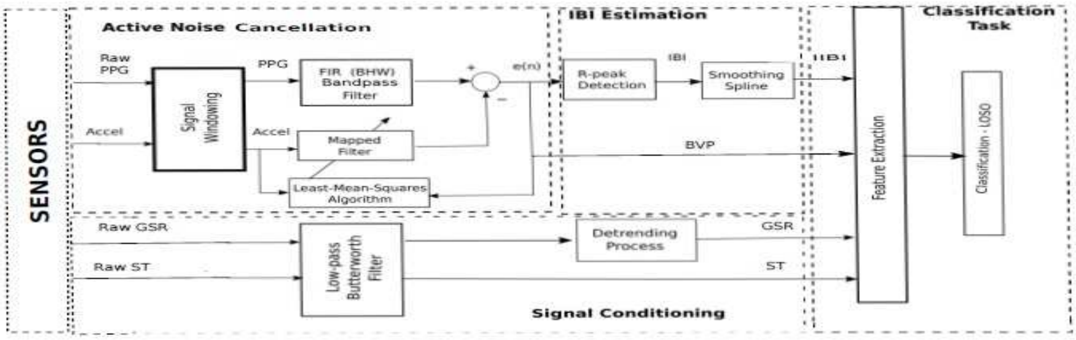


Fig. 4.11 Pipeline for Hypertension prediction using wearable Devices signals. This pipeline conforming blocks are: 1) Active noise Cancellation based on LMS, 2) R-peak IIBI estimation using PPG decontamination, and 3) GSR and ST signal features extraction.

The only difference with the HEAL-T process is the active-noise cancellation using the LMS for artifact removal proposed in (Han and Kim, 2012). Specifically, This method consists of a LMS adaptive algorithm as we mentioned above, and from a initial conditions we minimize the error with respect to the desired filter desired response represented with a FIR filter with [0.5-5] Hz bandwidth.

4.4.1 Feature Extraction

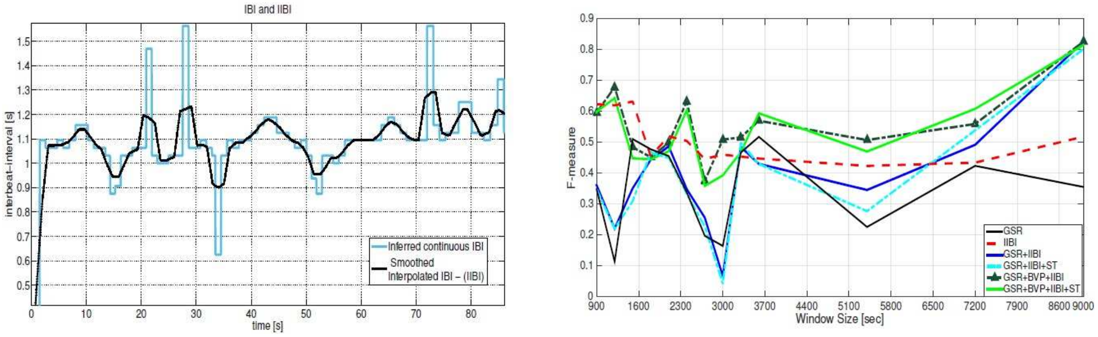
For the feature extraction we use a windowing range from 15 minutes to 2 and half hours across the entire trial per subject. Concordantly, we proceed with the feature extraction using features from the cleansed BVP, IIBI, GSR, and ST signals after pre-emphasis. We will describe the features extracted as follows:

1. **Cleansed BVP:** we calculate the statistical features such as mean, SD, min and max per trial.
2. **GSR:** we calculate statistical as well mean, SD, min and max, and instantaneous change features such as the duration and amplitude of a the startle response on the GSR signal trial. Features from Skin Conductance Level (SCL) are also extracted following (Ghosh, Danieli, and Riccardi, 2015).
3. **ST:** we extrated the mean, standard deviation, maximum, and minimum, for non-normalized and the normalized signal.

4. **IIBI:** The IIBI signal positivities are correlated with sympathovagal activation and the sympathetic baroflex function activation. 17 time-domain features per window such as: the minimum of the Heart Rate estimated per window, Root Mean Square of the Successive Difference of the NN interval (RMSSD) obtained from the cleansed BVP dichrotic nothces, Standard Deviation of the NN interval (SDNN), Percentage of Consecutive NN intervals which differ by more than 50 (pNN50), and 30 (pNN30) milliseconds. On the other hand, some frequency-domain features related to sympathetic and parasympathetic neural activity are extracted from the IIBI signal such as: the ratio of the Low Frequency and High Frequency (LF/HF), and the statistical features for Low and High frequency ranges such as mean, variance, max and min peaks. An example of IIBI trial is shown in Figure 4.12a

4.4.2 Classification Essential Hypertension

Our classification problems is a 2-class hypertensive and normotensive subject detection used a Leave One Subject Out (LOSO) cross-validation. Each test fold contains trials from either a hypertensive subject or a normotensive participants exclusively. We calculate the confusion matrix by combining the individual classes per fold for each subject. True Positives across the confusion matrix per subject contribute as hypertensive subjects classified correctly as hypertensive, and viceversa. We perform classification with both individual and combined signal features. We evaluate feature-level fusion from all the different signals on five different classifiers such as kNN, Naive Bayes, Decision Trees, Linear Kernel SVM, and two ensemble learning algorithms such as Adaptive boosting (Adaboost), and Random Forest classifiers. The ensemble based classifiers outperform the rest of the classifiers for both individual and fusion of features, being Adaptive Boosting the best performance.



(a) (b)

Fig. 4.12 IBI and IIBI signals for a long trial in the Essential Hypertension detection in Figure 4.12a, and the performance for all the feature-level combinatios evaluated in this study such as GSR, IIBI, GSR+IIBI, GSR+IIBI+ST, GSR+BVP+IIBI, and GSR+BVP+IIBI+ST

Considering the single BVP signal's features we always obtain low classification accuracies, with BVP features being the highest F-measure in 0.62. However, the feature-level combination of different signal improves significantly the classification results. GSR-IIBI and GSR-BVP+IIBI combinations provide the best discrimination between hypertensive and normotensive/Controls participants. The best F-measure of 0.83 obtained feature-level combinations is using a features from the cleansed BVP, GSR and IIBI signals. The F-measure variability in terms of window size is plotted in Figure 4.12b showing only the combination modalities. The F-Measure is calculated using the Equation 4.13 using each subject *k* resulting confusion matrix.

$$F_{measure_k} = \frac{2 * Recall_k * Precision_k}{Recall_k + Precision_k} \tag{4.13}$$

4.5 HR and IBI calculation in industry One-LVL company

In 2018 the company One-LVL, located in Austin, Texas, contacted us in the Signals and Interactive System Lab (Sislab), in Trento, Italy with the purpose. The purpose of the contact was to develop a system that can compete with other top systems which can measure HR and IBI reliably using light implementation algorithms on wearable devices, more precisely a wristband.

For a month we developed and enhanced the HEAL-T pipeline (Torres et al., 2016) with the most realistic data possible acquired from One-LVL company. We enhanced the HEAL-T performance based on 3 important items for HR prediction for an in-the-wild reported by (Sun and Zhang, 2015) as follows, the itemize has been taken literally from the paper:

- "About 75% of spectral peaks that have good amplitude are true peaks. Peaks with good amplitude are defined as peaks with the highest amplitudes in their corresponding time windows."
- "About 84% of peaks that have good positions are true peaks. A peak with a good position refers to the one with the shortest distance from its previous true peak"
- "About 96% of true peaks have good amplitude and good position..."

Assuming these conditions from a modification in HR searching peak remarks explained in the previous section it will be possible to obtain good performances even if the Accelerometer is modelled with an adaptive filter or not. For this implementation we adjust the possible HR ranges per subject adjusting manually the possible HR ranges for subjects that shown a large change in estimated HR spectrum points.

An important anomaly observed to do the regression, is the multiple conditionals should be done to separate the HR spectrum with the contaminated signal included Accelerometer spectrum even after we implement an adaptive filter modelling the Accelerometer as noise as we explained above. Figure 4.13 show how BVP and Accel spectrum are entangled. The M-FOCUSS implementation assures a sparse spectrum but requires a large computational complexity and resources (Zhang, 2015) to be used in from a wearable device.

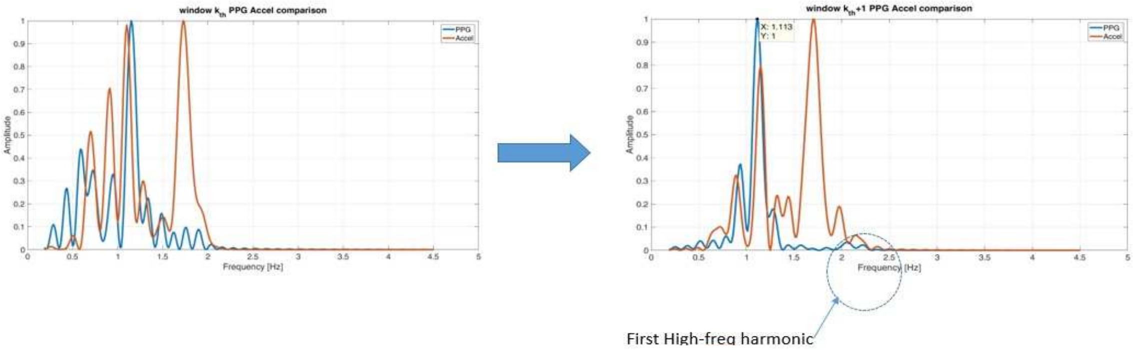
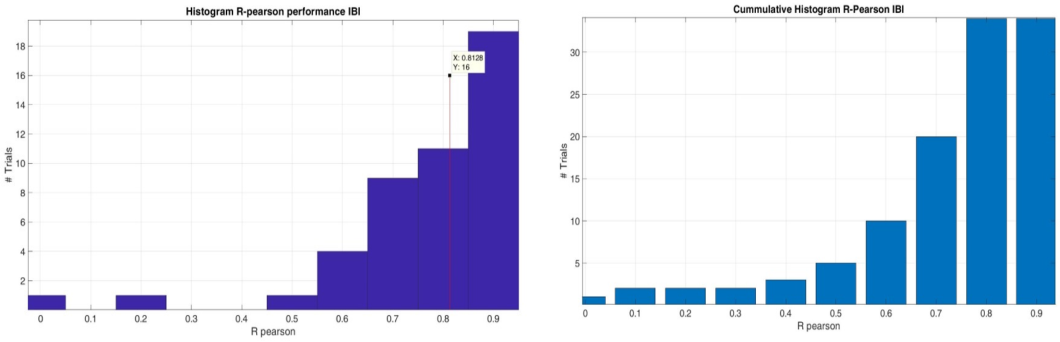


Fig. 4.13 Accelerometer and BVP spectrum changes from a window k_{th}, to a window $k_{th}+1$.

For this One-LVL competition they share 100 trials collected from 16 possible events per trial. 50 out of the 100 trials were selected for this evaluation excluding two sessions data with excessive or unknown movement artifacts. Lie, Sitting, Start Computer and Lie trials were included in this analysis where we found unexpected and non-periodic movements from the Accelerometer. The main objectives of this phase I study for One-LVL evaluation were:

1. Analyze in-vivo Photoplethysmographic (PPG) data from portable wristband prototype for estimating Interbeat-Intervals (IBI) and Heart-Rate (HR) robustly.
2. Analyze errors, motion artifact, or data inconsistencies presented in the data collection assuming events in real life scenarios.
3. Achieve a minimum of Pearson $R = 0.9$ measuring IBI-IBI or HR-HR correlation, and analyzing twelve HR-related metrics proposed by One-LVL.

Grouping all the results found from the 50 trials analyzed in phase I we obtained the histogram, and cumulative histograms in Figures 4.14a and 4.14b from the complete twelve HR-based outcome measures proposed by One-LVL researchers such as: very-low Frequency average (vLF), amount value of the Low-Frequency ranges (aLF), amount value of the High-Frequency ranges (aHF), amount value for the entire spectrum (aTotal), Percentage of power High-frequency range (pHF), Normalized Low-frequency power (nLF), Percentage of sucessive differences greater than 10 milliseconds (pNNx), standard deviation of nearest notches interval (SDNN), standard deviation of sequential 5-minute notches interval average (SDANN), root mean square of successive notches differences intervals (RMSSD), standard deviation of the Poincaré plot first dimension (SD1), and standard deviation of the Poincaré plot second dimension (SD2) (Handouzi et al., 2014).
We also included the R-Pearson values for the Bland-altman plot that proposed by our individual research. The complete results for this Bland-Altman analysis with the corresponding difference plots are reported in Figure 4.15. The overall results averaging the correlation R-Pearson per trial is reported in the first row of the Table 4.8, the standard deviation results calculated from the R-Pearson values per trial are reported in the second row of Table 4.8. Ironically, when the HR predicted values for all the trials are concatenated assuming a sequence as One-LVL proposed we can observe some metrics such as aTotal, SDNN, SDANN, SD2, and R-IBI greater than 0.90, being this latter the R value from the Bland-Altman IBI shown in Figure 4.15a.

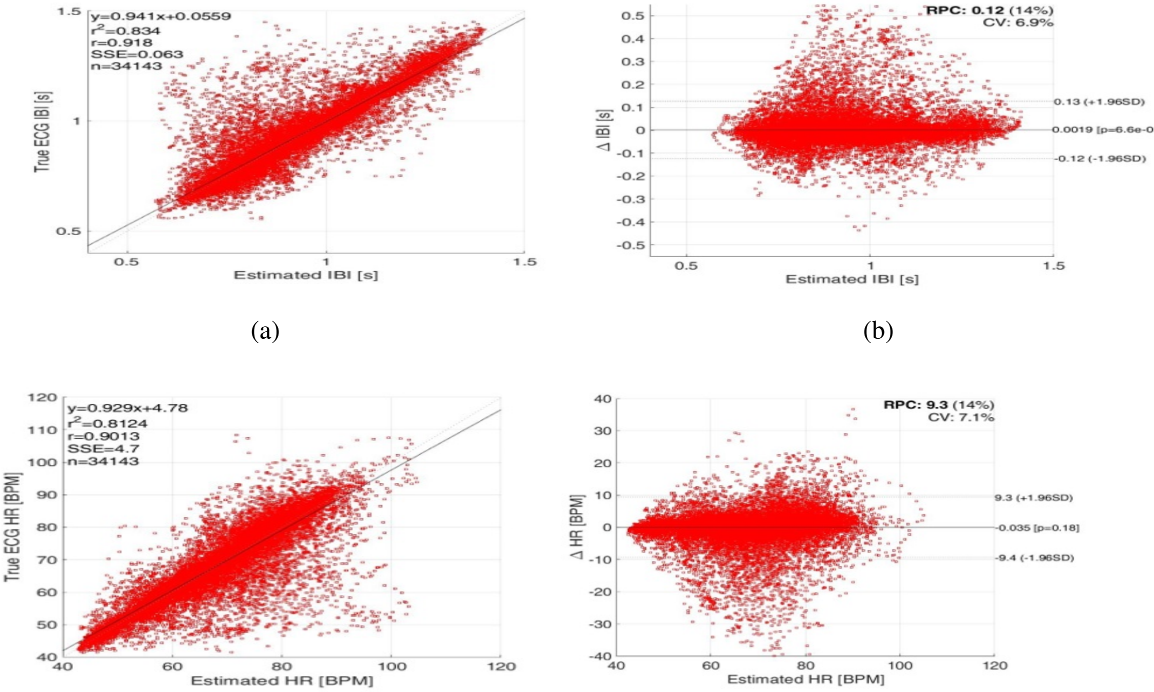


(a)

(b)

Fig. 4.14 Histogram for all the windows and all the 12 R-Pearson values associated with HR-based outcome measures proposed by One-LVL.

The Bland-Altman plot is a measurement of agreement proposed for a variability-based quantification of two outcome measures distribution, and how these two measures are more or less constant between across themselves, and no whether or how much these outcome measures are close enough (Bland and Altman, 2002; Giavarina, 2015). For our particular research this measure is more adequate for a HR and IBI calculation in-the-wild.



(a)

(b)

(c)

(d)

Fig. 4.15 Bland-Altman plots for IBI (Figure 4.15a) and HR (Figure 4.15c) prediction v.s the IBI and HR deduced from the ECG ground-truth given by One-LVL company. In Figures 4.15b and 4.15d we show the differences plot and how separated are the prediction in comparison with the ground-truth.

The Bland-Altman plotting can be considered a measure of agreement more desirable in long term for clinical trials due to its flexibility and the clinical trials need for measuring bias between two outcome measure means. Some studies suggest that a measured or p-values or R-Pearson are not following a correct matching between variables' means, in contrast with the Bland-Altman's R value. Our results confirms the importance of measures such as Bland-Altman metrics instead of other indicadors from spectrum metrics such as HF, or LF outcome measures that are showing only a partial evaluation of the HEAL-T pipeline on a realistic environment.

Table 4.8 Average and standard deviation results for the R-Pearson across the 12 HR-based proposed outcome measures, plus the R value obtained from the Bland-Altman plot denoted as R-IBI

R/Metr	**vLF**	**aLF**	**aHF**	**aTotal**	**pHF**	**nLF**	**pNNx**	**SDNN**	**SDANN**	**RMSSD**	**SD1**	**SD2**	**R-IBI**
Av R	0.807	0.686	0.390	0.845	-0.049	-0.048	0.356	0.888	0.855	0.822	0.827	0.888	0.810
STD R	0.378	0.472	0.675	0.324	0.723	0.734	0.681	0.213	0.336	0.304	0.294	0.212	0.109
Conc. R	0.894	0.842	0.548	**0.918**	0.112	-0.021	0.703	**0.947**	**0.908**	0.855	0.855	**0.946**	**0.918**

Chapter 5

ConvNet Pipeline for EEG-based Enhanced Emotion Decoding in Autism

This chapter starts the central and main part of this dissertation where we will describe the Deep ConvNet pipeline for EEG-based successful emotion decoding including Autism population from clinical lab-control and time-locked trial.

Due to a concrete and evident necessity of measure the statistical incidence of multiple neuro-correlates included in the FER elicited ASD group neural activity, we propose to measure and compare the neurocognitive effects, and the performances obtained from a FER task elicited by DANVA 2.0 imageset (Nowicki, 2000) with a novel Deep ConvNet pipeline to decode multiple emotion categories in a robust way (Schirrmeister et al., 2017; Weitz et al., 2018).

Using our Deep ConvNet architecture we outperform significantly the FER human performance in children and adults ASD, thus supporting the Deep ConvNet as a good candidate for perceptual classifier capable to fill the multiple FER behavioral deficits observed in ASD groups.

In the following subsections we will describe the methodologies use for construct and train the Deep ConvNet pipeline for emotion decoding using EEG features, the corresponding performances intra-subject, and across subjects, considering quantifiable comparison with the corresponding FER human performances for each subject, and across all the participants for non-ASD/TD controls and ASD participants.

5.1 Demographics and Behavioral variables

For this study we include a complete sample of 192 participants to evaluate the classifier performance. This 192 participants sample is divided in three subsamples: 1) A eighty-eight children participant sample to evaluate the performance comparison, and the feature importance analysis. This sample was composed of fourty-eight non-ASD/TD participants, and 40 ASD participants with a clinical age $Age = 15.69 \pm 1.28$ for TD participants, and $Age = 14.47 \pm 1.55$ for ASD.

For performance results replication, specifically, for the TD/ASD Deep ConvNet performance generalization, and the ADOS Calibrated Severity Scale (ADOS-CS) statistical correlation analyses we include two complementary participant samples such as 2) a sample including sixty-nine adults participants composed of fourty-two non-ASD/TD and twenty-seven ASD participants with a clinical age of $Age = 20.74 \pm 3.11$ for the TD group, and $Age = 22.97 \pm 4.94$ for the ASD group, and 3) a third sample including only thirty-six ASD participants with a clinical age between 3 and 16 years old 13.41 ± 1.96.

All the three participant samples grouped are a total of 192 TD/ASD participant included in this study, and we included the Intellectual Quotient (IQ), the percentage of male and female per group, and ADOS-CS being this latter significantly different between groups for the sample # 1 (F(1,87)=2.345, p=0.0434) and # 2 (F(1,68)=2.001, p=0.0415). The IQ measure was not significantly different for all the samples. Along this chapter we will report the Deep ConvNet performance results for the three samples described here.

Samples	**Sample #1**				**Sample #2**				**Sample #3**	
	TD N = 48		**ASD N = 40**		**TD N = 42**		**ASD N = 27**		**ASD N = 36**	
	μ or #	**σ or %**	**μ or #**	**σ or %**	**μ or #**	**σ or %**	**μ or #**	**σ or %**	**μ or #**	**σ or %**
Age (years)	15.69	1.28	14.47	1.55	20.74	3.11	22.97	4.94	13.41	1.96
Male N and %	29	60.42%	32	80.00%	31	60.42%	20	80.00%	25	69.44%
ADOS-CS**	3.33	2.71	8.15	2.05	4.45	2.99	7.77	2.34	8.08	3.04
IQ	107.82	14.03	100.78	16.54	115.77	17.74	113.56	16.78	116.66	15.55

Table 5.1 Demographics, ADOS-CS, and IQ for all the participants across all the samples included in this study.

5.1.1 Experimental Protocol and Data Collection

48 emotional faces from DANVA-2 image-set (Nowicki, 2000) were presented to all the 192 participants grouping all of them across the three participant samples mentioned above. The

participants had EEG visits in the Social Competence and Treatment Lab (SCTL), in Stony Brook University, NY, USA under the approval of Stony Brook Medicine Ethical Commitee. The participants performed a FER task annotating the presented face between four different emotions such as happy, sad, angry, and fear. DANVA-2 emotional face dataset does not contain a neutral mental state, implying a variety of four different emotional states and a clean elicitation of four extreme mental/emotion states dedicated to evaluate TD/ASD neural and behavioral outcome measures.

EEG signals were recorded using a BrainVision 32 channels ActiChamp recorder system downsampling the trials from 1KHz to 500Hz sampling frequency. The two electro-oculogram (EOG) channels were removed from the initial analysis focused on the Prep pipeline as we will describe below, but included again for ADJUST blinking artifact removal. The rest of the pipeline execution use the 30 EEG channels only including neural activity after the pre-emphasis and artifact removal methodologies as we will explain below.

The data were collected in anonymous way only identifying each participant with automatic generated code, and the behavioral files containing the annotation, and the neural files including EEG signals were saved separately but identified with the corresponding code per participant.

5.2 ConvNet pipeline Description

The complete pipeline for EEG-based pipeline is shown in Figure 5.1 and it is composed of four important substages such as 1) the EEG filtering stage or the pre-emphasis stage using signal conditioning methods, 2) the artifact removal phase, 3) the ZCA whitening transformation, and 4) the training of the Deep ConvNet described in Figure 5.1 and the sections below.

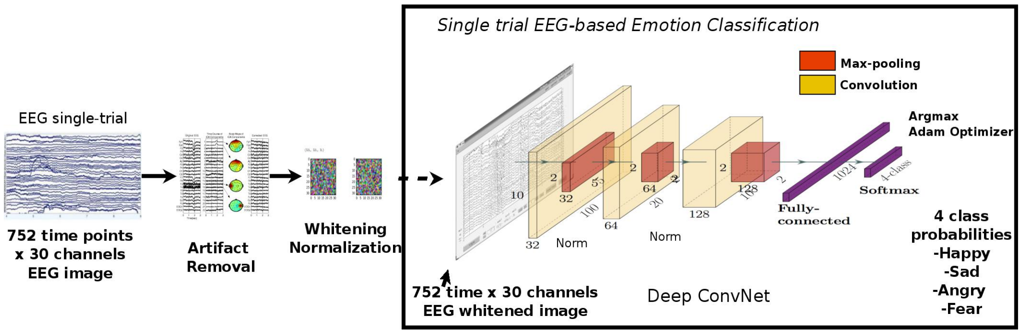


Fig. 5.1 Pipeline for emotion decoding composed 1) EEG Filtering and Pre-emphasis, 2) the artifact removal process composed of a subsequent usage of the Prep pipeline including Koethe's cleanraw and Artifact Subspace Removal (ASR) for bad-channel removal, and the ADJUST EEGlab plugin for automatic noisy ICs removal, 3) the ZCA whitening normalization process to increase the class separability and the high-frequency neural activity excerpts per trial, 4) the Deep ConvNet composed of 3 conv-pool layers going from high to low in terms for conv-pool dimensionality, and low to high in terms of the number of filters per conv-pool layer, two local normalization layers, and a fully connected layer with 1024 units.

5.2.1 EEG Filtering and Artifact Removal

Each raw EEG trial was processed using EEGlab (Delorme and Makeig, 2004) Matlab toolbox assigning an EEGlab structure per participant/code, and for each emotion. We used a 150 coefficient Blackman-Harris-Window band-pass filter with a pass-band between [0.1-30] Hz. The filtered EEG segments composed of 32 channels were initially referenced to Cz, and re-referenced to maximize the neural activity in superior channels such as T9 -T10 obtaining a new 30 channels EEG structure. Each channel was composed of 875 time-points based on a 500Hz sampling rate. The EEG time-locked trial cover the time range between -200 and 1550 ms relative to the stimulus onset. Each structure groups the 12 corresponding epochs associated with a specific emotion. Concordantly, each EEG structure had a data field with a size of 30 channel × 875 time-points × 12 emotion epochs conforming a total of 48 epoched trials, and 4 structures per participant one for each emotion.

The neural activity baseline was removed between -200 and 0 ms relative to the onset. A bandpass filter with a passband between 0.1 and 30Hz was applied to each EEGlab structure to conserve important neural activity on face perception such as N170, and LPP windows(Dawson, Webb, and McPartland, 2005; Dawson et al., 2002; Mayor Torres et al., 2018). Subsequently, an automatic channel rejection and an artifact removal process were applied to each EEGlab structure.

First, the Prep pipeline (Bigdely-Shamlo et al., 2015) was used to remove noisy and artifactual channels based on the Koethe's cleanraw function and Artifact Subspace Removal

(ASR) method (Torres et al., 2018). Second, an Independent Component Analysis (ICA) decomposition (Hyvärinen and Oja, 2000) was applied to 2D reshaped EEG structure to calculate a decomposition matrix per structure, subsequently, the decomposition ICA matrix and the 3D shaped EEG structure was used by the ADJUST EEGlab plugin (Mognon et al., 2011) to classify artifactual independent components (ICs) using spatio-temporal high order statistical moment ICs' features. If a maximum is observed in features such as *Temporal Kurtosis*, *Spatial Average difference*, *Maximum Epoch Variance*, or *Generic Discontinuities Spatial Feature* a horizontal/vertical eye blink artifact can be detected. Therefore, the ICs classified as artifactual based on the mentioned features were excluded from the subsequent ICA composition process obtaining a clean EEG data structure.

5.2.2 ZCA Transformation

The artifact-free EEG structure was normalized using a ZCA whitening normalization. The ZCA-like whitening normalization is also known as *Mahalanobis Zero Phase Whitening* (Coates and Ng, 2011, 2012) and is used as previous step to create a 2D representation which maximizes the average cross-covariance between each dimension of the whitened X_{zca} and the original EEG cleansed data per trial X. Equation 5.1 represents the new X_{zca} whitened EEG image representation where $S_x = VDV^T$ represents the eigenvalues decomposition of the EEG cleansed matrix epoch composed of channels × time-points data field and denoted with X. ε_{zca} is denoted as the contrast bias to move the resulting EEG image's X_{zca} contrast around the cross-covariance matrix trace. We set ε_{zca} value in 0.01.

$$X_{zca} = \frac{VV^T X}{\sqrt{D + \varepsilon_{zca} I}} \tag{5.1}$$

X_{zca} is then obtained decomposing and integrating the eigenvalues of the cross-covariance matrix S_x. Applying the ZCA whitening normalization we convert a new EEG 2D feature set in a high-frequency amplified "EEG image" which propagates better feature separability across the max-pool layers of the ConvNet (Huang et al., 2018). The ZCA whitening normalization allows a zero-phase and a minimal rotation of the feature input-map changing the amplitude adequately to enhance our current pipeline decode the emotion successfully from the neural activity. This whitening normalization process is iterated across all the 48 epoched trials, and per subject following the cross-validation modality explained below.

5.2.3 Deep ConvNet training

The implementation of the Deep ConvNet was coded in the Tensorflow Python library (Abadi et al., 2016), and the file management for the accessing whitening images was provided by other Python Libraries. The Deep ConvNet was composed of three convolutional-pool layers. The first convulitional-pooling (conv-pool) block had a convolutional-layer with a kernel-size of 100x10 units and 32 filters, and a max-pool layer with a size of 5x2 units connected to local response normalization layer (Schirrmeister et al., 2017). A second conv-pool block was composed of a convolutional-layer with a kernel-size of 20x5 units, and a max-pool layer with a size of 2x2 units also connected to a second local response normalization layer. A third conv-pool block composed of a convolutional-layer with a size 10x2, and a max-pool layer with a size 2x2 times 128 filters depth. Each conv-pool block has a stride factor of 2, and non zero-padding using the option VALID from Tensorflow, and thus dividing the output size for each dimension to a half after each conv-pool block. The third max-pool layer is connected to a fully-connected (FC) *Softmax* layer with 1024 units to compute the final four emotion classes probabilities for *happy*, *sad*, *angry*, and *fear*.
To illustrate a bit the Deep ConvNet arithmetic (Dumoulin and Visin, 2016) we can model the final output-size after a conv-pool block described in Equation 5.2 where i_{conv}, and k_{conv} the input size and the kernel size for each convolutional layer, and k_{pool}, O_{pool}, and S_{pool} the pooling size, the output size after the pooling layer, and the stride factor after the pooling layer respectively. Running this ConvNet training in Seawulf (single process) occupies 12.3% of memory from a Tesla K80 GPU node. On the other hand, running this ConvNet in HPC Trento occupies 55.78% of memory from a Tesla V100 GPU node.

$$O_{pool} = \frac{[i_{conv} - k_{conv} + 1] - k_{pool}}{S_{pool}} + 1 \tag{5.2}$$

Getting into the Deep ConvNet training process, we initialize our conv-pool blocks following an important initialization settings described in (Parcollet et al., 2018). The initialization was used as a critical process for weight and biases values convergence on this type of Deep ConvNet architecture. For our specific case we use the same initialization procedure equivalent to only real modelled Deep ConvNet following Equation 5.3 with a uniform random generated angle θ.

$$\omega = |\omega| cos(\theta) \tag{5.3}$$

Our choice to initialize the convolutional kernels in the first step were the Glorot's uniform initializer or also called Xavier Uniform initializer (Glorot, Bordes, and Bengio, 2011) from Tensorflow. The biases were initialized with random normal distributions with 0.1 standard deviation. Concordantly, the training process decreases the loss-function using a global-step weight decay based on the stochastic Adam optimizer (Kingma and Ba, 2014) modulating the weight changes from the FC layer, and starting with an initial learning rate of 0.00001.

$$\theta_t \leftarrow \theta_{t-1} - \alpha_t \frac{\hat{m}_t}{\sqrt{\upsilon_t} + \varepsilon} \tag{5.4}$$

Equation 5.4 and 5.5 show the Adam's update rule for the parameters, the kernel and biases weights θ_t, and the learning rate α_t respectively. $\hat{m}_t$ and υ_t are the bias-corrected estimators calculated from the parameters distribution and bounded gradients derived from the previous epoch learnt parameters, and the input feature-set.

$$\alpha_t = \alpha_{t-1} \frac{\sqrt{1 - \beta_2^t}}{1 - \beta_1^t} \tag{5.5}$$

β_1 and β_2 are the weight decay hyper-parameters calculated from the bias-correction either which are used to update the learning rates across the Adam Optimizer procedure (Goroshin et al., 2015). The training for all our evaluations in the three participant samples, and the feature importance evaluation was executed using 4 mini-batches randomly distributed across the epochs/training iteration. thus constituting the leave-one-trial-out (LOTO) per subject cross-validation modality iterating the batches on 47 out 48 different training trials per trial. A drop-out constant of 0.25 is applied per each training epoch. We will describe the LOTO cross-validation in the next subsection.

All the units in the conv-pool blocks use Rectified Linear Unit (ReLU) activation functions (Martinez, Bengio, and Yannakakis, 2013). A maximum of 500 iterations were set to train each iteration in the cross-validation modality with an early stop criterion described in (Schirrmeister et al., 2017).

5.2.4 Leave-One-Trial-Out per subject (LOTO) Cross-validation

To guarantee a correct execution of LOTO per subject cross-validation we created a .csv file including the 47 training normalized images for the training-set, and a test-set .csv file including the test normalized image to evaluate the Deep ConvNet classifier. As expected 47 trials for train and 1 for test per subject cross-validation.

The learning process read the training-set and test-set .csv files and used the features from the ZCA whitening normalization when the training process ends the process generates an output file grouping the results for each iteration for each subject. The results are composed of the trial number, the loss value, and the four test probabilities assigning 0 or 1 accuracy if the maximum probability computed by the classifier match the real trial class (*hit*) or not (*miss*).

All these values are computed per cross-validation epoch following Equation 5.6, 5.7, and 5.8 are used to compute the performance metrics such as the Accuracy, Precision, and Recall values finally obtained averaging the true-positives, true-negatives, type I (false positive), and type II (false negatives) errors per subject. We denote t_p as true-positives, t_n as true-negatives, f_p as false-positives, and f_n false-negatives epochs for subject j_{th}.

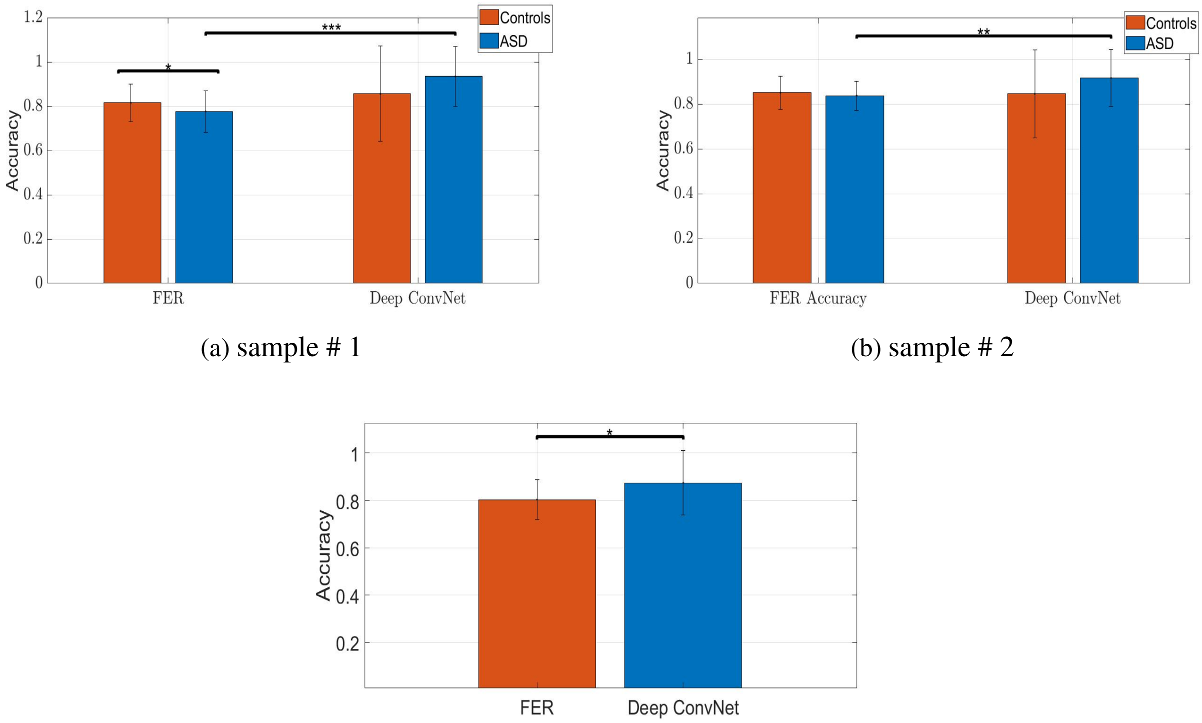
$$Accuracy_j = \frac{\sum_n t_p + \sum_k t_n}{\sum_n t_p + \sum_k t_n + \sum_p f_p + \sum_q f_n} \tag{5.6}$$

$$Precision_j = \frac{\sum_n t_p}{\sum_n t_p + \sum_p f_p} \tag{5.7}$$

$$Recall_j = \frac{\sum_n t_p}{\sum_n t_p + \sum_p f_n} \tag{5.8}$$

5.3 Performances Evaluation

The main results observed in this performance subsection is an overall higher accuracy observed in the ASD groups in comparison with the non-ASD/TD groups for all the three participant samples using our proposed Deep ConvNet performance, especially comparing FER and Deep ConvNet metrics. Figure 5.2a shows the barplots for FER average accuracies in red, and Deep ConvNet average accuracies in blue for the sample # 1 including 88 participans, and respectively Figure 5.2b for adults sample or sample # 2 with 69 participants, and Figure 5.2c the ASD only sample or sample # 3 with 36 participants. Table 5.1 shows the demographic details of these three participant samples.



(a) sample # 1

(b) sample # 2

(c) sample # 3

Fig. 5.2 Barplots showing the mean and the standard deviation for the TD group in red, and ASD group in blue. The black line marks the significant differences found between the accuracy groups denoted as FER or for Deep ConvNet classifier modalities. The number of asterisks are the number of zeros after the comma of the p-value comparing the groups using one-way ANOVA.

All the multivariable comparison used for the analysis below are one-way ANOVA. We found significant differences in the sample # 1 ,F(1,79)=21.54, p=0.0000213 between ASD FER accuracies and the ASD Deep ConvNet accuracies.

Comparing TD and ASD FER accuracies we found significant differences F(1,87)=4.69, p=0.0342 for sample # 1. However, for the sample # 2 with adult participants we did not find statistical differences between TD and ASD accuracies, F(1,68)=1.31,p=0.1053, implying an increased variance on the TD FER and Deep ConvNet or machine accuracies. For same sample # 2 the difference between ASD FER and ASD Deep ConvNet accuracies is significant, F(1,52)=8.15, p=0.0062, supporting a Deep ConvNet model extension in terms of performance for an older ASD sample thus complementing the non-behavioral difference between adults groups. Evaluating the sample # 3 -a similar age range sample as sample # 1- FER and Deep ConvNet accuracies we found a significant difference but with an increased variance across both groups, F(1,68)=6.8,p=0.0112.

We evaluate the comparisons between TD FER accuracies and the corresponding Deep

ConvNet accuracies and we did not find any difference on sample # 1, F(1,95)=1.131, p=0.1256, and nor for sample # 2 F(1,83)=0.027, p=0.878.

On the other hand, we evaluate the difference between TD and ASD FER accuracies across the three samples, thus finding significant differences only for the sample # 1, F(1,87)=4.69, p=0.0342, being TD>ASD, and non-significant differences for sample # 2, F(1,68)=2.71, p=0.1031 being TD>ASD. Sample # 3 we only have an ASD group and we can not evaluate differences between groups in this case.

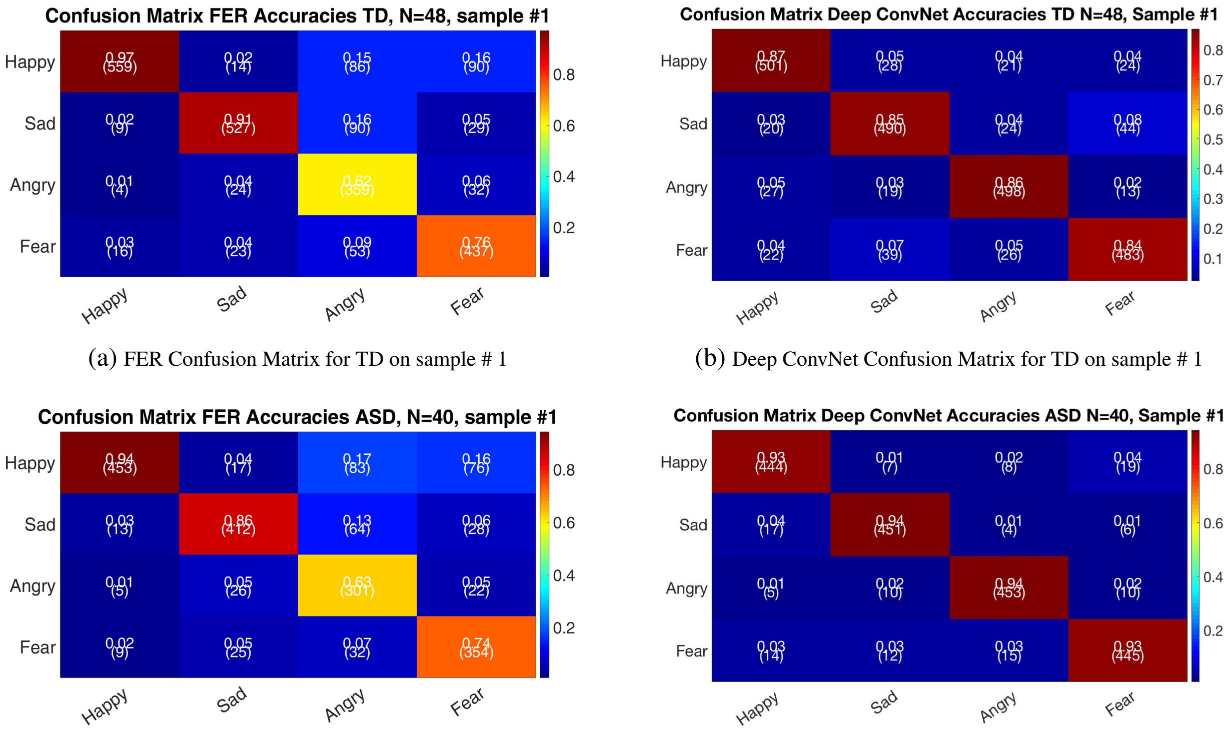
Evaluating the Deep ConvNet accuracies side between TD and ASD groups on sample # 1 and # 2 we did not find any significant difference due to the TD performance variance. For sample # 1 we obtained, F(1,95)=1.46, p=0.2281, obtaining TD<ASD, and for sample # 2, F(1,68)=2.71, p=0.1039, being TD< ASD. All these p-values are uncorrected due to they are comparisons between groups adding the participant factor assuming a full ANOVA model comparison.

5.3.1 Confusion Matrices - Performances

The accuracy metrics are extracted from each participant confusion matrix after the LOTO per subject modality for the Deep ConvNet evaluation as well as measuring the FER performance contructing a confusion matrix per participant as we mentioned above. In this subsection we will report the grouped confusion matrices for FER and Deep ConvNet performances adding the members of each confusion matrix calculated per subject.

The power of these results is shown in these FER and Deep ConvNet confusion matrices. Figures 5.3, 5.4, and 5.5 show all the confusion matrices patterns for TD/ASD groups and the three samples. Figures 5.3a, 5.3c, 5.4a, 5.4c, and 5.5a show an interesting pattern in the FER confusion matrices for all the samples. The main metrics' decreasing contribution is observed in the negative emotions such as **angry**, and **fear** showing an accuracy dropping of more than 20% in comparison with the Deep ConvNet accuracies across all the samples, and for both TD and ASD groups.

This decreasing effect is not observed in emotions such as **happy**, and **sad**. For these particular emotions the accuracy is the same or even higher than the Deep ConvNet accuracies. This suggests a robust generalization of negative emotions such as **angry**, and **fear** appraisal deficits using the Deep ConvNet on TD and ASD participants.



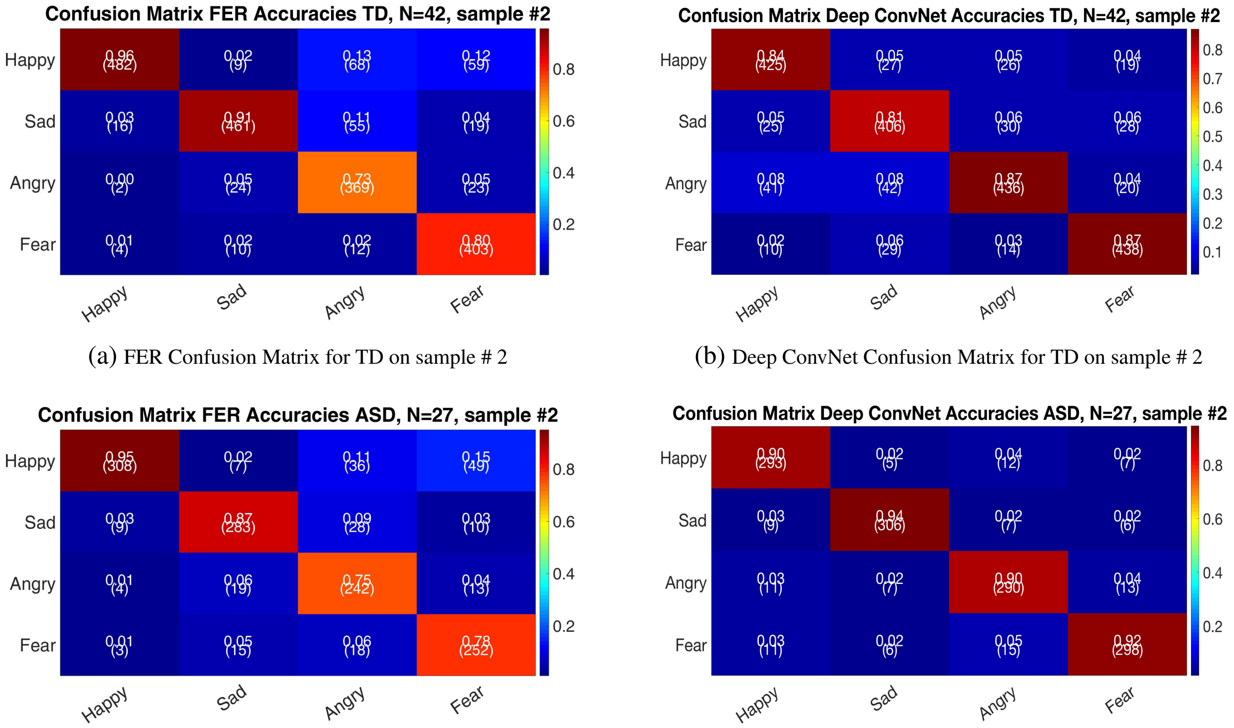
(a) FER Confusion Matrix for TD on sample # 1

(b) Deep ConvNet Confusion Matrix for TD on sample # 1

(c) FER Confusion Matrix for ASD on sample # 1

(d) Deep ConvNet Confusion Matrix for ASD on sample # 1

Fig. 5.3 Confusion matrices for the sample # 1, and for both groups TD and ASD. The matrices are calculated grouping each individual confusion matrix per subject. The colormap is jet and the colorbar show the performance between 0 and 1 going from blue being the lowest, and darker red the highest. The differences are critical for angry and fear emotions contributing to the accuracy dropping for FER metrics.



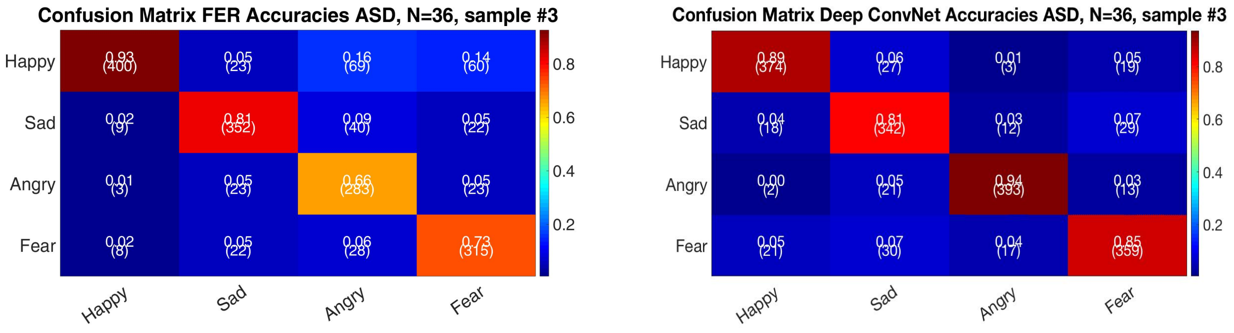
(a) FER Confusion Matrix for TD on sample # 2

(b) Deep ConvNet Confusion Matrix for TD on sample # 2

(c) FER Confusion Matrix for ASD on sample # 2

(d) Deep ConvNet Confusion Matrix for ASD on sample # 2

Fig. 5.4 Confusion matrices for the sample # 2, and for both groups TD and ASD. The matrices are calculated grouping each individual confusion matrix per subject. The color-map is jet and the colorbar show the performance between 0 and 1 going from blue being the lowest, and darker red the highest. The differences are critical for angry and fear emotions contributing to the accuracy dropping for FER metrics.



(a) FER Confusion Matrix for ASD on sample # 3

(b) Deep ConvNet Confusion Matrix for ASD on sample # 3

Fig. 5.5 Confusion matrices for the sample # 3, and for both groups TD and ASD. The matrices are calculated grouping each individual confusion matrix per subject. The color-map is jet and the colorbar show the performance between 0 and 1 going from blue being the lowest, and darker red the highest. The differences are critical for angry and fear emotions contributing to the accuracy dropping for FER metrics.

The TD participants perform better in terms of negative emotions such as **angry**, and **fear** FER accuracies, however, this deficit is observed in both groups showing a classifier generalization independently from the group as we can see in the Deep ConvNet confusion matrices. We can confirm this effect in Figures 5.3b, 5.3d, 5.4b, 5.4d, and 5.5b. All these resulting confusion matrices are obtained grouping each confusion matrix per subject, and per group individualizing the hit and misses in the corresponding emotion class.

5.3.2 Table Performances

Tables 5.2 and 5.3 show the overall performance metrics for FER and Deep ConvNet classifier respectively. In these tables Accuracy (Acc), Precision (Pre), Recall (Re), and F1 score (F1) are reported and these metrics are calculated following the Equations 5-7 for each participant confusion matrix. Evaluating statistically using ANOVA we found always a significant difference between the ASD Deep ConvNet metrics, and the ASD FER metrices, for all cases F(1,>=35) 4.21, p<0.05. For TD we do not found any significant difference due to the high variance in the performance metrics.

Table 5.2 Average and standard deviation of the overall FER task performances metrics for the all the samples are shown in this table. The results are computed averaging the Accuracy (Acc), Precision (Pre), Recall (Re), and F1 score (F1) from all the confusion matrices constructed per subject.

	TD				**ASD**			
Samples / Groups	**Acc**	**Pre**	**Re**	**F1**	**Acc**	**Pre**	**Re**	**F1**
Sample #1	0.815±0.083	0.808±0.079	0.802±0.077	0.807±0.079	0.776±0.093	0.774±0.089	0.768±0.088	0.771±0.088
Sample #2	0.846±0.074	0.858±0.067	0.847±0.073	0.852±0.070	0.837±0.064	0.853±0.062	0.840±0.062	0.846±0.062
Sample #3	–	–	–	–	0.817±0.077	0.8363±0.070	0.818±0.074	0.827±0.072

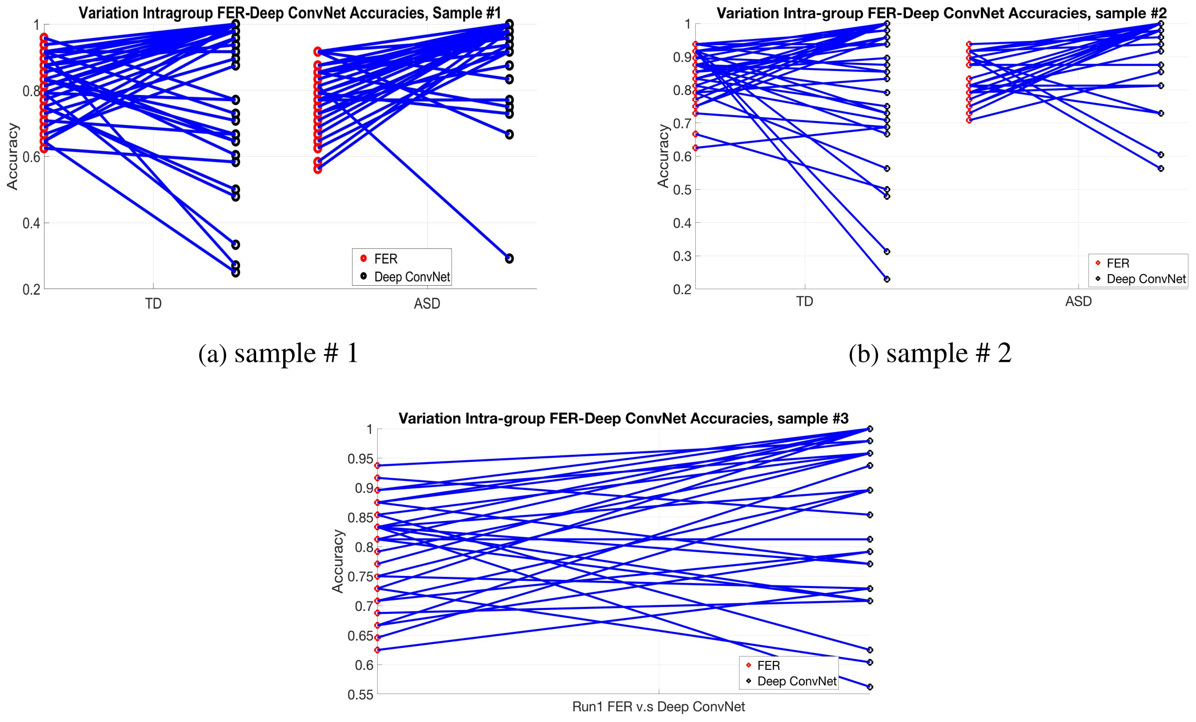
Table 5.3 Average and standard deviation of the overall Deep ConvNet performances metrics for the all the samples are shown in this table. The results are computed averaging the Accuracy (Acc), Precision (Pre), Recall (Re), and F1 score (F1) from all the confusion matrices constructed per subject.

	TD				**ASD**			
Samples/Groups	**Acc**	**Pre**	**Re**	**F1**	**Acc**	**Pre**	**Re**	**F1**
Sample #1	0.860±0.213	0.864±0.201	0.860±0.204	0.862±0.202	0.934±0.134	0.935±0.132	0.933±0.134	0.934±0.132
Sample #2	0.847±0.198	0.856±0.191	0.848±0.197	0.852±0.194	0.915±0.127	0.9207±0.121	0.915±0.127	0.918±0.124
Sample #3	–	–	–	–	0.909±0.114	0.911±0.113	0.907±0.116	0.909±0.114

5.3.3 Performance Interaction - FER (Human) v.s Deep ConvNet (Machine)

In this section we also describe the numerical interaction between the FER and Deep ConvNet accuracies showing effect plots per subject and intra-group for each sample. To relate the

performance variables we plot a blue line to join the FER accuracy point in red with the corresponding Deep ConvNet accuracy point in black.
Figure 5.6a, 5.6b, and 5.6c show these effect plots also denoted as *Spaghetti Plots* (Potter et al., 2009) or an important tool for visualization of statistical ensembles. These plots show an expected TD high-variance, however, most of the lines goes up linking FER with Deep ConvNet accuracies and supporting in most cases the Deep ConvNet as a classifier which can successfully decode the emotions having a lower FER accuracy or not indistinguishably. This effect is a constant across the samples, except in sample #3 where we can see a more sparse effect between the variables. For sample #3 we only compute the Run_1 denoted by the first phase of the experiment applied to these participants.



(a) sample # 1

(b) sample # 2

(c) sample # 3

Fig. 5.6 Intra-subject effect plots linking the FER and the Deep ConvNet average accuracies computed per subject for all the samples. Sample #3 only has an ASD group and we only report a single effect plot this group. The lines going up are observed in both TD and ASD groups yielding a higher variance in TD groups.

The results shown in Figure 5.6 support the model isolation between the statistical model which produce the behavioral performance in ASD participants and the Deep ConvNet model which can extract important information for decode negative emotions properly, and overcome the deficits shown in the confusion matrices of Figures 5.3, 5.4, and 5.5. In the following chapter we will show the statistical interactions between the FER (Human) and Deep ConvNet (Machine)accuracy performances, and the ASD severity measure such as the ADOS-CS, and others such as the AQ and the SCQ scores related to social competence measures.

Chapter 6

Correlation between Deep ConvNet parameters and ADOS-CS

In this chapter we will analyze statistical correlation between Performance variables described in the chapter above such as FER and classifier accuracies, and behavioral outcome measures such as ADOS-CSS or the ADOS calibrated severity score (Gotham, Pickles, and Lord, 2012), the score of the Autism-Spectrum Quotient (AQ) (Baron-Cohen et al., 2001), and Social Communication Questionnaire (SCQ) (Bölte, Holtmann, and Poustka, 2008).
We found an overall **non-significant correlation between the FER (human) and Deep ConvNet (machine) performances, and the Autism severity scores such ADOS-CS, SCQ, and AQ** for Adults specifically.
Following the same effect plots (*Spaghetti Plots*) analysis used between the FER and Deep ConvNet accuracies in the chapter above we paired the FER and Deep ConvNet accuracies with the ADOS-CS calculated for each sample in the screening time. First, we will describe a typical a screening process using ADOS-2 and the calculating of the ADOS-CS used for the three participant samples included in this dissertation.

6.1 ADOS-CS evaluation

In Social Competence and Treatment Lab (SCTL), StonyBrook, NY, USA the screening process is very important before implement any neural measurement, or behavioral intervention visit to a new participant sample as a clinical study. The PhD fellows and research associates are formally trained to fill up . In (Gotham, Pickles, and Lord, 2009, 2012) ADOS-2 assessment has been evaluated with more than 2000 participants across multiple races, and ages providing a large statistical power. As we reported on Table 2.1 and chapter

2 there is multiple methodologies for ASD diagnosis and screening such ADI-R, SSR, and CELF-4.

Particularly, ADOS as a semi-structured assessment has an important statistical validity (Gotham et al., 2007). Each item from the language and developmental level measurement is divided in a 4-point scale with 0 meaning *no abnormality observed*, and 3 *moderate or severe abnormality observed*. For each ADOS module the final score is obtained adding the values of each item per module by the trained evaluator.

With purpose of reduce demographics variance and increase the sample representative level Autism researchers calibrate the ADOS-2 raw score using eighteen specific age/language cells mapping the raw ADOS score on each cell. Following the Table 6.1 and using the raw ADOS score per module we can infer the corresponding ADOS-CS assigned per participant. All the ADOS-CS values used in the subsections below are discrete integers between 1-10 meaning a larger severity with a larger number and viceversa.

Table 6.1 ADOS-CS calculation extracted from (Gotham, Pickles, and Lord, 2009) and sumarize the calibration algorithm for normalize the ADOS-raw score using age, and language levels obtained from the ADOS-2 itself. The values inside the table are the ADOS raw scores, the values listed on the column one is the resulting ADOS-CS. For this table NS is No Spectrum or TD, and AUT Autism participants.

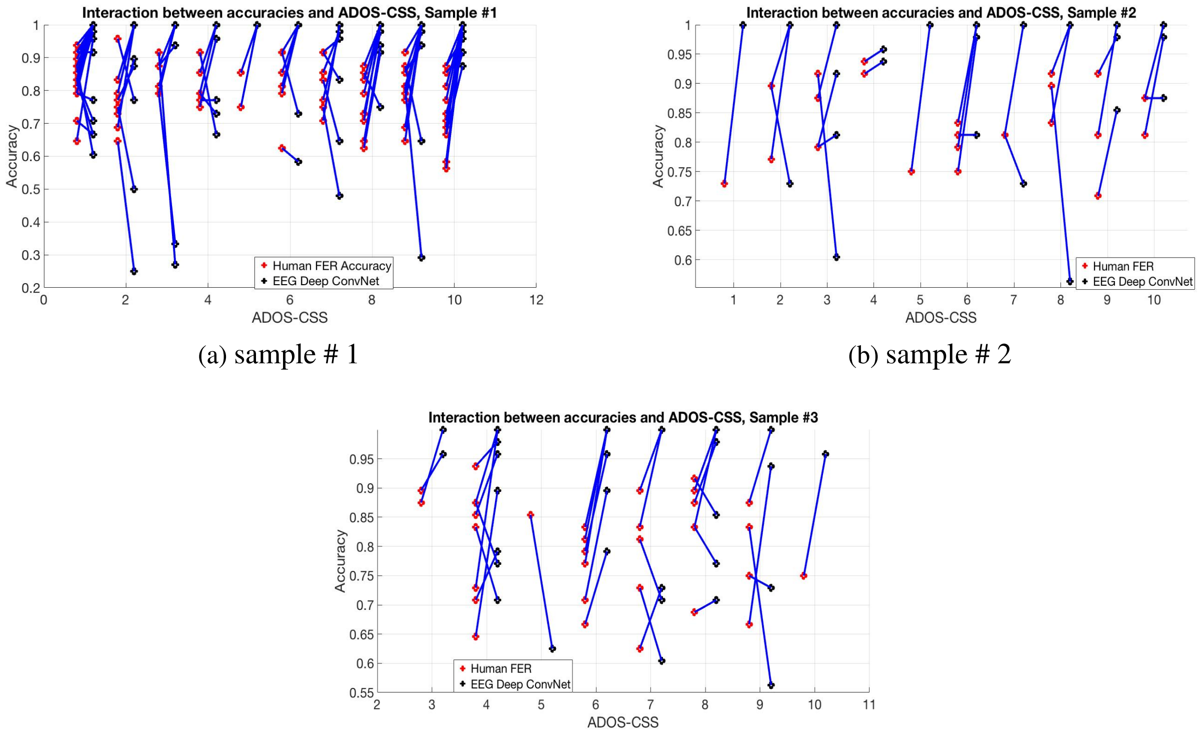
		ADOS Raw Score																	
		Module 1, No Words				**Module 1, Single Word**					**Module 2, Phrases**						**Module 3, Fluent**		
ADOS	**CSS**	**2y**	**3y**	**4-5y**	**14y**	**2y**	**3y**	**4y**	**5-6y**	**14y**	**2y**	**3y**	**4y**	**5-6y**	**7-8y**	**16y**	**2-5y**	**6-9y**	**16y**
NS	**1**	0–6	0–6	0–3	0–3	0–3	0–4	0–2	0–2	0–2	0–2	0–3	0–3	0–3	0–2	0–2	0–3	0–2	0–3
	2	7–8	7–8	4–6	4–6	4–5	5–6	3–4	3–4	3–5	3–5	4–5	4–5	4–5	3–5	3–5	4	3–4	4
	3	9–10	9–10	7–10	7–10	6–7	7	5–7	5–7	6–7	6	6	6	6–7	6–7	6–7	5–6	5–6	5–6
ASD	**4**	11–13	11–14	11–12	11–13	8–10	8–9	8–9	8–10	8–9	7–8	7–8	7	8	8	8	7	7	7
	5	14–15	15	13–15	14–15	11	10–11	11	11	10–11	9	9	8–9	-	-	-	8	8	8
AUT	**6**	16–19	16–20	16–19	16–19	12–13	12–14	12–15	12–16	12–18	10–11	10–12	10–13	9–14	9–14	9–14	9–11	9–10	9–10
	7	20–21	21–22	20–21	20–22	14–16	15–17	16–18	17–19	19–20	12	13–14	14–16	15–16	15–17	15–17	12	11–12	11–12
	8	22	23	22–23	23–24	17–19	18–19	19–20	20–21	21	13–14	15–16	17–18	17–20	18–21	18–20	13–15	13–14	13–14
	9	23–24	24	24–25	25	20–21	20–21	21–22	22–23	22–23	15–17	17–18	19–20	21–22	22–23	21–23	16–17	15–17	15–17
	10	25–28	25–28	26–28	26–28	22–28	22–28	23–28	24–28	24–28	18–28	19–28	21–28	23–28	24–28	24–28	18–28	18–28	18–28

Autism and ASD are considered ASD and Pervasive Developmental Disorder. Not Otherwise Specified (PDD-NOS) participants groups respectively for (Gotham, Pickles, and Lord, 2009, 2012). For simplification purposes we only use TD and ASD groups grouping all of the ADOS-CS values for the statistical correlation analysis explained below.

6.2 Interaction Between Deep ConvNet and FER Accuracies with ADOS-CS

The ADOS-CS spectrum reflects the level of Autism severity based on language, cognitive competence impairments, and age. The ADOS-CS can be calculated for TD and ASD based on the Table 6.1. For our particular case sample #1 has ADOS-CS values for TD and ASD groups. However, sample #2 and sample #3 have ADOS-CS scores reported for ASD groups only. This limitation is given by sample #2 include adults TD participants (older than 18 years) and the ADOS-CS can not be extended for a NS adult participants. As for sample #3 it has only an ASD group.

Figures 6.1a, 6.1b, and 6.1c show the interaction effect across the participants that belong to the 10 different ADOS-CS values between 1-10. We organize the effects grouping the participant's performances from 1 to 10 ADOS-CS values and similar to the effect plots in Figure 5.6 red is FER accuracy, and black is Deep ConvNet accuracy per participant. In these reported Figures TD and ASD performances are grouped all together to find a correlation per sample, as we will analyze in the next subsection.



(a) sample # 1

(b) sample # 2

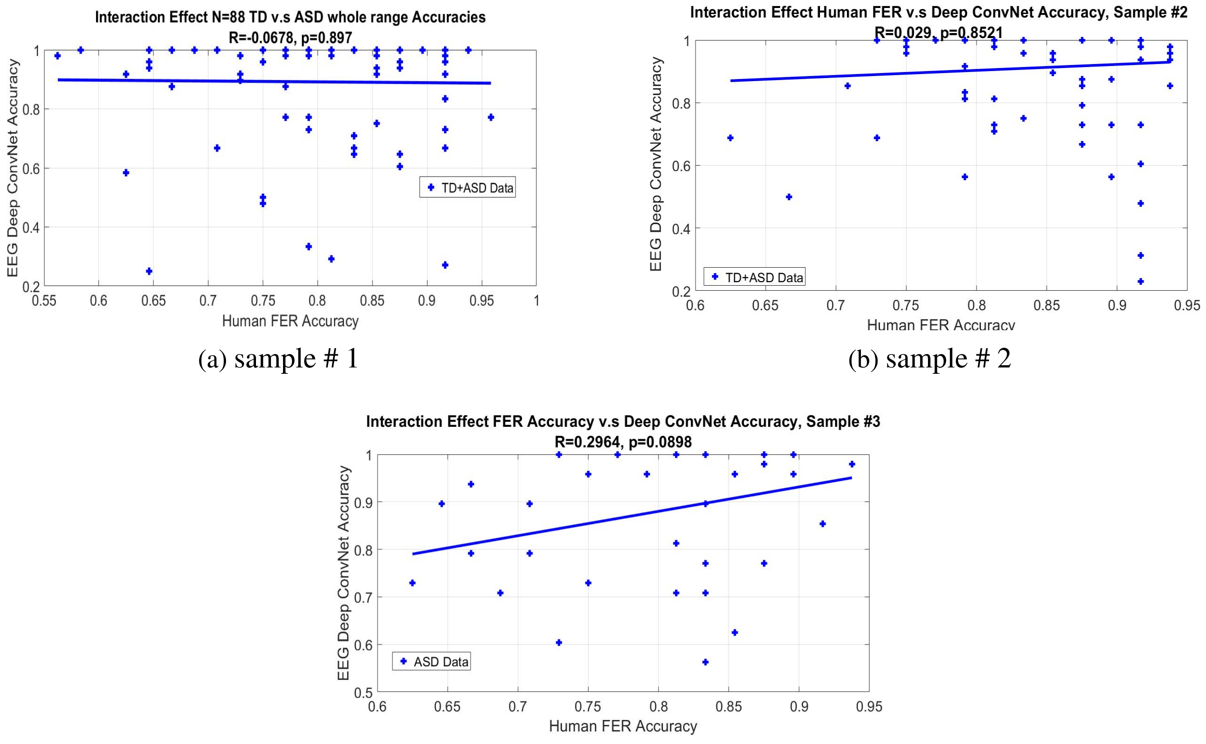
(c) sample # 3

Fig. 6.1 Interaction effect between FER and Deep ConvNet accuracies, and the ADOS-CS spectrum explicit on the x-axis. The variation across ADOS-CS scores show high and low Deep ConvNet accuracies indistinguishably, as well as high and low FER accuracies without finding any negative or positive correlation across the variables.

High and low FER and Deep ConvNet accuracies are presented in both low and high ADOS-CS scores showing a considerable level of sparsity across the ADOS-CS spectrum. For sample #1 including TD participants we can observe a large variance thus suggesting FER, Deep ConvNet accuracies, and the ADOS-CS scores as three different models. However, we expect a negative correlation between FER and ADOS-CS for one of the samples, or at least ASD groups. We will support this hypothesis in the next subsection evaluating the statistical correlation between FER, Deep ConvNet performances and ADOS-CS.

6.2.1 FER v.s Deep ConvNet

To visualize the effects of the *Spaghetti Plots* shown above we evaluate the statistical correlations using a Generalized Linear Model (GLM) explained in the Appendix A. The results obtained using the GLM analysis are shown in Figures grouping TD and ASD groups within the corresponding FER and Deep ConvNet accuracies for the three samples respectively. We can observe none significant regression, or a positive/negative significant R-Pearson value to be considered significant in this analysis.
Figures 6.2a, 6.2b, and 6.2c show the linear regressions across TD and ASD groups, and FER and Deep ConvNet accuracies across for the three samples in this study respectively.



(a) sample # 1

(b) sample # 2

(c) sample # 3

Fig. 6.2 FER v.s Deep ConvNet accuracies linear regressions for sample #1 (Figure 6.2a), #2 , and #3 . Dot points represent the pair (FER x-axis, Deep ConvNet y-axis), and the line is a robust linear regression calculated using *fitlm* from Matlab package, and the model explained in Appendix A.

For sample #3 we found a little positive correlation but not significant. These results can be justified because sample #3 was recruited in a specific location in Virginia, US, in early 2000s and with a very different experimental setting than sample #1. For this particular case we can suggest that neural activity, and FER human performances are more in synchrony in comparison with other samples.

These results suggest that FER and Deep ConvNet accuracies represent two different numerical models, thus showing that is possible to decode successfully the emotion using our Deep ConvNet-based pipeline and assuming a complete different numerical representation from the neural activity information decoding comparing the Deep ConvNet performances with the FER human behavior performances across TD and ASD groups.

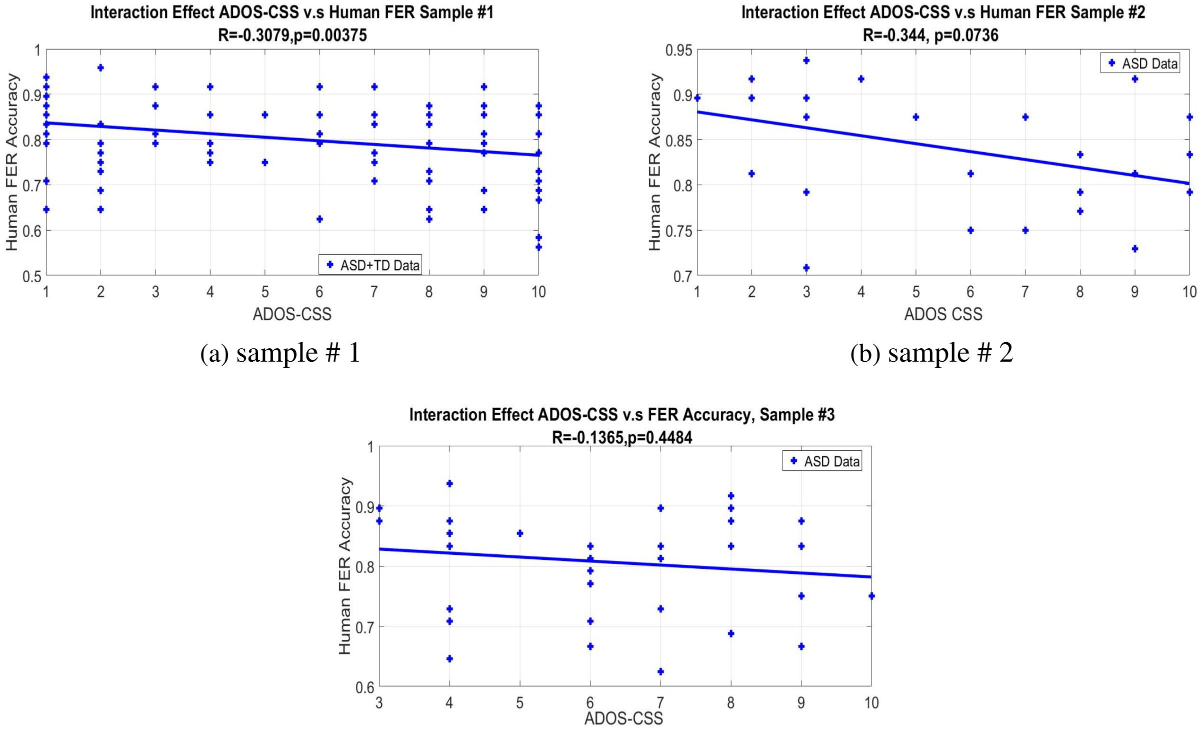
All the intercepts were significantly estimated except for sample #2 ADOS-CS v.s Deep ConvNet evaluation. As expected β_1 was always negative for the negative correlation value observed between the FER accuracy and the ADOS-CS. The β values explained in Appendix A are reported in Tables 6.5, 6.6, and 6.7 for sample #1, #2, and #3, specifically in the intersection between row one and column one.

6.2.2 FER v.s ADOS-CS

To evaluate the correlation between performances and ADOS-CS variables we use the R-Pearson values defining the positivity or negativity of the correlation, and the p-value obtained pairing all the possible scores per subject between FER and Deep ConvNet accuracies, and the ADOS-CS as we explained below.

Our golden-standard for this analysis is to find at least a negative correlation between FER accuracies and ADOS-CS, observed in previous studies (Clarkson et al., 2019). Any significant correlation is a plus for analysis expecting always negative R-Pearson comparing performances with the severity scores.

Figure 6.3a, 6.3b, and 6.3c show the linear regressions measuring the relationship between FER accuracies and ADOS-CS for all the samples included in this study. As expected all correlations between FER performances and ADOS-CS are negative, thus finding significant correlations for sample #1, and nearly significant for sample #2. In Tables 6.2 and 6.3 we found the R-Pearson and p-values for the regressions between ADOS-CS and FER performances for sample #1 and #2. For #3 the correlation is negative but not negative enough for being significant showing a more sparse FER performances in comparison with sample #1 and #2. The R-Pearson and p-values are reported in Table 6.4 for sample #3.



(a) sample # 1

(b) sample # 2

(c) sample # 3

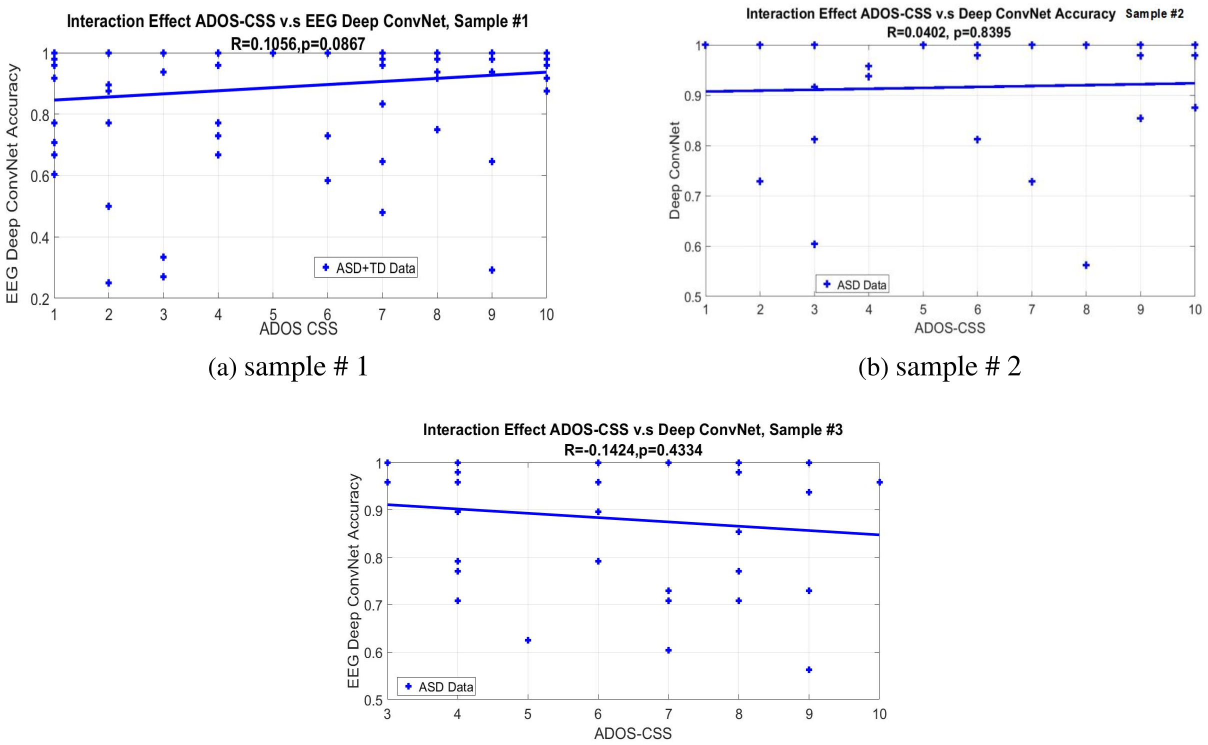
Fig. 6.3 ADOS-CS scores v.s FER accuracies linear regressions for sample #1 (Figure 6.3a) , #2 (Figure 6.3b) , and #3 (Figure 6.3c) . Dot points represent the pair (ADOS-CS x-axis, FER y-axis), and the line is a robust linear regression calculated using *fitlm* from Matlab, and the model explained in Appendix A.

We suggest that the significant p-value found for sample #1 evaluating statistical correlation between ADOS-CS and FER accuracies are related to the number of points or subjects (N=88) being covered by the ADOS age cells. Sample #2 and #3 only include ASD participants eliminating the β_2 and β_3 factors from the corresponding regressions. These specific regressions are considered with less statistical power and robustness in comparison with Sample #1. The β factors are reported in Tables 6.5, 6.6, and 6.7 for sample #1, #2, and #3, specifically in the intersection between row three and column one.

6.2.3 Deep ConvNet v.s ADOS-CS

The Deep ConvNet accuracies depends on the neural activity but we expect differences between ADOS-CS and Deep ConvNet accuracies due to the multiple conv-pool blocks dedicated to discriminate intermediate features in our emotion decoding pipeline. The Deep ConvNet model will be considered an isolated model from the ASD behavioral perfomance, deficits, and Autism severity measures.

Figures 6.4a, 6.4b, and 6.4c show the linear regressions between ADOS-CS and Deep ConvNet accuracies. In this linear models we did not find any significant or considerable correlation between both variables across the three samples. Tables 6.2, 6.3, and 6.4 show the R and the p-values for all the samples in this study as we mentioned above. Sample #3 shows a non-significant negative correlation similar to the FER case due to the different experiment setting in comparison with Sample #1 as we mentioned above.



(a) sample # 1

(b) sample # 2

(c) sample # 3

Fig. 6.4 ADOS-CS scores v.s Deep ConvNet accuracies linear regressions for sample #1 (Figure 6.4a) , #2 (Figure 6.4b) , and #3 (Figure 6.4c) . Dot points represent the pair (ADOS-CS x-axis, Deep ConvNet y-axis), and the line is a robust linear regression calculated using *fitlm* package from Matlab, and the model explained in Appendix A.

Tables 6.5, 6.6, and 6.7 shows the such as β_1, β_2, and β_3 and the intercept b with the corresponding p-values in the intersection between row two and column one. As we mentioned above the for sample #2 and #3 with ASD group only, the linear model is reduced using β_1 and b only.

Table 6.2 Linear regression R-Pearson correlation value and the corresponding p-value between the variables in row and columns for sample #1. A positive R value represent a positive slope, and negative R value represent a negative slope in the linear regression. These values show the statistical relationship between human and machine accuracies. Only the highlighted values are significant correlations. The partially highlighted values are near to be significant in the evaluation.

Sample #1	**FER**		**Deep ConvNet**		**ADOS-CS**	
	R	p	R	p	R	p
FER	–	–	-0.0678	0.8972	-0.3079	0.003775
Deep ConvNet	-0.0678	0.8972	–	–	0.1056	0.0867
ADOS-CS	-0.3079	0.003775	0.1056	0.0867	–	–

Table 6.3 Linear regression R-Pearson correlation values and the corresponding p-value between the variables in row and columns for sample #2. A positive R value represent a positive slope, and negative R value represent a negative slope in the linear regression. These values show the statistical relationship between human and machine accuracies. Only the highlighted values are significant correlations. The partially highlighted values are near to be significant in the evaluation.

Sample #2	**FER**		**Deep ConvNet**		**ADOS-CS**	
	R	p	R	p	R	p
FER	–	–	0.0290	0.8521	-0.3472	0.0765
Deep ConvNet	0.0290	0.8521	–	–	0.0402	0.8395
ADOS-CS	-0.3472	0.0765	0.0402	0.8395	–	–

Table 6.4 Linear regression R-Pearson correlation value and the corresponding p-value between the variables in row and columns for sample #3. A positive R value represent a positive slope, and negative R value represent a negative slope in the linear regression. These values show the statistical relationship between human and machine accuracies. Only the highlighted values are significant correlations. The partially highlighted values are near to be significant in the evaluation.

Sample #3	**FER**		**Deep ConvNet**		**ADOS-CS**	
	R	p	R	p	R	p
FER	–	–	0.2964	0.0898	-0.1364	0.4484
Deep ConvNet	0.2964	0.0898	–	–	-0.1464	0.4334
ADOS-CS	-0.1364	0.4484	-0.1464	0.4334	–	–

Table 6.5 Estimated parameters for all the linear regressions in this Chapter comparing correlations between Deep ConvNet, FER human accuracies, and the ADOS-CS scores for sample #1.

Sample#1	**Deep ConvNet**							
	β_1	p	β_2	p	β_3	p	b	p
FER	0.4783	0.1322	0.7246	0.0444	-0.8082	0.07184	0.8567	0.0759
ADOS-CS	0.0028	0.7323	0.3087	0.2375	-0.3143	0.3281	0.8464	6.35E-35
Sample#1	**FER**							
	β_1	p	β_2	p	β_3	p	b	p
ADOS-CS	-0.0077	0.0788	–	–	–	–	0.8813	4.72E-22

Table 6.6 Estimated parameters for all the linear comparing correlations between Deep ConvNet, FER human accuracies, and the ADOS-CS scores for sample #2.

Sample#2	**Deep ConvNet**							
	β_1	p	β_2	p	β_3	p	b	p
FER	0.1841	0.6217	0.5646	0.3084	-0.5873	0.3662	0.6817	0.0335
ADOS-CS	0.0018	0.8393	–	–	–	–	0.9054	1.25E-14
Sample#2	**FER**							
	β_1	p	β_2	p	β_3	p	b	p
ADOS-CS	-0.0077	0.0788	–	–	–	–	0.8813	4.72E-22

Table 6.7 Estimated parameters for all the linear comparing correlations between Deep ConvNet, FER human accuracies, and the ADOS-CS scores for sample #3.

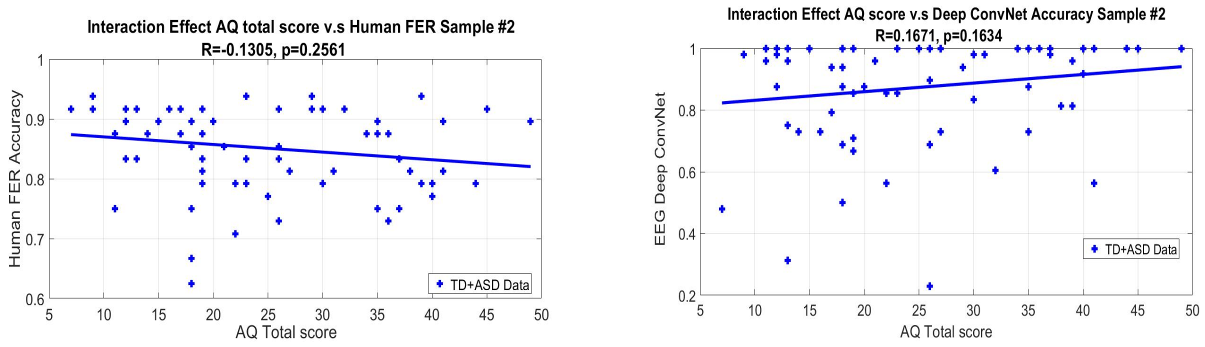
Sample#3	**Deep ConvNet**							
	β_1	p	β_2	p	β_3	p	b	p
FER	0.4814	0.0834	–	–	–	–	0.4869	0.0324
ADOS-CS	-0.0096	0.4214	–	–	–	–	0.9353	2.01E-13
Sample#3	**Deep ConvNet**							
	β_1	p	β_2	p	β_3	p	b	p
ADOS-CS	-0.0052	0.4765	–	–	–	–	0.8371	5,08E-15

6.2.4 Including AQ and SCQ scores

In this subsection in order to include TD participants for Sample #2 we include other severity measures such as the Autism-Spectrum Quotient (AQ) (Baron-Cohen et al., 2001), and for sample #3 the Social Communication Questionnaire (SCQ) (Bölte, Holtmann, and Poustka, 2008).

The AQ is a psychological self-assessment which was primarily evaluated to diagnose Autism-Spectrum (AS), and High-Functioning Asperger (HFA) groups with different economic and demographic conditions with a great validity. The AQ is composed of 50 items, 10 for each module evaluated such as *Communication*, *Social Skills*, *Imagination*, *Local Details*, and *Attention Switching* being this assessment only adequate for adult participants. The Total AQ

score is reported summing the scores for al the 50 items.
The score is calculated per item from a 4 point-scale. For a half of the items the *"Slightly Agree"* and *"Definetely Agree"* score 1, and for the other half *"Slightly Disagree"* and *"Definetely Disagree"* score 1 too. As this assessment is dedicated to evaluate behavioral outcome measures we repored the Total AQ score for Sample #2.
On the other hand, SCQ is a parent report questionnaire for screening and severity measures of ASD participants using 40 binary scaled items for SCQ screening open the door for TD and ASD participants having a last differentiation specificity of 0.96. An average SCQ score is reported for the Sample #3 regressions.
A higher AQ and a lower SCQ total are a representation of a higher ASD severity using a self and parent report options instead of a evaluator annotator as the ADOS-CS. In Figures 6.5a and 6.5b we show the linear regression of FER accuracies related to the AQ total scores, and the Deep ConvNet accuracies related to the same AQ scores from Sample #2 including only adults participants.
In Sample #2 AQ linear regressions we did not find any significant correlations in terms of the p-values, however, we found a consistent negative correlation comparing the FER performances and the AQ total scores, $R = -0.1305$,$p = 0.2561$, and an inconclusive little positive correlation between the Deep ConvNet accuracies, and the same AQ scores, $R = 0.1671$, $p = 0.1634$ thus supporting again the different model established by the Deep ConvNet which is learning features and patterns in a different way in comparison with the human brain in TD and ASD participants.



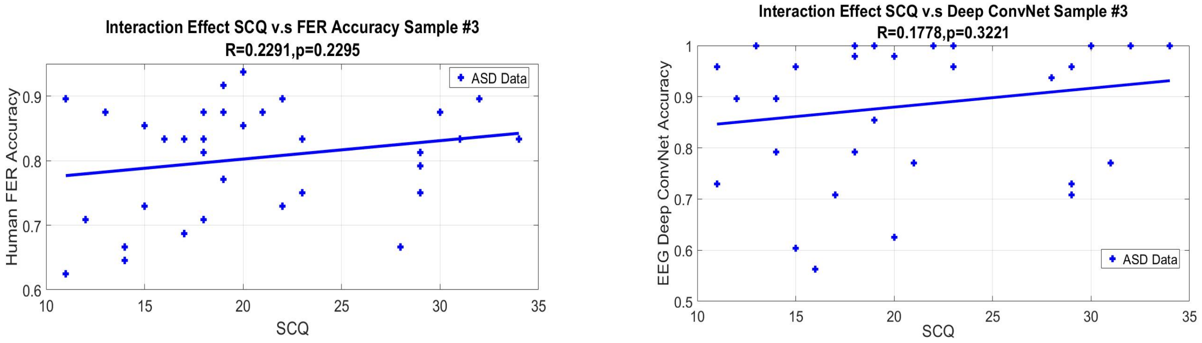
(a) AQ v.s FER accuracies

(b) AQ v.s Deep ConvNet Accuracies

Fig. 6.5 Linear regression between the FER accuracies v.s AQ total scores in Figure 6.5a, and between Deep ConvNet accuracies v.s AQ total scores in Figure 6.5b with all the data from Sample #2 TD and ASD data included.

Figures 6.6a and 6.6b show the linear regressions evaluating the relationship between the FER accuracies and SCQ average scores, as well as the relationship between the Deep

ConvNet accuracies and the same SCQ average scores respectively. The correlations are both positives for FER and Deep ConvNet accuracies but they are not significant and again they do not offer a conclusion about the statistical interaction of SCQ and the human, $(R = 0.2291, p = 0.2225)$, and machine, $(R = 0.1778, p = 0.3221)$, performances.



(a) SCQ v.s FER accuracies

(b) SCQ v.s Deep ConvNet Accuracies

Fig. 6.6 Linear regression between the FER accuracies v.s SCQ total scores in Figure 6.6a, and between Deep ConvNet accuracies v.s SCQ total scores in Figure 6.6b with all the data from Sample #3.

6.3 ConvNet and Behavioral Models, are linked?

The correlation results reported in previous two sections suggest that Deep ConvNet and the subsequent complete pipeline proposed in this dissertation encode successfully the neural activity representation to create and independent model through the higher performances from the Deep ConvNet.

This neural activity statistically associated with the emotion decoding is then completely isolated model comparing them with the machine accuracies. The FER human performances are correlated negatively as expected with the severity scores such as ADOS-CS, AQ and SCQ (Clarkson et al., 2019).

This supports our pipeline as a transparent system being able to generalize and compensate the neural and behavior deficits found in ASD emotion appraisal mentioned in the previous chapters.

The statistical results show negative Pearson correlation between FER accuracies and ADOS-CS in Figure 6.3 and Tables 6.2, 6.3, and 6.4. The only significant correlation found across all the variables and samples was between FER and ADOS-CS on sample #1.

Sample #3 shows some small negative and positive correlation comparing the Deep ConvNet accuracies with the ADOS-CS. But this correlation is not significant because of the data sparsity and the sample #3 heterogeneity. In the following chapter we will show the similarities between TD and ASD neural activity sensitivity, and the neural activity corresponding relevance with its corresponding location in space and time using the more robust saliency methods.

Chapter 7

Saliency Maps Evaluation - EEG Features Relevant Measures

In this chapter we will explore the brain network differences between TD and ASD groups (Black et al., 2017) visualizing the Deep ConvNet classifier sensitivity and the relevance maps obtained from the Deep ConvNet using layer-wise and on-pixel relevance propagation, inverse deconvolution operators, statistical linear modelling, and EEG features occluding methods.

For this preliminary study we use the 88 TD/ASD participants observing plausible/significant differences between groups on the relevance maps as we will explained below. This proposed saliency maps can be considered a novel set of methodologies for measuring feature importance going from EEG single trials 2D input-map to the classification output in a Deep ConvNet, and visceversa (Kapishnikov et al., 2019).

Our definition of relevance-map is a feature relevance quantification using the training parameters from the Deep ConvNet and the 2D input image, and for our specific case we measure similarities and differences as well in brain network activation from FER-elicited patterns in TD and ASD groups.

For this saliency-map analysis we evaluate four different type of saliency methods. First, we propose to investigate saliency maps which include parameter optimization, and linear/non-linear constraints applied to all the conv-pool blocks on a trained Deep ConvNet classifier, thus propagating feature-maps relevance quantification from the classifier decision to the input layer.

These methods are called the Layer-wise Relevance Propagation (LRP) (Bach et al., 2015; Binder et al., 2016) methodologies. These type of methods use constraint and inverse operator models to propagate the relevance from the output layer's decisions through the hidden conv-pool blocks to the input feature-map.

On other hand, we use other methods such as the dedicated Deconvolution process (Zeiler and Fergus, 2014; Zeiler, Taylor, and Fergus, 2011) using Deconvolution and Un-pooling inverse operators for each conv-pool blocks. We will also focus on the deconvolution based drawbacks and the evident need to add a double parameters per each conv-pool block to obtain a single relevance-map.

Third, we study gradient-based methods such as the broadly used such as Gradient-weighted Class Activation Map Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018), and other new evaluations such as Smooth-Grad (Smilkov et al., 2017) using the learnt parameters of the last fully-connected weights from a function Deep ConvNet in the particular case of our pipeline. However, we will focus on the gradient-based model drawbacks specially when the number of classes is high, and the input space representation is very entangled as the EEG single-trial.

A final type of saliency method is the systematic occluding of the input feature-set (Huang et al., 2018; Zeiler and Fergus, 2014), this method will particularly occlude the transformed EEG input 2D feature-set using square or rectangular occluding sections. The output performances are modulated depending on the size and the amount of occluders distributed across the input feature set.

In the following sections we will discuss the results for the Saliency Methods used for feature-relevance calculation in Sample #1.

7.1 Saliency Methods - Results

In this section we will specify the results for LRP, Deconvolution-based, Gradient-based, and Occluding based results. As we mentioned a select the sample #1 with 88 participants to test the saliency methods generalizing the relevance-maps for a children and adolescents sample who includes ADOS-CS values for TD and ASD groups, and thus allowing the possibility to associate relevance-maps with the corresponding ADOS-CS scores.

The relevance/saliency maps calculated from proposed by the methodologies mentioned above have different settings and different requirements for a correct relevance calculation, thus implying different interpretations between methods.

In the following subsections we will explain the main core of the saliency methods models, and the important parameters optimized for consolidating the saliency maps more robust.

7.2 iNNvestigate package

The iNNvestigate package (Alber et al., 2019) is a set of enhanced and modified saliency maps implemented by Google brain research group. In this dissertation we implement the complete library on our 3-conv-pool block network pipeline to compare an initial set of saliency maps such as Integrated Gradients, Deconvolution (mentioned above), PatternNet and Pattern-Attribution (Kindermans et al., 2017a,b), and LRP presets using stabilizers and the flat $\alpha\beta$ rule balancing positively and negatively relevance propagations (Montavon, Samek, and Müller, 2018). We will explain this in detail in the following subsections and Appendix B.

These iNNvesitage methods represent the most significant improvement and most reliable saliency maps options. iNNvestigate authors suggest to improve the salience maps reliabilities using the package as critical improvement critical for ML studies which include clinical trials (Hooker et al., 2018; Kindermans et al., 2017c).

7.2.1 Previous Saliency Methods

As we mentioned above popular saliency maps such as Grad-CAM and Grad-CAM++ have been used for generalize neural activity in clinical trials (Andreotti, Phan, and De Vos, 2018). However, in our particular experiment we can not propagate the gradient values from the last max-pool layer because our input 2D feature-set is a single-channel image with a size 752 time points $\times$ 30 channels, thus the gradient calculated from single-channel image on the fully-connected layer output is always zero cost function.

The existing Grad-CAM and Grad-CAM++ implementations use RGB, CMYK, and HSV squared images, or other multi-channels image representation propagating a gradient different than zero. This same situation is evaluated in (Andreotti, Phan, and De Vos, 2018) where the input representation is an input image composed of three channels: one channel EEG, a second channel EOG, and a third channel electro-myography (EMG) emulating a RGB representation. These input images are composed of a Frequency $\times$ Time input image composed of multiple biosignals EEG, EOG, and EMG channels.

In the following subsections we will discuss the most critical and robust saliency methods evaluated by the iNNvestigate package such LRP its foundations and the corresponding presets, PatternNet and Pattern-Attribution, and Smooth-Grad, with an extra evaluation using the RemOve-and-RetrAin (ROAR) (Hooker et al., 2018) method to debug the level of certainty of the resulting relevance-maps obtained by these established methods.

7.2.2 Layer-Wise Relevance Propagation (LRP)

In this subsection and Appendix B we will explain the evolution and multiple models who constraint and construct the LRP methodologies for relevance-map calculation on a Deep ConvNet classifier.

To define the LRP foundations (Bach et al., 2015) defines a constraint for a single and multiple hidden-layer ConvNet network based on the Lebesgue measure and the conservation law (Szepessy, 1989). Equation 7.1 shows the conservation law for a multiple-layer network quantifying the classifier output $f(x)$ to calculate the corresponding relevance-maps denoted as R_d^l where l is the l_{th} layer or conv-pool block and d the dimensionality unit defined as a pixel, a neuron, or a set of neurons. L is the total amount of layers in the Deep ConvNet.

$$f(x) = \sum_q R_q^1 = \sum_{d \in (l+1)} R_d^{l+1} = \sum_{d \in (l)} R_d^l = \ldots = \sum_d R_d^L \tag{7.1}$$

Following Equation 7.1 LRP is a generalized model for a discriminative multi-layer network using a *"message passing"* back-propagation methodology where the relevance calculated from the neuron j to the neuron i can be calculated as the sum of the input relevances $\sum_i R_{i \leftarrow j}^{l,l+1}$ from predecessor layer. Equation 7.2 described this propagation using the same notation of Equation 7.1.

$$\sum_i R_{i \leftarrow j}^{l,l+1} = R_j^{l+1} \tag{7.2}$$

To generalize more the network structure we can define the relevance $R_{i \leftarrow j}^{(l,l+1)}$ based on the layers and neuron's parameters as follows $z_{ij} = x_i^l \omega_{ij}^{(l,l+1)}$ where $\omega_{ij}^{(l,l+1)}$ are the learned weights connecting layer l and i. These weights are adjacent to the inputs for neuron j all related to the index i and the variable x_i^l. Following this new parameter definition we can rewrite Equation 7.3 assigning to $f(x) = \sum_i z_{ij}$ from the top layer to the bottom generalizing the message passing as Equation 7.3. For a sum across all the units per layer we add the bias term as b_j for the layer j_{th}

$$R_i^l = \sum_j \frac{z_{ij}}{\sum_i z_{i,j}} R_j^{l+1} = \sum_j \frac{x_i^{l+1} \omega_{ij}^{(l,l+1)}}{\sum_i x_i^l \omega_{ij}^{(l,l+1)} + b_j} R_j^{l+1} \tag{7.3}$$

Equation 7.3 can be affected for low and high relevance values propagated from the top layer. Therefore, in order to not have unbounded relevance values in the feature-input layer (Binder et al., 2016) introduce a stabilizer variable ε having always a positive sign.

LRP-z and LRP-ε

The LRP-z method is defined as the most simple LRP method without any relevance numerical balance method, thus calculating the relevance propagation using the linear propagation defined in Equation 7.3.
Equation 7.3 can be rewritten and split in two part Equations, where a positive and negative relevance values are controlled through the propagation process. These two types of relevance can be balanced per layer as we explain with the $\alpha\beta$ rule in the Appendix B and sections below. The split Equation 7.3 is expressed in Equation 7.4 and Equation 7.5.

$$R_i^l = \sum_j \frac{z_{ij}}{\sum_i z_{i,j} + \varepsilon \texttt{sign}\left(\sum_i z_{ij}\right)} R_j^{l+1} \qquad \sum_i z_{ij} \geq 0 \tag{7.4}$$

$$R_i^l = \sum_j \frac{z_{ij}}{\sum_i z_{i,j} - \varepsilon \texttt{sign}\left(\sum_i z_{ij}\right)} R_j^{l+1} \qquad \sum_i z_{ij} < 0 \tag{7.5}$$

The numerical stabilizer ε is included in Equation 7.3 obtaining a split function in Equation 7.4 and 7.5 that yields the conservation law. The LRP-ε method is configured by the numerical stabilization on the denominator of Equations 7.4 and 7.5 making the relevance-maps cleaner as we will see in the sections below.

Flat $\alpha\beta$ Rule

With the purpose of establishing a bilateral control on the relevance propagation across classifier's layers the positive z_{ij}^{+}, and the negative z_{ij}^{-} part of the relevance adjustment parameters. These parameters α and β are set to adjust the positive and negative relevance values controlling the amplitude overflow with a numerical constraint $\alpha - \beta = 1$. Including these adjust parameters the new expression for the relevance propagation is re-written in Equation 7.6.

$$R_i^l = \sum_j \left[\alpha \frac{z_{ij}^{+}}{\sum_j z_{ij}^{+}} + \beta \frac{z_{ij}^{-}}{\sum_j z_{ij}^{-}} \right] \tag{7.6}$$

Following the model on Equation 7.6 the relevance can be propagated through a Deep ConvNet's layers limiting the initial relevance calculation using a Taylor-type decomposition (Nik and Soleymani, 2013) converting the classifier's output decision to a vector of relevances with the same dimensionality of the predecessor layer transforming the domain across layers $f : R^N \rightarrow R^M$ where N is the dimensionality of the decision space, and M the dimensionality of the previous layer. We will describe this model in detail in Appendix B.

LRP numerical balance - A,B Presets

iNNvestigate authors include the LRP with flat rule adding the stabilizer with a value of $\varepsilon = 0.1$ following the enhanced model, Equation 7.6 can be transformed to Equation including the stabilizer in the denominator to clean the relevance propagation from the positive and negative signs.

Previous implementations have include LRP without numerical balance using EEG features with 2D arrangement for motor imagery decoding (Sturm et al., 2016). However, without the adequate numerical balance the LRP-z or the LRP-ε approaches won't describe the more reliable locations for the more relevant or unrelevant features.

In the iNNvestigate package proposed a modification of α and β adjustment parameters. The $\alpha\beta$ preset A set the values in $\alpha = 1$ and $\beta = 0$, and the preset B set the values in $\alpha = 2$ and $\beta = 1$ (Montavon, Samek, and Müller, 2018). To illustrate the LRP-z reliability limitations we show the results from (Weitz et al., 2018) in Figure 7.1 where a face emotion decoding problem becomes very tricky without using an adequate numerical balance such as the A and B presets.

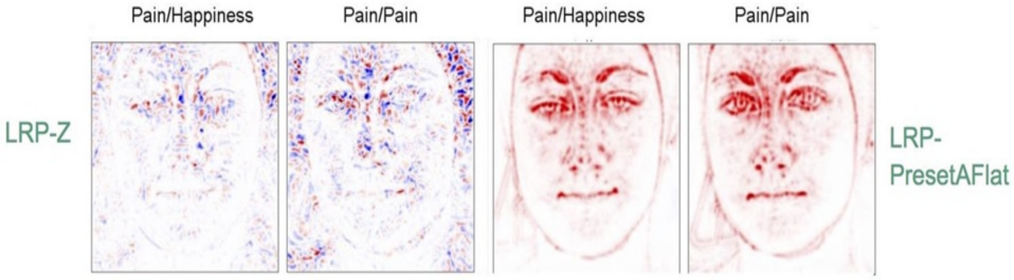


Fig. 7.1 The relevance-map propagation entanglement produced without using an adequate numerical balance including the adequate $\alpha\beta$ preset. These results are extracted from (Weitz et al., 2018)

In the following subsections we will illustrate the more robust methods such as PatternNet, Pattern Attribution, and Smooth-Grad. The Deconvolution-based methods will be described with more detail in the Appendix B. We will also show the results for the most reliable methods applied to the sample #1 described in the previous chapter with the comparison between TD and ASD groups.

7.2.3 PatternNet and Pattern Attribution

PatternNet and Pattern-Attribution (Kindermans et al., 2017a) are robust saliency methods that can isolate the input signal model representation from any disturbing perturbation during the training process using linear modelling. These methods optimize the learning representation making it more disentangled from common linear and multidirectional disturbances across each feature-space per conv-pool blocks.

The initial representation of inputs per layer in the Deep ConvNet is modeled using a typical linear model x affected by any kind of deterministic/non-deterministic disturbance commonly called by Google Brain researchers a *"distractor"* or d. The linear model from input feature-space has a signal component denoted as s.

PatternNet and Pattern-Attribution methods optimize the linear model to increase the statistical correlation between the signal component s and the linear representation x associated with the corresponding layer output y.

The linear model is defined as $x = s + d$, and each component has a parametric direction parameters denoted as a_s and a_d for the signal s and the distractor d. Therefore, the linear model can be re-written as $x = a_s y + \varepsilon a_d$ where ε is a multidimensional noise source.

With this new linear model PatternNet and Pattern-Attribution propose methodologies to find a set of filter weights ω to maximize the signal component taking into account the conv-pool layer model $y = \omega^T x$.

In (Kindermans et al., 2017a) the authors first analyze gradient-based methods where the propagation is not deriving the level of signal included in x, but only optimizing the filter values based on the input-output relationship $\frac{\partial y}{\partial x} = \omega$.

With a gradient-based model the distractor can not be detected properly using relevance propagation. Thus, re-formulating the derivative propagation the new linear model will be affected by the distractor. Changing the approach such as Deconvolution and Guided-Backpropagation process use a similar gradient propagation in comparison with gradient-based methods without isolating the level of signal propagated through the Deep ConvNet layers.

As a third case the authors analyze relevance-based methods such as LRP-z and LRP Deep-Taylor approaches. Both methods are sensitive to the Taylor constraint root value denoted by x_0 as we explain in the Appendix B. To simplify the distractor model for LRP methods $x_0 = d$. Following the isolation purpose PatternNet and Pattern-Attribution estimate a signal contribution from a different region associated with a completely different model for the distractor.

PatternNet and Pattern-Attribution model the distractor assuming this equivalences $y = \omega^T x$, $y = \omega^T s$, and $\omega^T d = 0$ to compute a new signal estimator denoted as $S(x) = s$, and a correlation estimator for quality measure denoted as ρ. This new estimator assumes a complete isolation from distractor using an estimator denoted by $\hat{d}$.

To compute ρ we assume that the distractor and the signal estimators should not be singular, and should be decomposed in eigenvalues. Equation 7.7 shows that a better signal estimator

$S(x)$ can be evaluated with a more reliable neuron-wise explanation if ρ is higher.

$$\rho(S) = 1 - \texttt{max}_{corr}\left(\omega^T x, u^T(x - S(x))\right) = 1 - \texttt{max}\left(\frac{u^T cov(d,y)}{\sqrt{\sigma_{u,d}\sigma_y}}\right) \tag{7.7}$$

To simplify the ρ quality criterion as LRP proposed we assign $\sigma_{u,d} = \sigma_y$ where u is the new modelled input. The estimator $S(x)$ can be reduced as a filtered signal estimator $S_\omega(x)$. Therefore, we define $S_\omega(x) = \frac{\omega}{\omega^T\omega}\omega^T x$ propagating the estimator through the network with an evident association.

With the $S_\omega(x)$ definition PatternNet and Pattern-Attribution methods propose two different estimators approaches that can increase the quality criterion expressed in Equation 7.7.

The first estimator is a covariance-based linear estimator between x and the corresponding output y. This covariance estimator is denoted as $S_a(x)$, and the covariance between this estimator and corresponding layer output y should be evaluated with a zero covariance between the distractor and the output $cov[d,y] = 0$. The new linear estimator model is defined in Equation 7.8 and modelling $cov[x,y]$ assuming the previous statements we can calculate the estimator a.

$$cov[x,y] = cov[S_a(x),y] \Rightarrow cov[x,y] = cov[a\omega^T x, y] \Rightarrow cov[x,y] = acov[y,y] \Rightarrow a = \frac{cov[x,y]}{\sigma_y} \tag{7.8}$$

a is equivalent to a filter estimator explained in (Kindermans et al., 2017b). This estimator is related to distractor components when dense ReLU layers are included as our pipeline. Thus, an alternative negative-positive new estimator should be defined to reduce the effect of ReLU as a distractor.

This new estimator is denoted as $S_{a+-}(x)$ including the negative and positive relevance values and balancing the distractor effect on ReLU layers. To define this new estimator the covariance between x and y, and the covariance between $S_a(x)$ and y must be defined in a bilateral way as well.

$$cov[xy] = \pi_+\left[E_+[xy] - E_+(x)E_+(y)\right] + (1-\pi_+)\left[E_-[xy] - E_-(x)E_-(y)\right] \tag{7.9}$$

$$cov[S(x)y] = \pi_+\left[E_+[S(x)y] - E_+(S(x))E_+(y)\right] + (1-\pi_+)\left[E_-[S(x)y] - E_-(S(x))E_-(y)\right] \tag{7.10}$$

Equations 7.9 and 7.10 re-define the covariances mentioned above including the numerical ratio π_+ to be propagated across the trained network. To complete the balance across the network we define the expected values $E[xy]$, $E[x]$, $E[y]$, $E[S(x)y]$, and $E[S(x)]$. All these new expected value should be redefined bilaterally too E_+ and E_-.

Equalling both covariances $cov[xy] = cov[S(x)y]$, and closing the range of the estimator for using only the positive relevance it is possible to define a new positive estimator denoted as a_+, assuming a minimum covariance between x and y with the distractor d.

$$a_+ = \frac{[E_+[xy] - E_+(x)E_+(y)]}{[\omega^T E_+[xy] - \omega^T E_+(x)E_+(y)]} \tag{7.11}$$

Equation 7.11 includes the formal definition of a_+ to modulate the final relevance propagation. Equation 7.11 also extend the definition of the optimized estimator from Equation 7.10 introducing the independent values a_+ and ω^T out of the brackets in Equation 7.11.

The next step is propagating a_+ through the Deep ConvNet to reduce significantly the distractor effect as well as increment the quality criterion ρ substantially as we can see on the experiments in (Kindermans et al., 2017a,b).

With the estimator a_+ defined we can point the PatternNet and Pattern-Attribution main difference in the propagation modalities. PatternNet uses a similar propagation as LRP DeepTaylor (Appendix B) but without propagating the filtered estimator.

PatternNet propagates a_+ through the layers modulating linearly the distractor incidence. As for Pattern-Attribution method the numerical incidence of the learnt weights ω^T is propagated using a linear product with a_+ denoted as $\omega^T a_+$ through the network.

The iNNvestigate package also includes other important methods such as the LRP flat presets A and B as we explained in the section above. We will report the results comparing the most important relevant maps across participants for the Sample #1.

7.3 Relevance-maps - Comparison TD and ASD

For the the relevance-maps show in the following Figures the dark red or hotter points can be defined as *"relevant"* and the blue or colder spots *"un-relevant"*. This rectangular form is denoted here as a heat-map with the size of the input feature-set 752 × 30 as we explained in the chapters above. As we explained in the chapters above the 752 points cover the time range between $[0 - 1500]$ ms, and the 30 channels are the final cleansed channels such as: FT9,F7,FC5,FP1,FZ,FP2,F4,F8,FC6,FT10,F4,F3,FC1,C3,FC1,FC2,C4,T7,CP5,CP1,CZ,CP2, P4,P8,CP6,T8,P7,P3,Pz,O1,O2, and Oz. All these methods represent a mathematical intuition about how the neural activity is decoding emotion through the Deep ConvNet training. To show the relevance-maps we use a 2D representation denoted as colormap with the channels in y-axis, and time in x-axis.

We use topo-maps or topographic plots from EEGlab (Delorme and Makeig, 2004) to compare the relevance between TD and ASD groups averaging the relevance level in 5 time ranges such as [0-500],[250-750],[500-1000],[750-1250], and [1000-1500] ms after the stimulus onset.

For the statistical comparison we use an ANOVA one-way with Bonferroni correction. The F values are grouped based on the initial comparison values, and only the p-values were corrected. To complement this evaluation we measure the differences in terms of statistical origin, in other words if TD and ASD relevance-maps correspond or not to the same type of CDF using the Kolmogorov-Smirnov test (K-S test) with a confidence value of 0.05 (Banerjee and Pradhan, 2018).

The relevance-maps are presented using the heat-maps and the topo-maps in the same Figure being the heat-map up and the 5 topo-maps down. The relevance-maps are also presented with an amplitude normalized between $[-1,1]$ following the normalization of (Bach et al., 2015).

In the following relevance-maps the values $R_q^1 \geq 0$ contribute positively in the correct emotion decoding *"relevant"*, and the values $R_q^1 < 0$ are *"un-relevant"* or do not contribute to the correct emotion decoding.

R_q^1 is calculated per method and group from the trained Deep ConvNet following the models explained in the sections above. For each method and group we report a difference relevance-map substracting the normalized relevance obtained for the TD group with the relevance-map obtained for the ASD group normalizing the final difference relevance-map between $[-0.1,0.1]$ for some methods and $[-0.02,0.02]$ when the difference is not significant.

7.3.1 LRP A,B flat presets results

In this section we will show the relevance-maps obtained using the LRP A and B flat preset as we explained in the section above. The methodology for showing results here will be grouping and averaging the relevance calculated per each subject, and for each *hit* registered by subject and for each emotion. In summary, we will show the relevance-map for *Happy*, *Sad*, *Angry*,*Fear*, and the *Average* relevance-map averaging the relevance-maps for all the classes.

Figures 7.2, 7.3, 7.4, 7.5, and 7.6 show the relevance-maps with the heat-maps and topo maps for the LRP preset A for *Average*, *Happy*, *Sad*, *Angry*, and *Fear* classes respectively. Evaluating the Bonferroni corrected ANOVAs across the 5 topo-maps, and for each class we did not found any significant difference between TD and ASD groups for emotions Happy, Angry, and Fear after correction F(1,87)<1.243, p>0.05.

We only found significant difference after correction in the emotion Sad for LRP A preset method and for the ranges **[1000-1500]ms, F(1,87)=13.54, p=0.0021, TD > ASD.**

Despite the non-significant differences in LRP A we found a consequent more relevant block in late components approximately after 1000ms for LRP A and LRP B presets in ASD groups observed in Angry and Fear emotions, and this effect is also replicated in the subsequent Average class.

We can correlate this feature importance obtained in LRP A and B presets with the correct emotion decoding observed in negative emotions in Chapter 5 in comparison with FER performances.

For LRP A some K-S tests show a null-hypothesis $h = 1$ acceptance, thus supporting that relevance-maps patterns between TD and ASD groups come from different statistical distributions, and without being significant in the ANOVA comparisons. For emotions such as Happy in ranges between [0-500]ms, p=0.0484, and between [750-1250]ms, p=0.0132, and for Sad in ranges between [1000-1500]ms, p=0.0016.

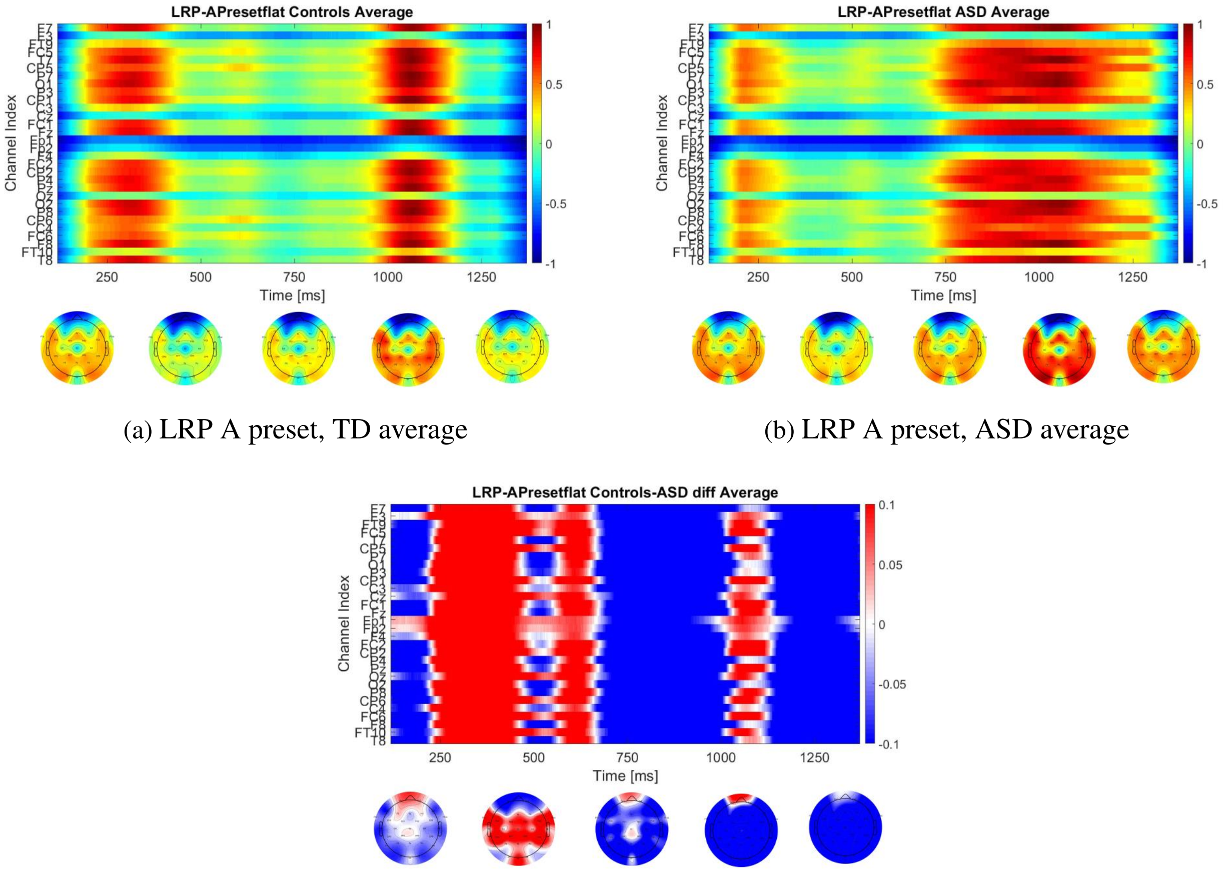
Despite LRP A and LRP B preset relevance-maps were similar between groups, we only found significant differences after correction for LRP preset B in the ranges due to the participant's high variability found in LRP A in comparison with LRP B presets. We found significant differences after correction in **[1000-1500ms], F(1,87)=7.889,p=0.0344, TD < ASD**, and **[0-500ms], F(1,87)=11.56, p=0.0033, TD > ASD** for the average class.

Figures 7.7, 7.8, 7.9, 7.10, and 7.11 show the relevance-maps with the heat-maps and topo maps for the LRP preset B method and for *Average*, *Happy*, *Sad*, *Angry*, and *Fear* classes. Although the LRP A preset and LRP B preset show similar relevance-map patterns, the results found for the LRP B preset are very different in comparison with the LRP A preset method.

Evaluating LRP B preset we found more significant differences after correction in emotions such as Sad in ranges between **[750-1250]ms, F(1,87)=8.491, p=0.0141, TD > ASD**, and between **1000-1500ms, F(1,87)=13.54, p=0.0005, TD > ASD** supporting again the differences in late components, Angry in ranges between **[0-500]ms, F(1,87)=10.85, p=0.0095, TD > ASD**, and between **[1000-1500]ms, F(1,87)=9.667, p=0.0102, TD < ASD**, and for Fear emotion in ranges between **[0-500]ms, F(1,87)=23.47, p=7.6e-6, TD < ASD,** between **[500-1000]ms F(1,87)=7.193, p=0.0263, TD > ASD** and between **[750-1250]ms, F(1,87)=9.313, p=0.0121, TD < ASD.** These differences support the late greater relevance component observed in ASD groups in comparison with TD, and a similar relevance pattern observed between Angry and Fear classes maps.

The K-S tests for the LRP B preset method are accepting the null-hypothesis for the same time regions where the ANOVAs are significant for 0-500 ms p=0.0198, and for 1000-1500ms p=0.0244 for the average class. In consonance with the ANOVA comparisons the K-S tests accept the null-hypothesis $h = 1$ for the same ranges and the same emotion classes such as Sad emotion between [0-500]ms,p=0.0484 and between [750-1250]ms,p=0.0132, Angry in ranges between [750-1250]ms, p=0.0078, and for [1000-1500]ms p=0.0006, and Fear covering all ranges such as [0-500]ms, p=5.5e-6, between [250-750]ms, p=0.0439, between [500-1000]ms, p=0.0088, between [750-1250]ms, p=0.0030, and between [1000-1500]ms, p=0.0251. These K-S p-values were not corrected due to they are multiple comparison but statistical origin inference analyses.

As an overall effect we observed a consolidated significance difference on negative emotions such as Angry and Fear, and the corresponding Average relevance maps comparing late time ranges such as [750-1250]ms and [1000-1500]ms across TD and ASD groups.

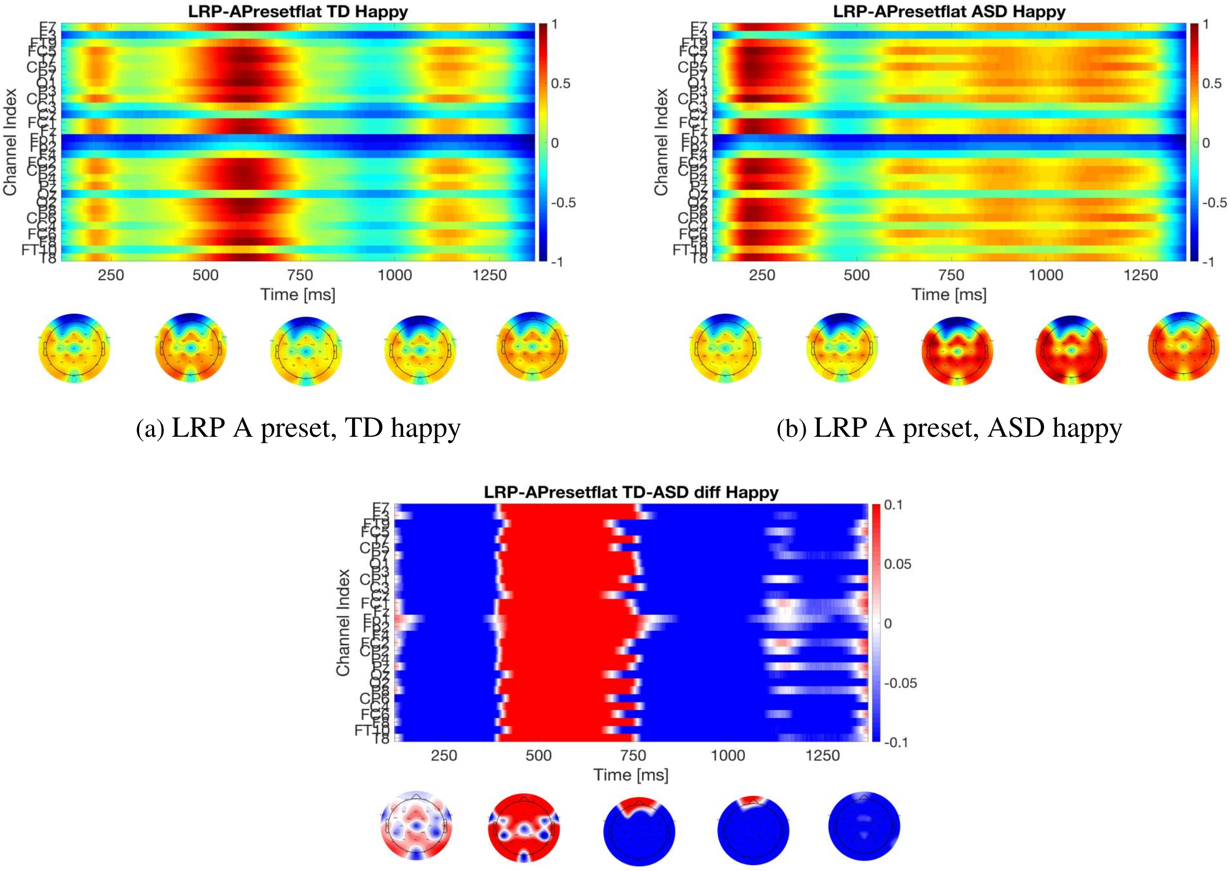


(a) LRP A preset, TD average

(b) LRP A preset, ASD average

(c) LRP A preset, TD-ASD diff average

Fig. 7.2 LRP A average class relevance-map for TD 7.2a, and ASD 7.2b, and the differences between TD-ASD 7.2c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

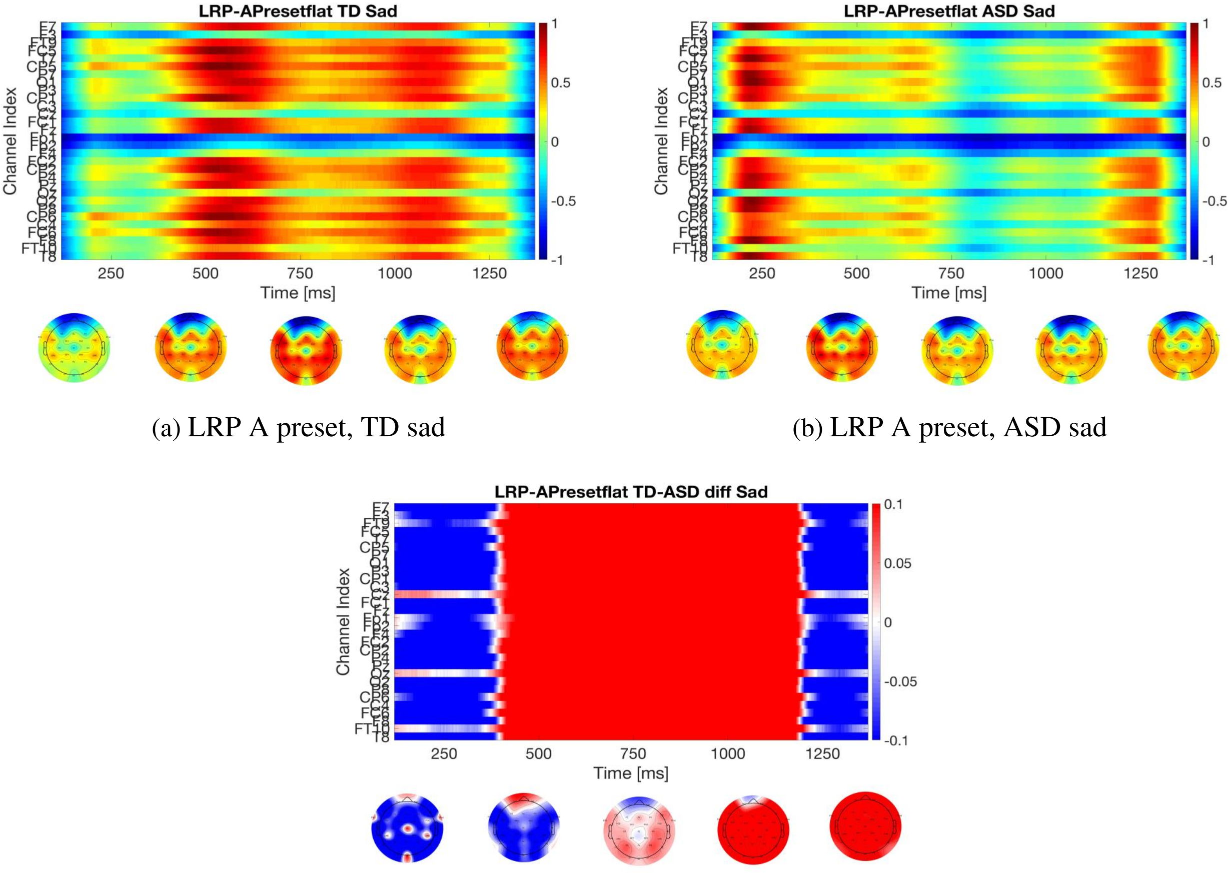


(a) LRP A preset, TD happy

(b) LRP A preset, ASD happy

(c) LRP A preset, TD-ASD diff happy

Fig. 7.3 LRP A Happy relevance-map for TD 7.3a, and ASD 7.3b, and the differences between TD-ASD 7.3c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

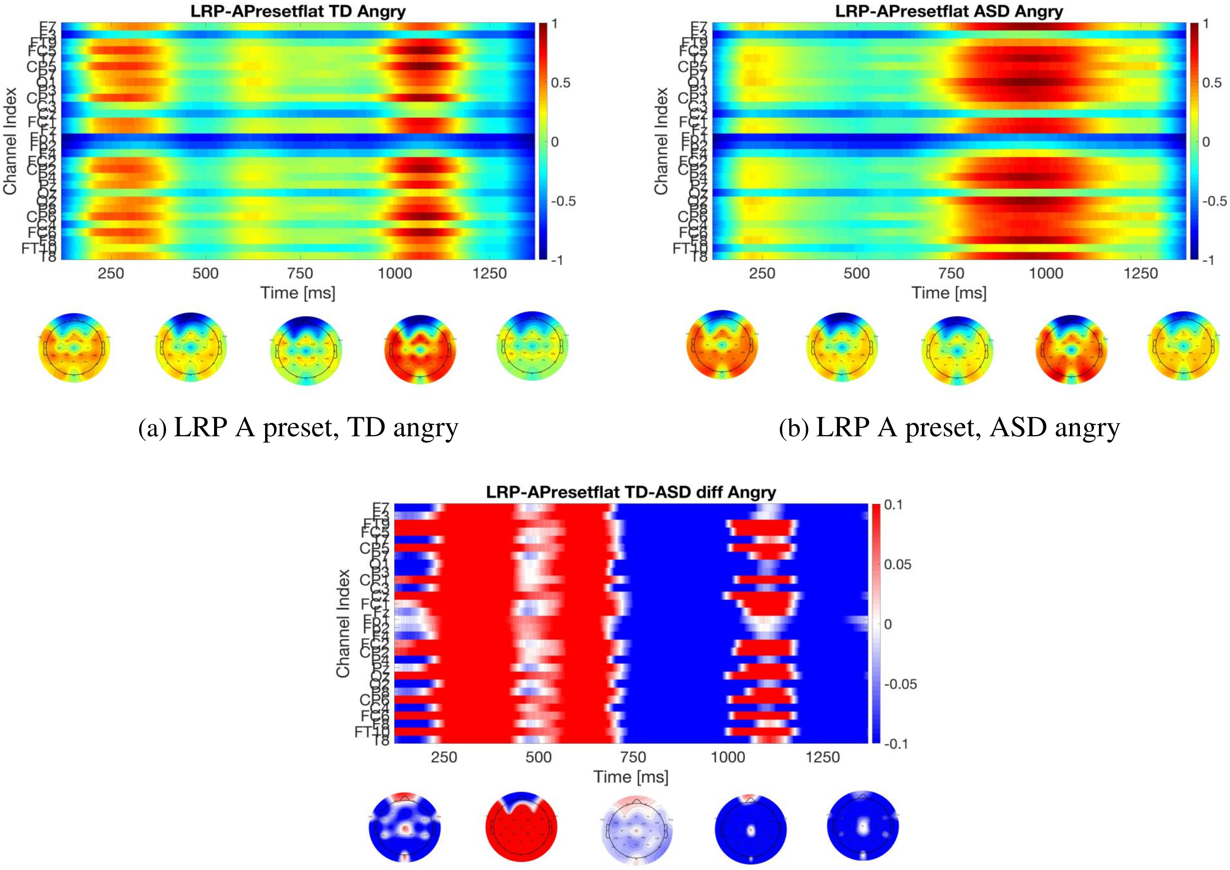


(a) LRP A preset, TD sad

(b) LRP A preset, ASD sad

(c) LRP A preset, TD-ASD diff sad

Fig. 7.4 LRP A Sad relevance-map for TD 7.4a, and ASD 7.4b, and the differences between TD-ASD 7.4c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

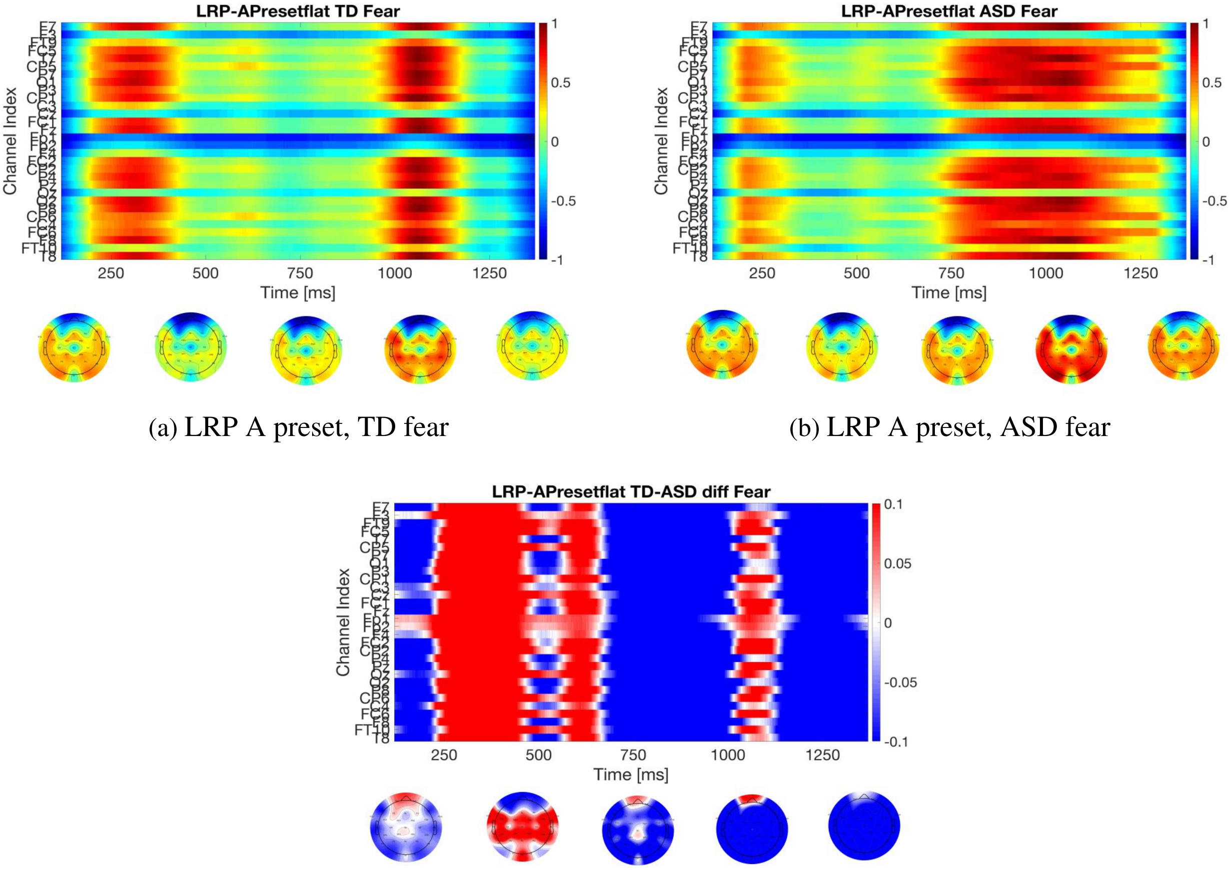


(a) LRP A preset, TD angry

(b) LRP A preset, ASD angry

(c) LRP A preset, TD-ASD diff angry

Fig. 7.5 LRP A Angry relevance-map for TD 7.5a, and ASD 7.5b, and the differences between TD-ASD 7.5c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

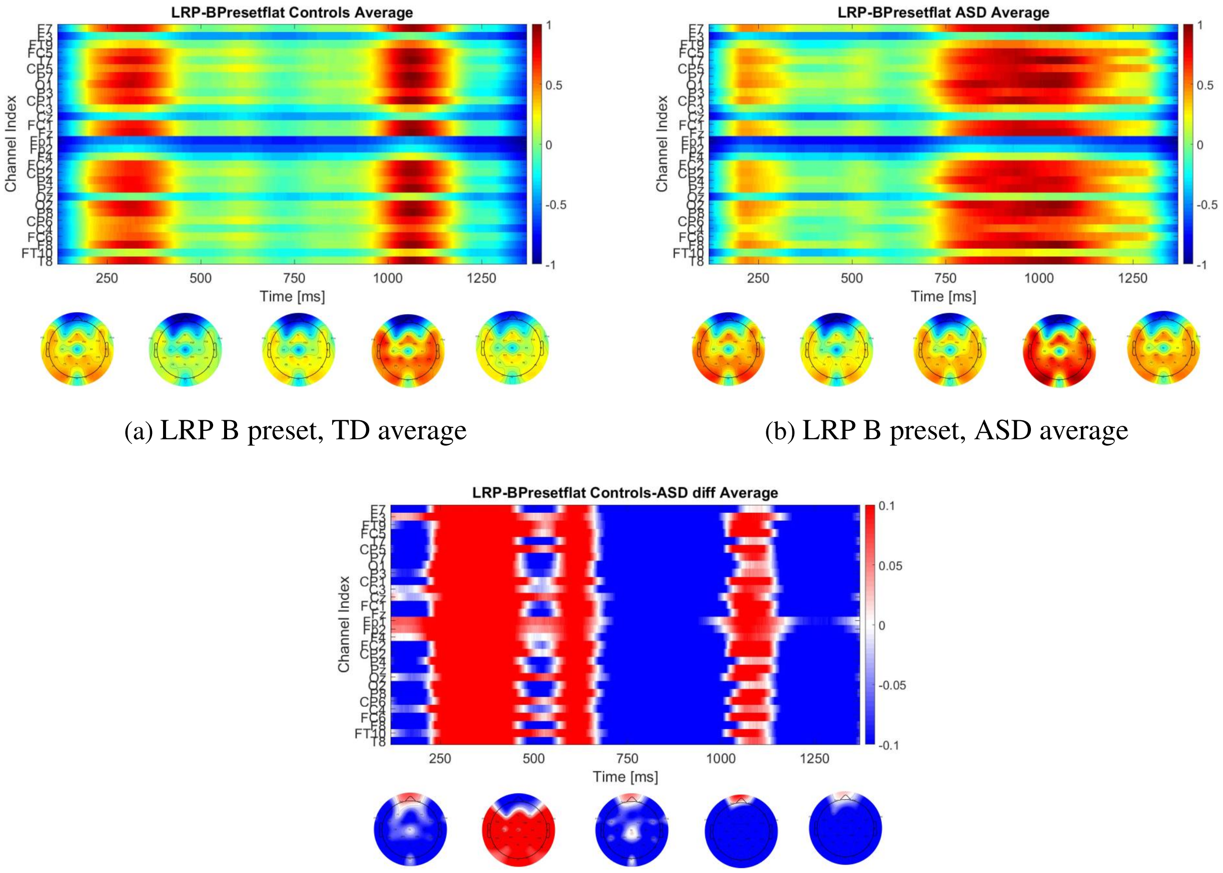


(a) LRP A preset, TD fear

(b) LRP A preset, ASD fear

(c) LRP A preset, TD-ASD diff fear

Fig. 7.6 LRP A Fear relevance-map for TD 7.6a, and ASD 7.6b, and the differences between TD-ASD 7.6c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

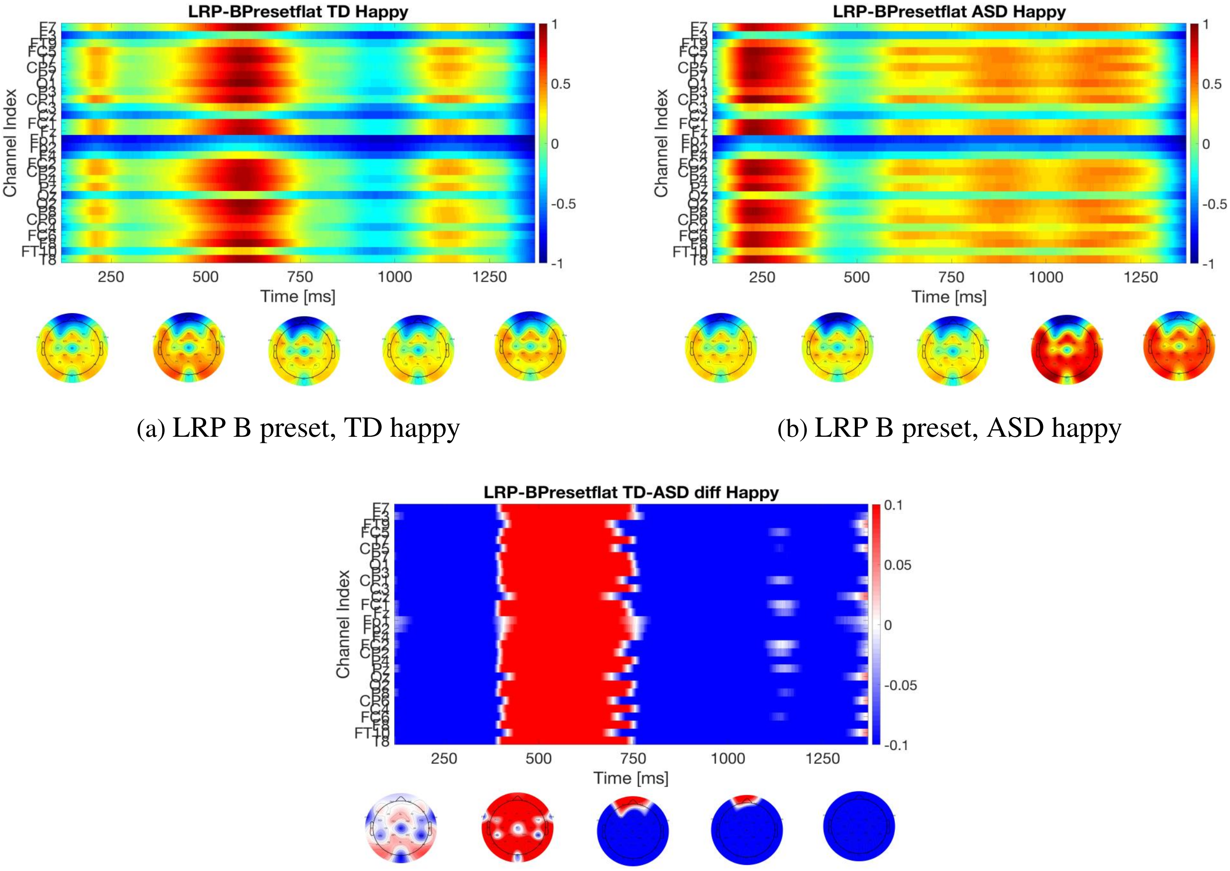


(a) LRP B preset, TD average

(b) LRP B preset, ASD average

(c) LRP B preset, TD-ASD diff average

Fig. 7.7 LRP B average class relevance-map for TD 7.7a, and ASD 7.7b, and the differences between TD-ASD 7.7c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

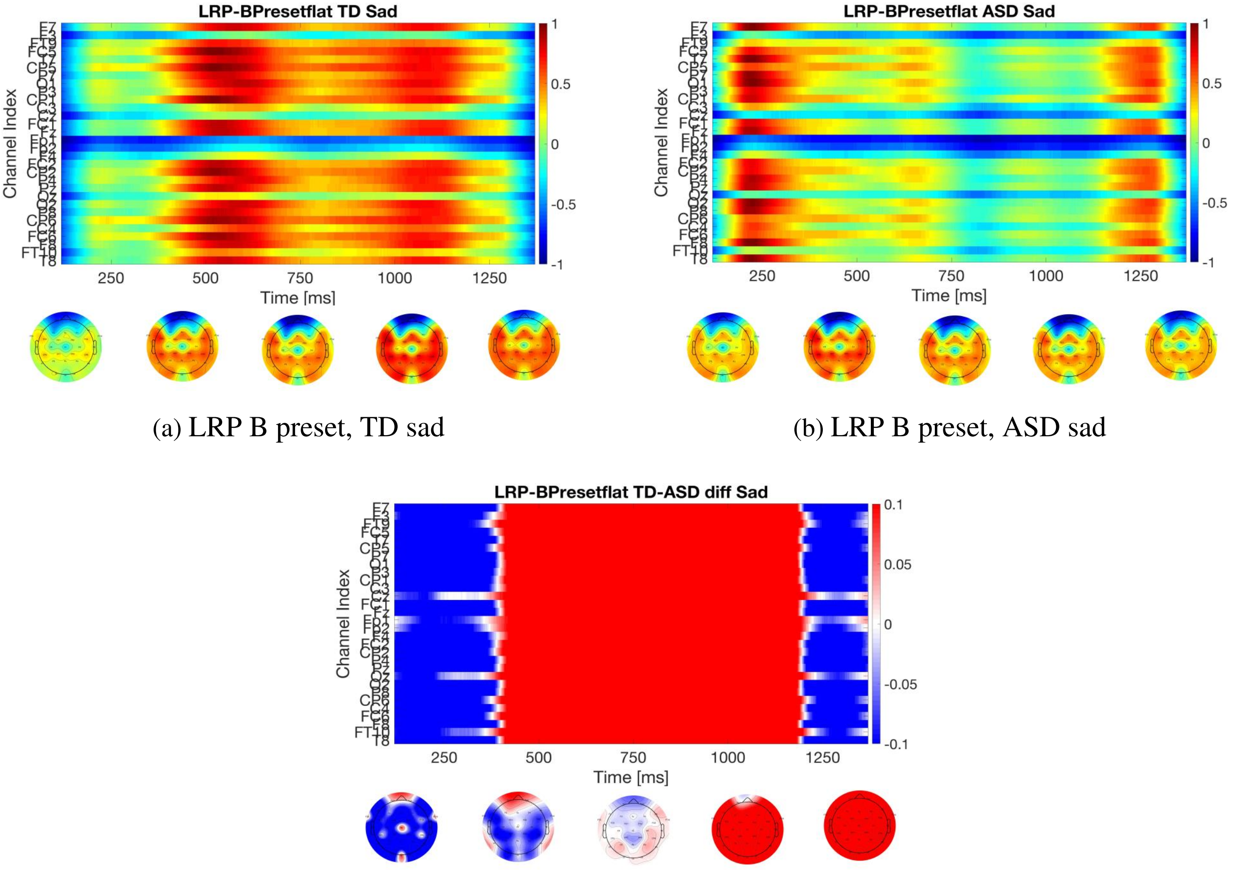


(a) LRP B preset, TD happy

(b) LRP B preset, ASD happy

(c) LRP B preset, TD-ASD diff happy

Fig. 7.8 LRP B Happy relevance-map for TD 7.8a, and ASD 7.8b, and the differences between TD-ASD 7.8c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

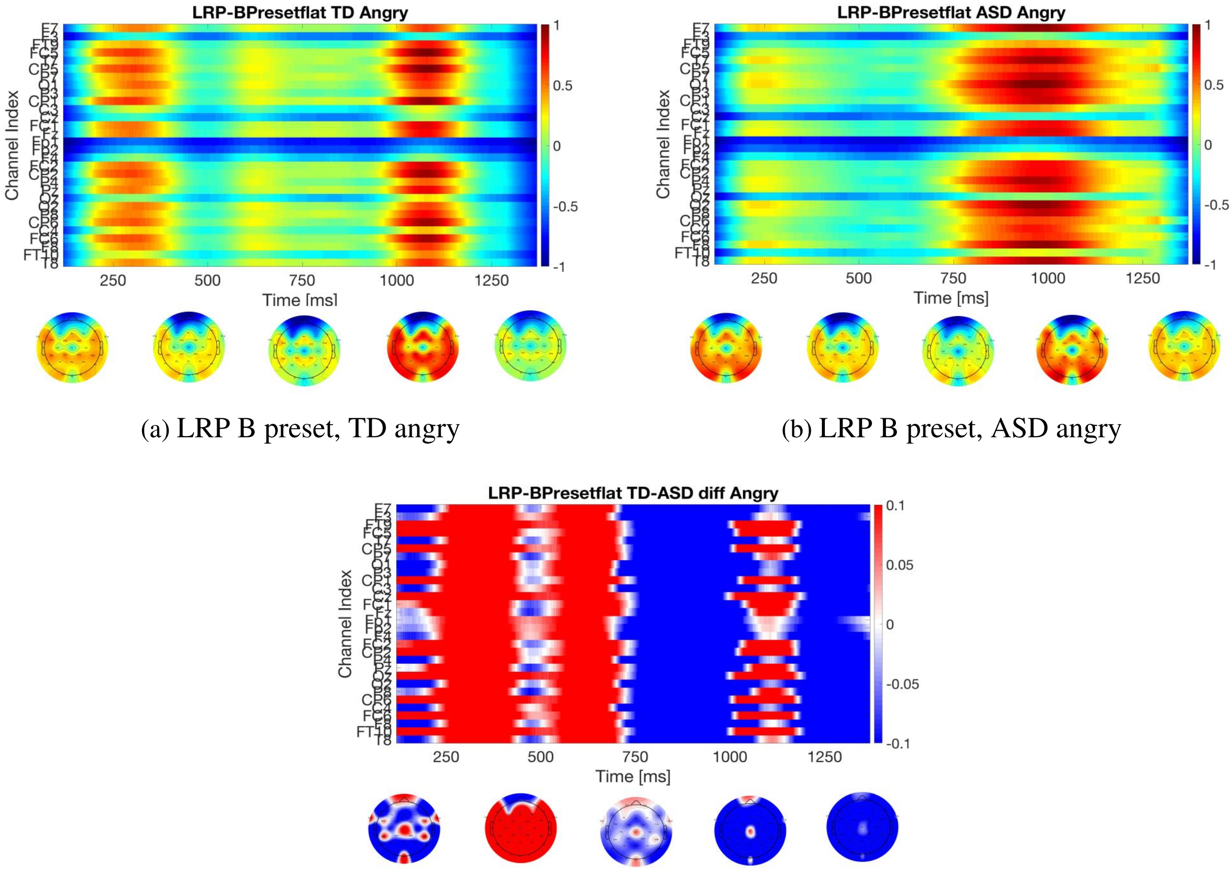


(a) LRP B preset, TD sad

(b) LRP B preset, ASD sad

(c) LRP B preset, TD-ASD diff sad

Fig. 7.9 LRP B Sad relevance-map for TD 7.9a, and ASD 7.9b, and the differences between TD-ASD 7.9c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

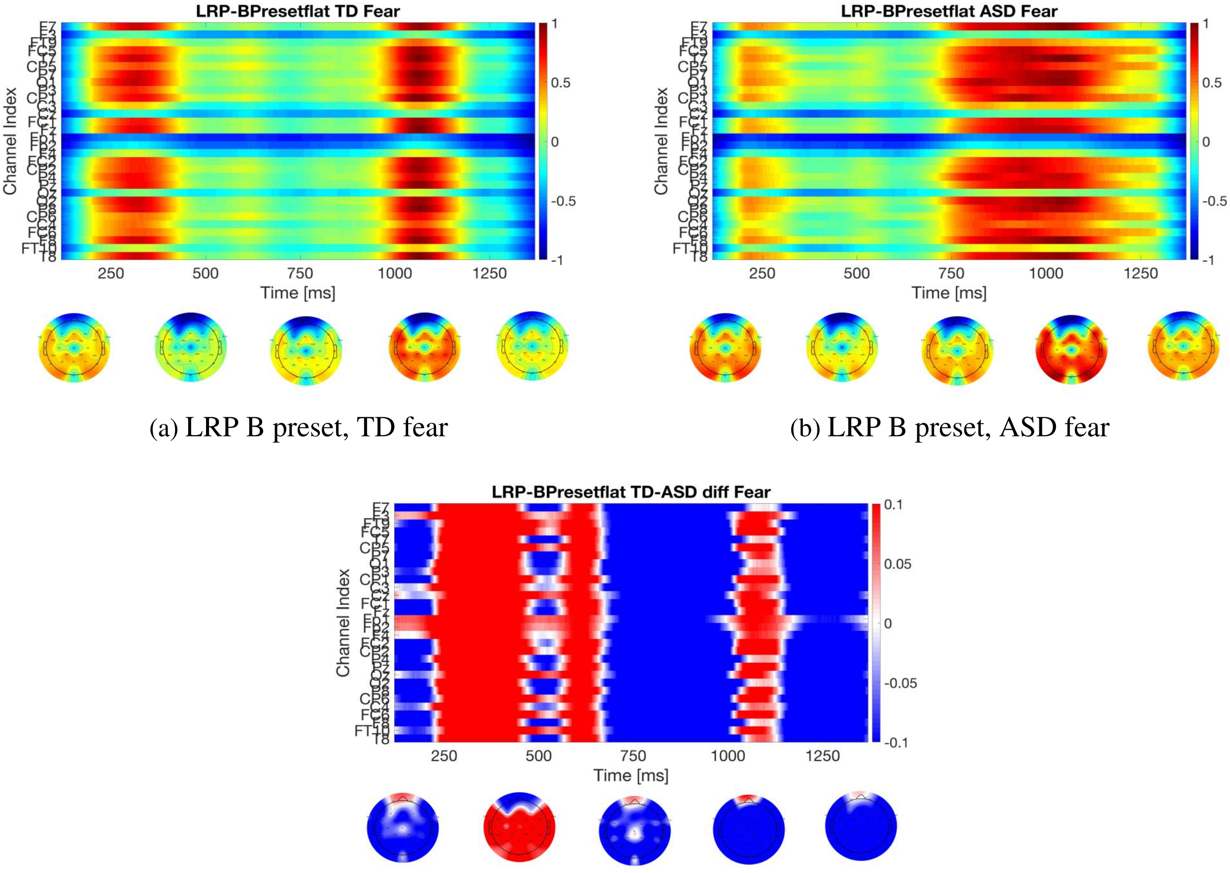


(a) LRP B preset, TD angry

(b) LRP B preset, ASD angry

(c) LRP B preset, TD-ASD diff angry

Fig. 7.10 LRP B Angry relevance-map for TD 7.10a, and ASD 7.10b, and the differences between TD-ASD 7.10c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.



(a) LRP B preset, TD fear

(b) LRP B preset, ASD fear

(c) LRP B preset, TD-ASD diff fear

Fig. 7.11 LRP B Fear relevance-map for TD 7.11a, and ASD 7.11b, and the differences between TD-ASD 7.11c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

The results obtained here for the LRP A and LRP B presets are the results with more significant p-values after correction in comparison with other methods comparing the topomaps relevance patterns between TD and ASD groups. In the following analysis we will use LRP B as the higher significance quota in comparison with other saliency methods.
As we mentioned above these results are confirming the late activation related with ASD neural emotion decoding in comparison with TD (Benning et al., 2016; Black et al., 2017). The spatial resolution of the input feature-set is not balanced in comparison with the time resolution. Therefore, we expect a less relevance resolution on the channels in comparison with the time domain (Selvaraju et al., 2016). We will show the results for PatternNet and Pattern-Attribution results in the following subsections.

7.3.2 PatternNet and Pattern-Attribution results

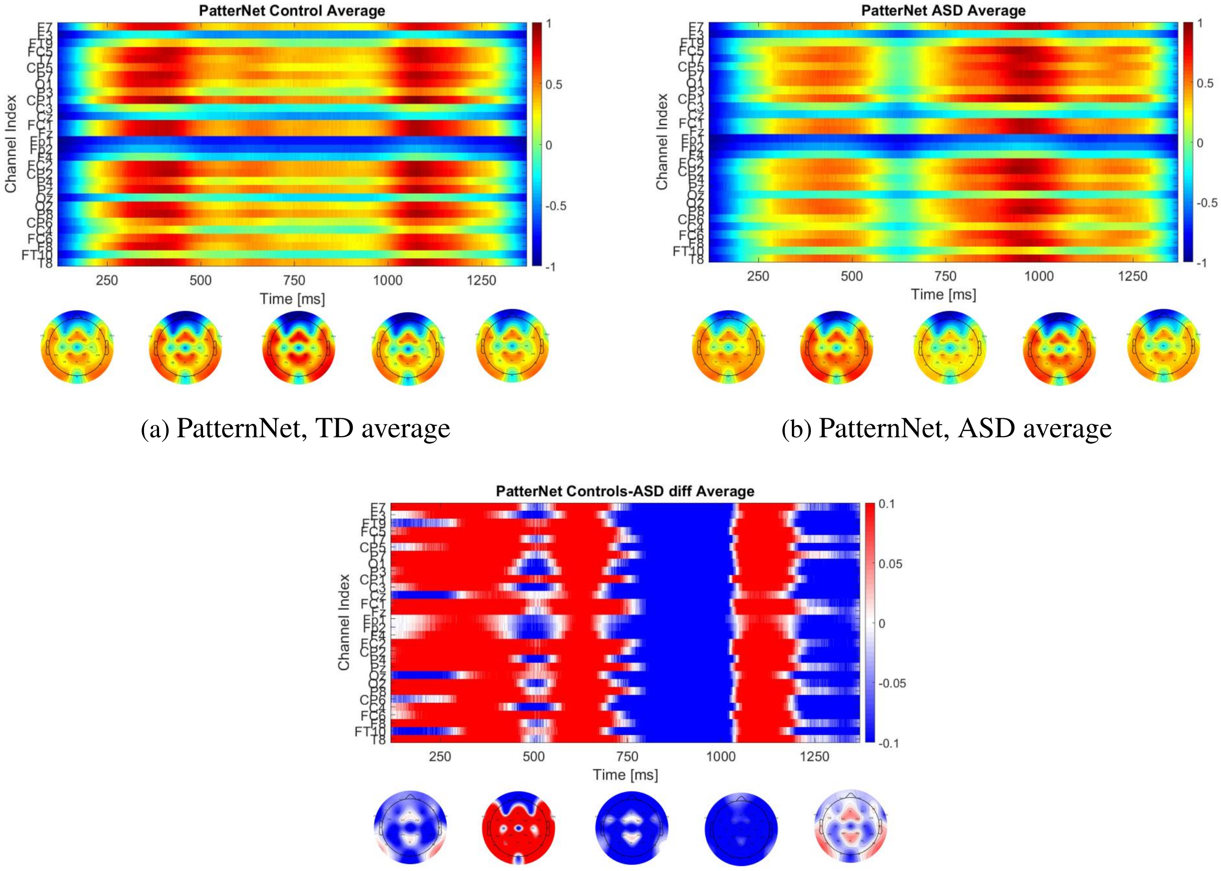
Figures 7.12, 7.13, 7.14, 7.15, and 7.16 show the relevance-maps for TD and ASD groups, and the difference TD-ASD relevance-maps for the PatternNet saliency method for *Average*, *Happy*, *Sad*, *Angry*, and *Fear* classes respectively.

On the other hand, Figures 7.17, 7.18, 7.19, 7.20, and 7.21 show the relevance-maps for TD and ASD groups, and the difference TD-ASD relevance-maps for the Pattern Attribution saliency method for *Average*, *Happy*, *Sad*, *Angry*, and *Fear* classes too.

For all the time ranges, PatternNet did not show any significant difference between the groups for any emotion class such as Average F(1,87)<0.126, Happy F(1,87)<1.334, Sad F(1,87)<0.556, Angry F(1,87)<0.775, and Fear F(1,87)<0.889 with all the p-values p>0.05 after correction.

The same results occur for Pattern Attribution were the propagation of the weights ω^T include a noisier relevance-map pattern in the propagation law across the Deep ConvNet as we can see in the corresponding Figures. We did not found any significant difference after correction for the Pattern Attribution method including Average class F(1,87)<0.034, and emotion classes such as Happy F(1,87)<0.222, Sad F(1,87)<0.045, Angry F(1,87)<0.067, and Fear F(1,87)<0.178 with all the p-values p>0.05 after correction.

The K-S tests were all $h = 0$ rejecting the null-hypothesis and we can not find any difference in the statistical distribution between TD and ASD groups and across all the emotions such as *Average*, *Happy*, *Sad*, *Angry*, and *Fear*.

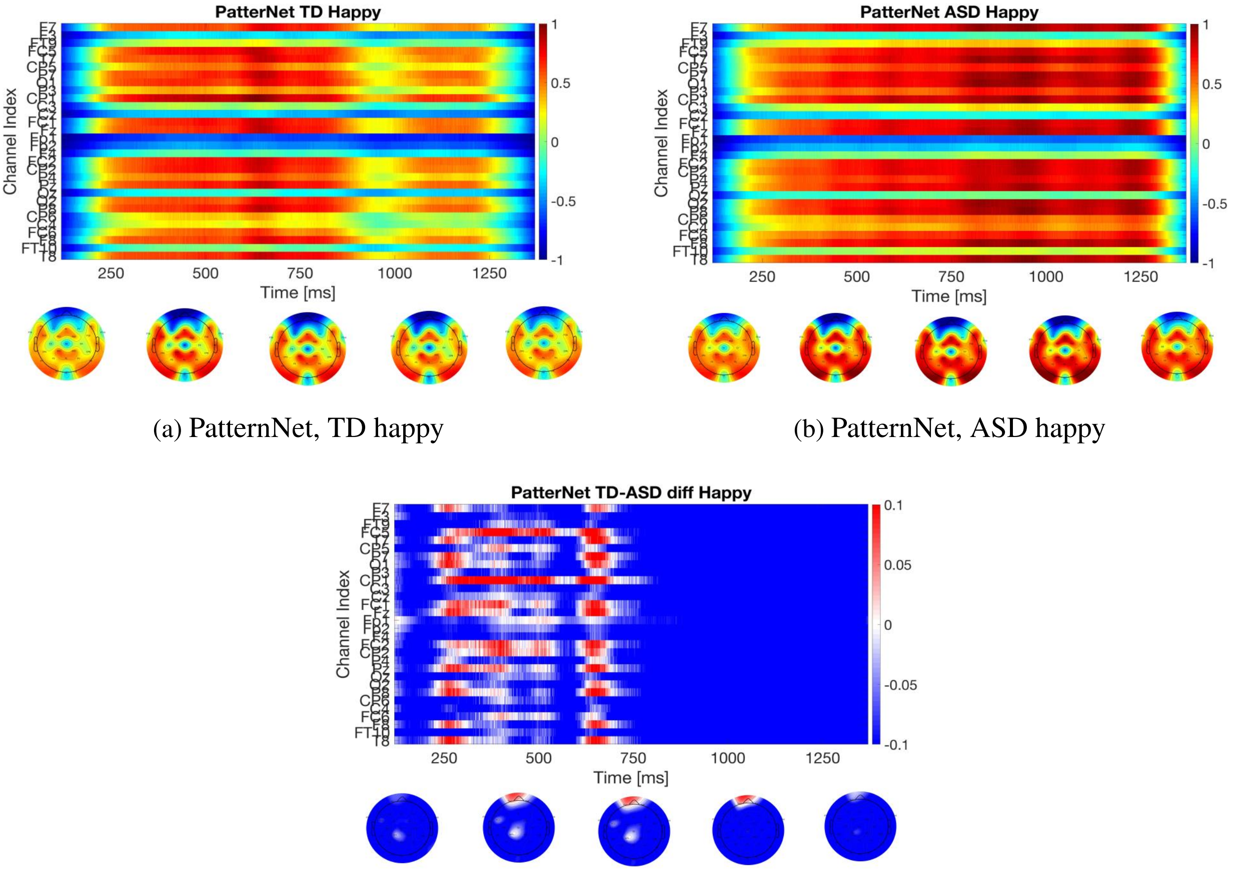


(a) PatternNet, TD average

(b) PatternNet, ASD average

(c) PatternNet, TD-ASD diff average

Fig. 7.12 PatternNet average class relevance-map for TD 7.12a, and ASD 7.12b, and the differences between TD-ASD 7.12c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

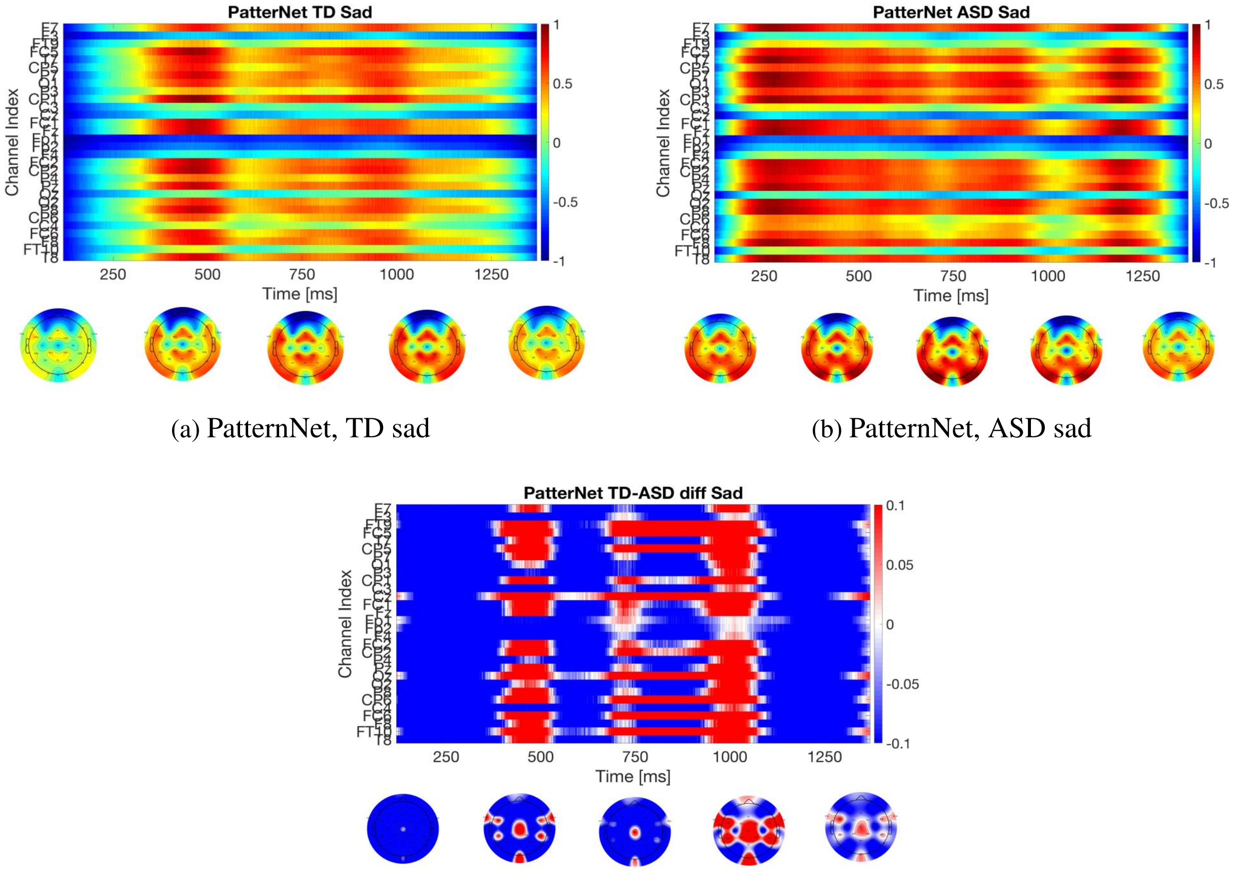


(a) PatternNet, TD happy

(b) PatternNet, ASD happy

(c) PatternNet, TD-ASD diff happy

Fig. 7.13 PatternNet Happy relevance-map for TD 7.13a, and ASD 7.13b, and the differences between TD-ASD 7.13c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

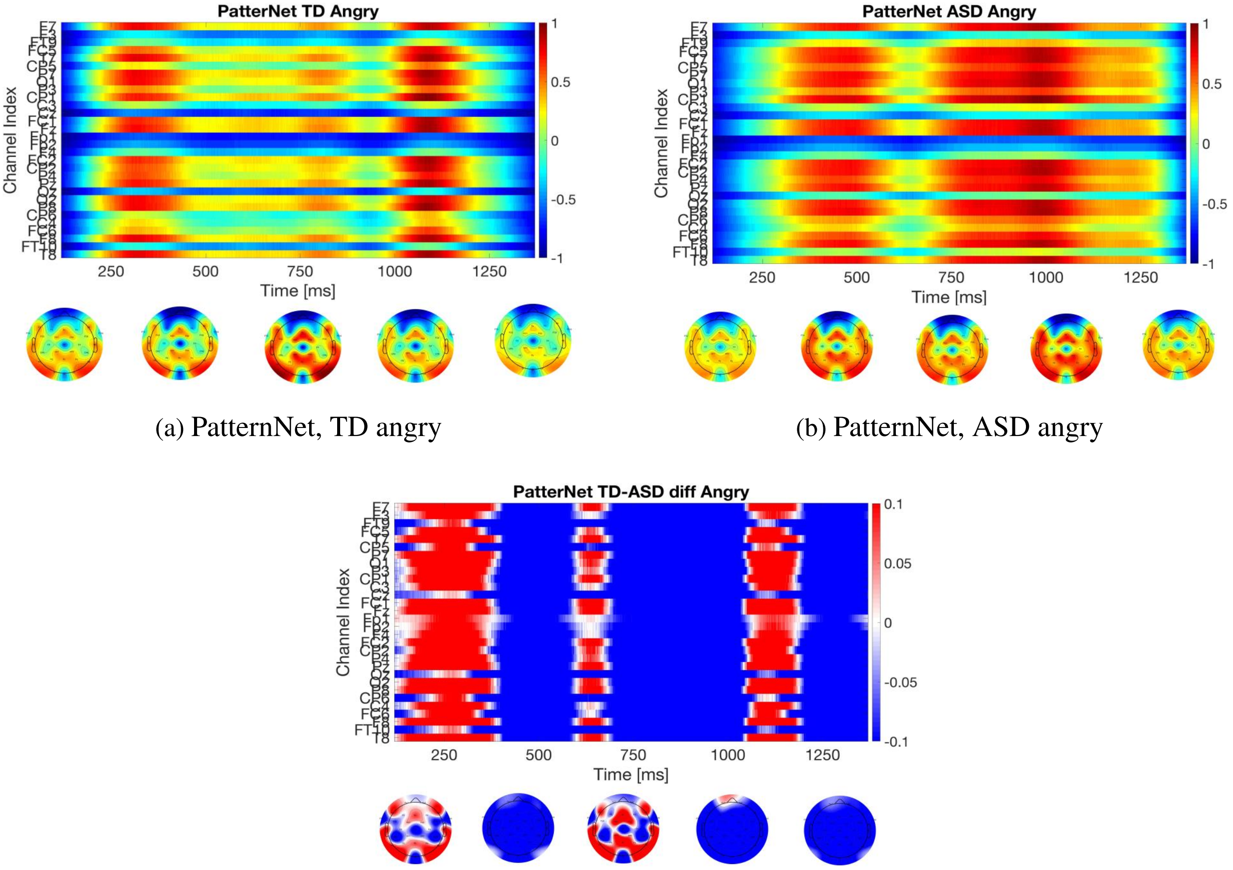


(a) PatternNet, TD sad

(b) PatternNet, ASD sad

(c) PatternNet, TD-ASD diff sad

Fig. 7.14 PatternNet Sad relevance-map for TD 7.14a, and ASD 7.14b, and the differences between TD-ASD 7.14c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

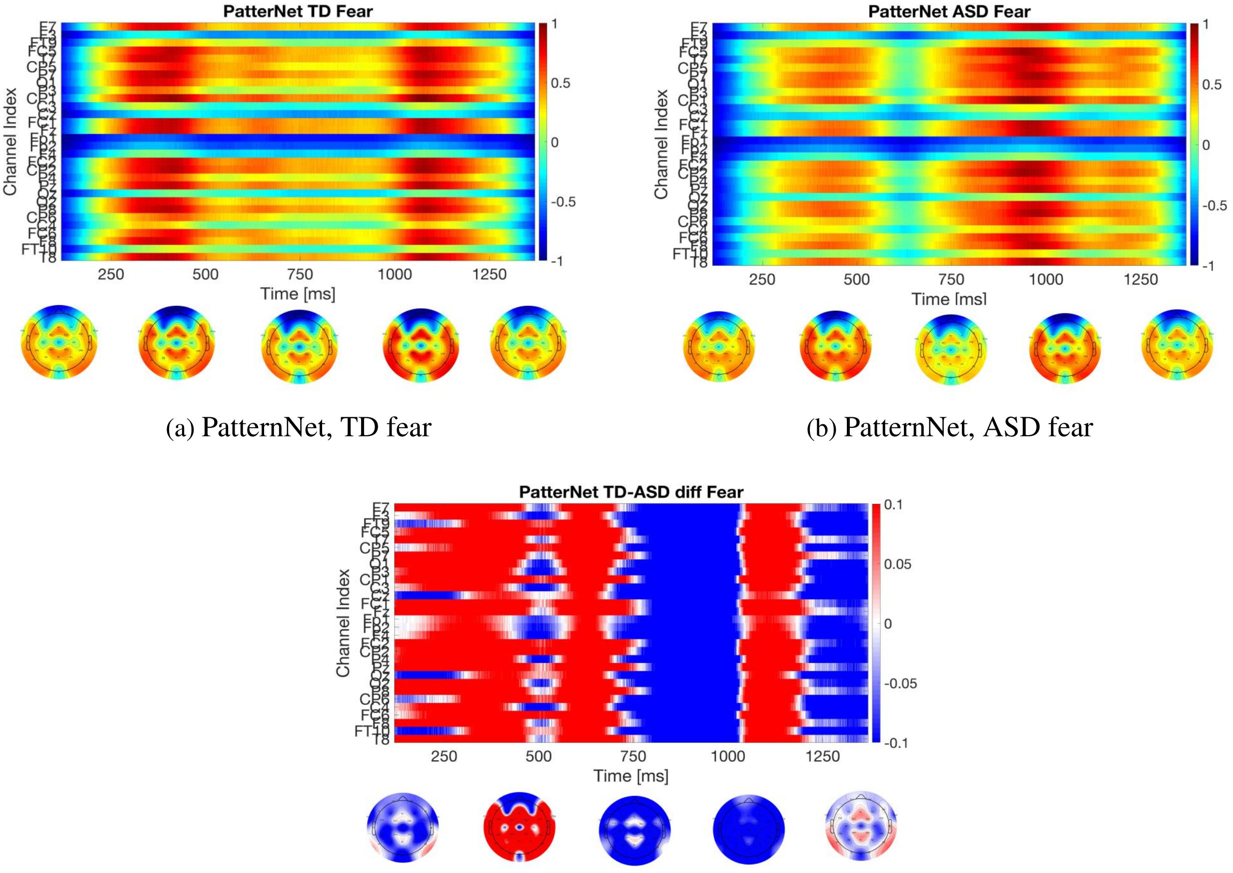


(a) PatterNet, TD angry

(b) PatterNet, ASD angry

(c) PatterNet, TD-ASD diff angry

Fig. 7.15 PatternNet Angry relevance-map for TD 7.15a, and ASD 7.15b, and the differences between TD-ASD 7.15c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

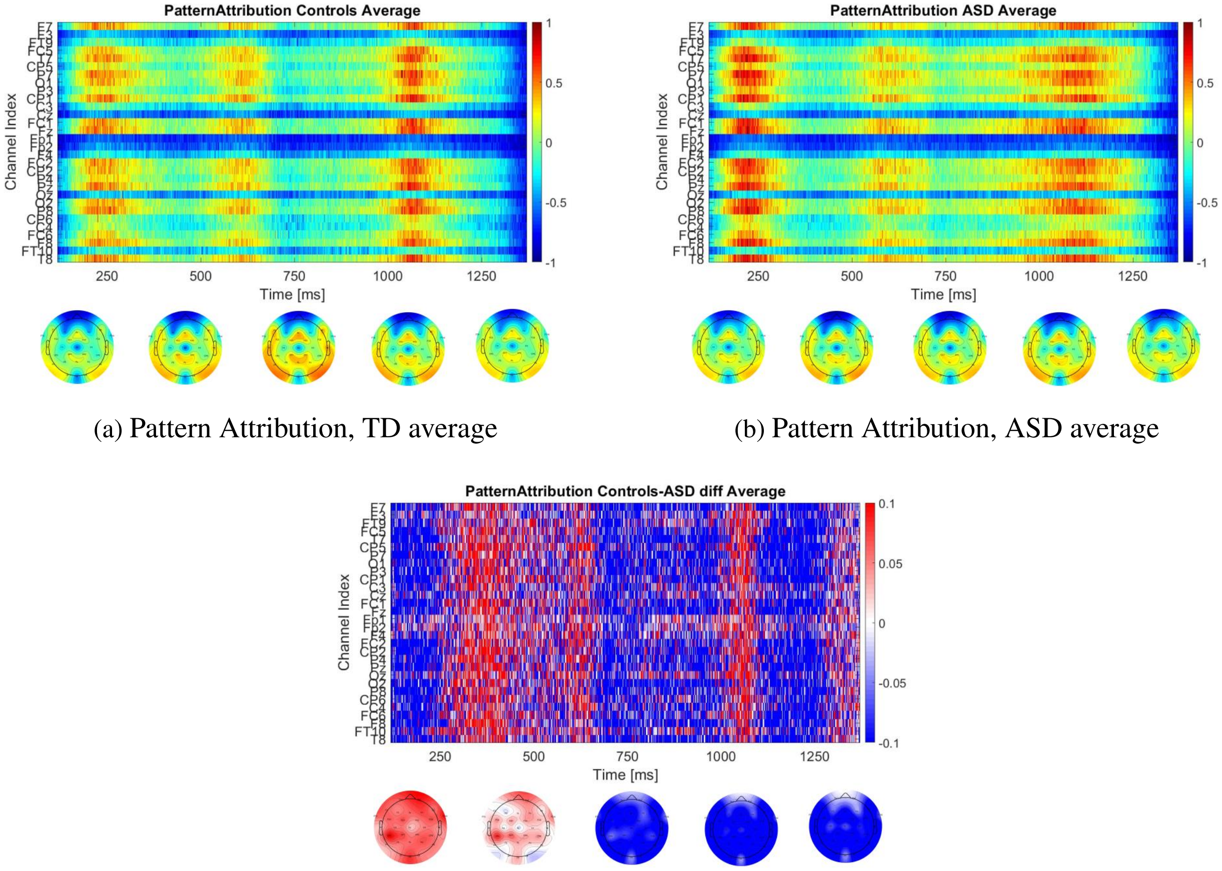


(a) PatternNet, TD fear

(b) PatternNet, ASD fear

(c) PatternNet, TD-ASD diff fear

Fig. 7.16 PatternNet Fear relevance-map for TD 7.16a, and ASD 7.16b, and the differences between TD-ASD 7.16c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

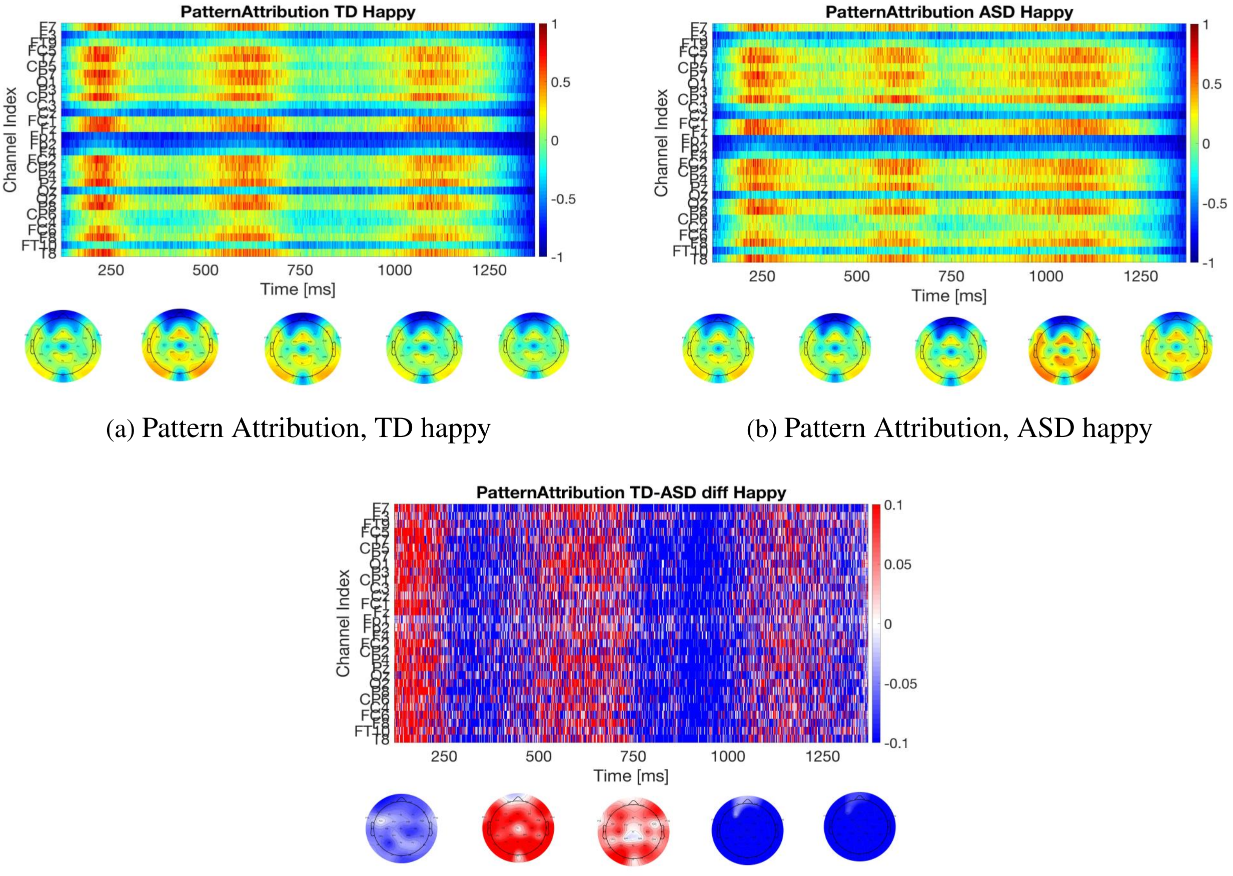


(a) Pattern Attribution, TD average

(b) Pattern Attribution, ASD average

(c) Pattern Attribution, TD-ASD diff average

Fig. 7.17 Pattern Attribution average class relevance-map for TD 7.17a, and ASD 7.17b, and the differences between TD-ASD 7.17c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

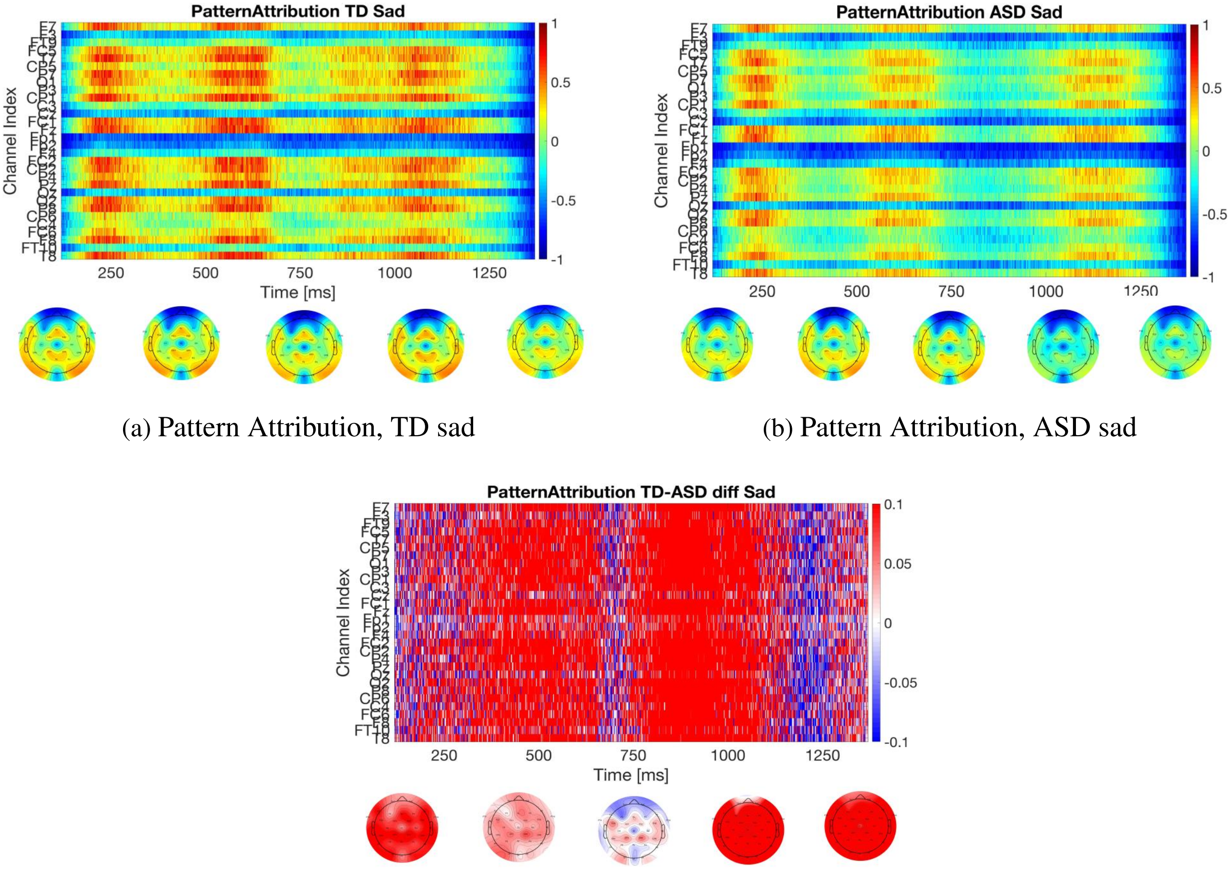


(a) Pattern Attribution, TD happy

(b) Pattern Attribution, ASD happy

(c) Pattern Attribution, TD-ASD diff happy

Fig. 7.18 Pattern Attribution Happy relevance-map for TD 7.18a, and ASD 7.18b, and the differences between TD-ASD 7.18c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

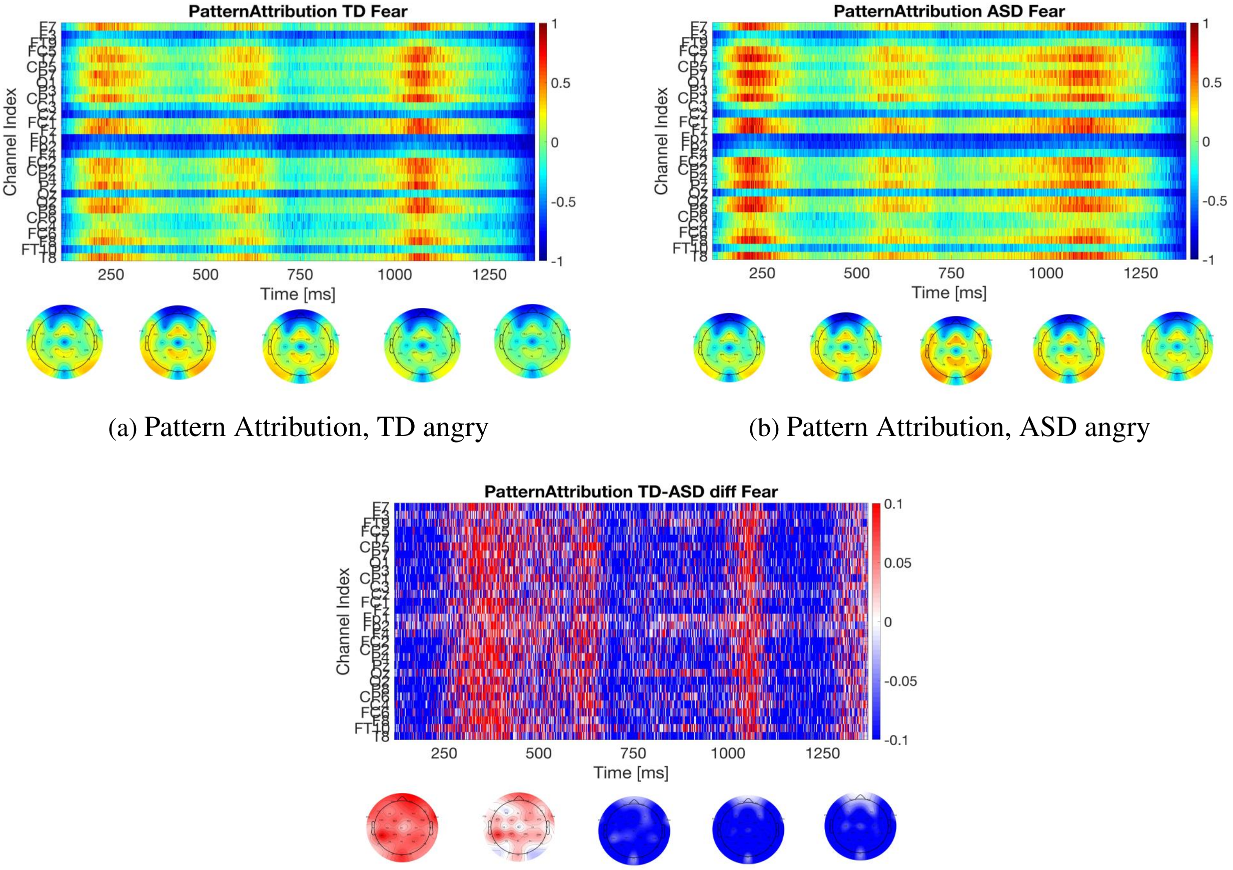


(a) Pattern Attribution, TD sad

(b) Pattern Attribution, ASD sad

(c) Pattern Attribution, TD-ASD diff sad

Fig. 7.19 Pattern Attribution Sad relevance-map for TD 7.19a, and ASD 7.19b, and the differences between TD-ASD 7.19c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

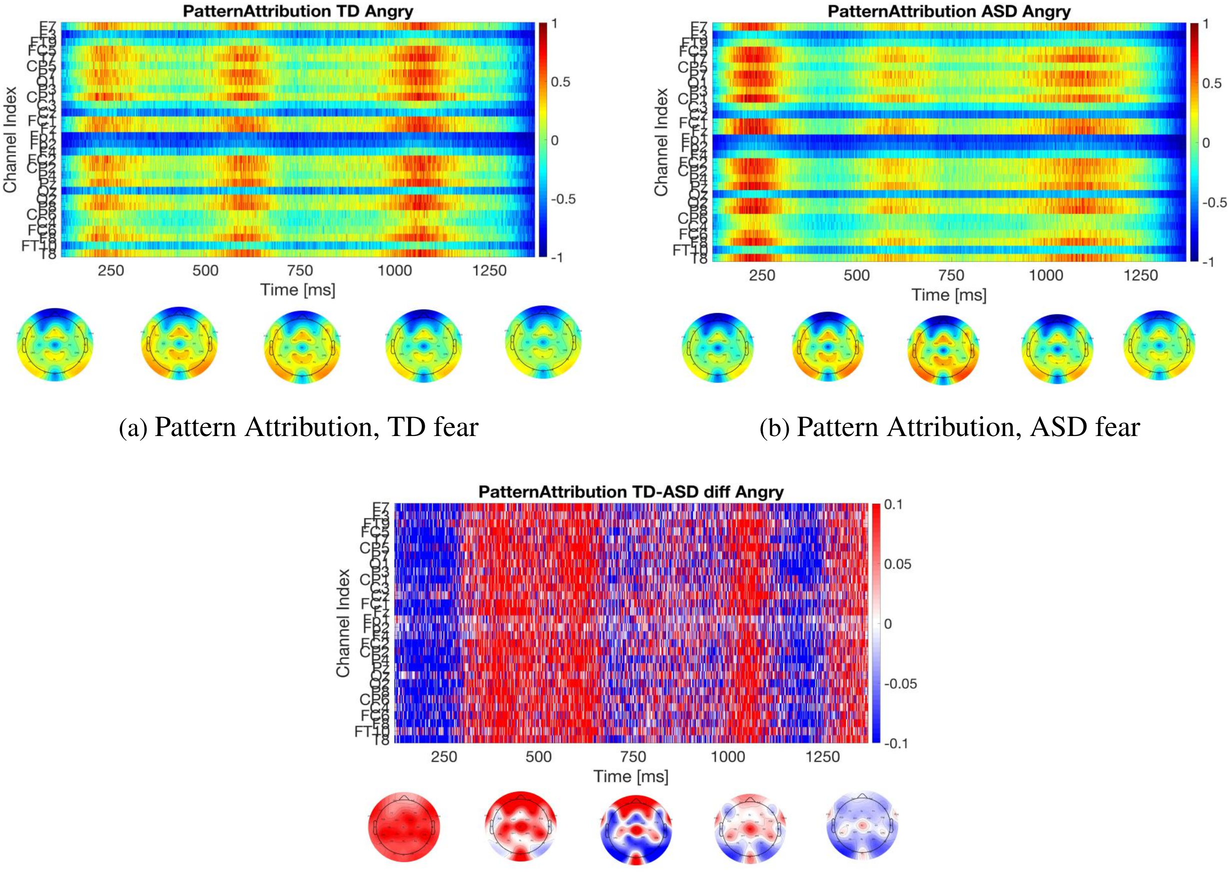


(a) Pattern Attribution, TD angry

(b) Pattern Attribution, ASD angry

(c) Pattern Attribution, TD-ASD diff angry

Fig. 7.20 Pattern Attribution Angry relevance-map for TD 7.20a, and ASD 7.20b, and the differences between TD-ASD 7.20c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.



(a) Pattern Attribution, TD fear

(b) Pattern Attribution, ASD fear

(c) Pattern Attribution, TD-ASD diff fear

Fig. 7.21 Pattern Attribution Fear relevance-map for TD 7.21a, and ASD 7.21b, and the differences between TD-ASD 7.21c. For the TD and ASD groups we use a jet colormap due to the relevance normalization between $[-1,1]$, and for the TD-ASD difference relevance-map we use the redblue colormap with 50 color scale between a range of $[-0.1,0.1]$.

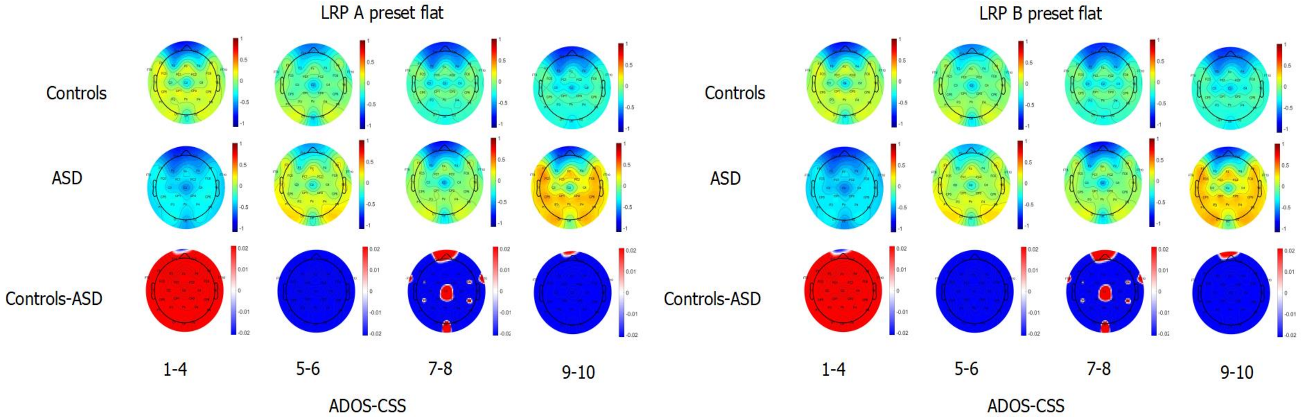
Now, in the Appendix B we report the rest of the results for the most important methods of the iNNvestigate package, and we will report the relationship between the ADOS-CS and the relevance-maps for the LRP A/B presets, PatternNet, and Pattern Attribution. We will also report the results for the ROAR methodology applied for Sample #1.

7.4 Topo-Maps - ADOS-CS

To evaluate the relationship between the relevance-maps and ADOS-CS scores we evaluate the linear regressions - similar to the linear models of Chapter 6- between the average value of the relevance-maps across all emotions (using the average class). Subsequently, we grouped p-values for each regression using a Bonferroni correction across the 5 time ranges

[0-500],[250-750],[500-1000],[750-1250], and [1000-1500] ms as we explained above. The results obtained after correction show no significance difference between TD and ASD groups. For all severity levels we grouped ADOS-CS values in this experiment such as low-severity 1-4, low-mid severity 5-6, mid-severity 7-8, and high-severity 9-10 in order to make the topo-maps smoother. We set these ranges due to the topo-map per subject or for each ADOS-CS value between 1-10 is showing an increased noise and non-deterministic level.

All the p-values found after the regression were p>0.05, and R correlation values between $0.0045 \leq R \leq 0.0932$ for the LRP A preset, $0.0025 \leq R \leq 0.0435$ for LRP B, $0.0098 \leq R \leq 0.0775$ for PatternNet, and $0.0056 \leq R \leq 0.0888$ for Pattern Attribution method showing again an isolated model in from the relevance-maps calculated from the trained Deep ConvNet in comparison with behavioral outcome measures such as ADOS severity measures.

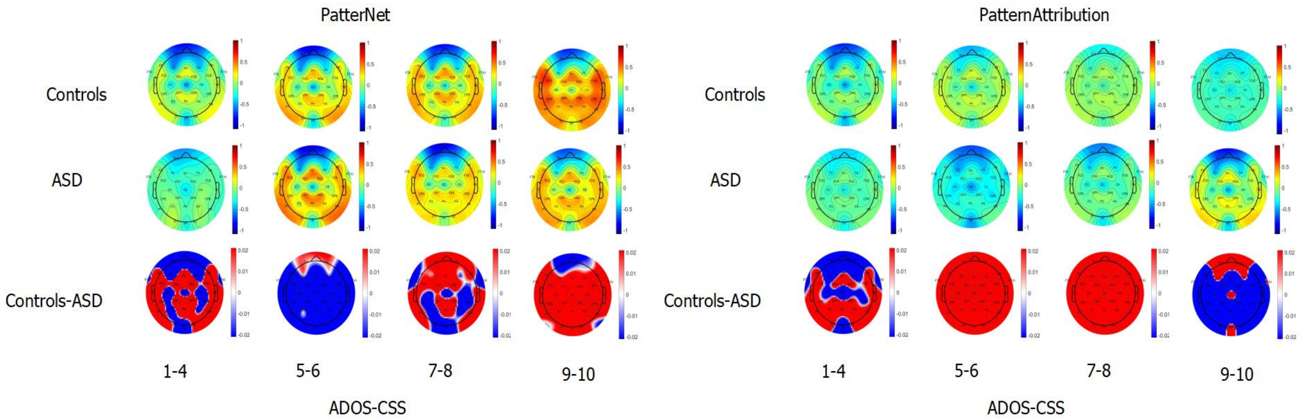


(a) LRP A preset v.s ADOS-CS

(b) LRP B preset v.s ADOS-CS

Fig. 7.22 Relevance topo-maps across the ADOS-CS spectrum in four groups low-severity 1-4, low-mid severity 5-6, mid-severity 7-8, and high-severity 9-10 for methods LRP A preset (Figure 7.22a), and LRP B preset (Figure 7.22b). TD and ASD groups are analyzed here in the first two rows. Colorbars are normalized between $[0,1]$ for TD and ASD topo-maps using the jet colormap, and the TD-ASD difference colorbar is normalized between $[-0.02, 0.02]$ using the redblue colormap.

As we can see also in Figures 7.22 and 7.23 the difference topo-map in the row three is showing minimum differences between $[-0.02, 0.02]$ supporting the non-significand regressions for LRP-based and estimator propagation methods.



(a) PatternNet v.s ADOS-CS

(b) Pattern Attribution v.s ADOS-CS

Fig. 7.23 Relevance topo-maps across the ADOS-CS spectrum in four groups low-severity 1-4, low-mid severity 5-6, mid-severity 7-8, and high-severity 9-10 for methods PatternNet (Figure 7.23a), and Pattern Attribution (Figure 7.23b). TD and ASD groups are analyzed here in the first two rows. Colorbars are normalized between $[0,1]$ for TD and ASD topo-maps using the jet colormap, and the TD-ASD difference colorbar is normalized between $[-0.02, 0.02]$ using the redblue colormap.

These results suggest that for low-severity ADOS-CS indicator TD>ASD, for low-mid severity we found that TD<ASD for the overall scalp, for mid severity we found TD<ASD for almost all the scalp, and for high severity we found TD<ASD for all the methods LRP A/B, and PatternNet, and Pattern Attribution methods. However, even if the relevance is going lower for a larger ADOS-CS values the correlation for each method is not enough negative for being significant. For Pattern Attribution method the previous relationship effect between the relevances and the ADOS-CS is not that evident due to the increased noise added to the Pattern Attribution model on ω^T.

7.5 RemOve And Retrain (ROAR) evaluation

Using the Sample #1 data we implemented ROAR (Hooker et al., 2018) methodology to measure the level of reliability and interpretability for our 3-layer Deep ConvNet trained to decode the four emotions explained above.

Following ROAR we use an initial training process based on (Torres et al., 2018, 2019) to calculate an initial relevance-map R^1_{ij} per trial from the most reliable methods of the iNNvestigate package such as SmoothGrad, SmoothGrad-Squared, PatternNet, Pattern-Attribution, and LRP presets A/B as we explained above (Alber et al., 2019).

As we mentioned in the section above the resulting relevance-map per trial is averaged for each participant, and finally an averaged relevance-map is obtained for each class Happy, Sad, Angry, and Fear: R_{happy}, R_{sad}, R_{angry}, and R_fear following Equation 7.12 for each participant averaging across the total 48 trials.

$$R_{av} = \frac{1}{48N}\sum_{i}^{N}\sum_{j=0}^{47} R_{ij} \tag{7.12}$$

A single and resulting relevance-map R_s is used for TD and ASD groups based on Equation 7.13 for each method mentioned above. To compute the removal part of the ROAR method we transform the resulting relevance-map R_s into a binarized-mask map R_b sorting the relevance values from high to low and removing the percentage of pixels which are considered important for a particular saliency method.

Therefore, in order to compare ROAR on our proposed EEG feature-set we compute a baseline based on a selection of thin slices testing the level of statistical correlation between the EEG channels (Krishna, Pasha, and Savithri, 2016). The thin slices construction consists in thin occluders with the size of an EEG channel, and 47 time points denoted here as a 47×1 slice.

Based on thin slices we select 47×1 slices of pixels indexes using a random uniform distribution of pixels in the mask denoting this as *random slices baseline* for each saliency method, or a method-based saliency slices baseline. We construct a second baseline based on the same 47×1 slices but sorted using relevance values for each saliency method denoting it as *slices* preffix complemented with the method name.

$$R_s = (R_{happy} + R_{sad} + R_{angry} + R_{fear})/4 \tag{7.13}$$

With the R_b sorted indexes we set a pixel rate removal r between 0 and 1 created for remove the pixels considered important by the corresponding saliency method, and the corresponding baseline relevance-maps. The pixel/feature removal follows Equation 7.14 obtaining a complete set of performance metrics for different r points such as 0.1, 0.2, 0.5, 0.7, 0.9 and 1 joining the points in the final plot. R_b is multiplied point by point to the input image to regulate the feature acceptance on the re-training.

$$R_b = \begin{cases} 1 & R_s \leq r \\ 0 & R_s > r \end{cases} \tag{7.14}$$

A summarized pipeline for ROAR methodology and our application on EEG data is shown in Figure 7.24.

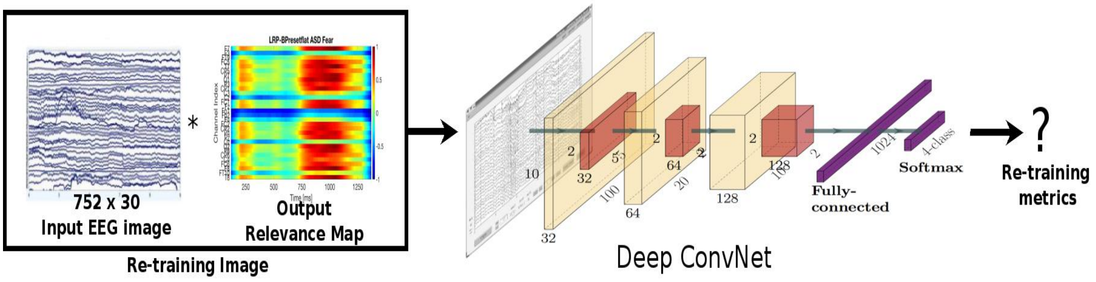


Fig. 7.24 RemOve And Retrain(ROAR) methodology pipeline. The pipeline sue a re-trained new input feature-set modulated by an averaged relevance-map calculated after the LOTO cross-validation. Using a binary mask we removed the features corresponding relevant channel and time point with the purpose of calculate new metrics using the original EEG image× binary mask per saliency method.

Using the LRP B preset method as baseline we show some binary masks R_b examples on Figure 7.25. These Figures show how the pixel rate removal value r change the binary-masks pattern diminishing the density of red points *"relevant"* in the R_b relevance-map having a complete feature removal when $r = 1$. These R_b or binary mask are shown only with the heat-map using the redblue colormap being the blue points 0 and the red points 1.

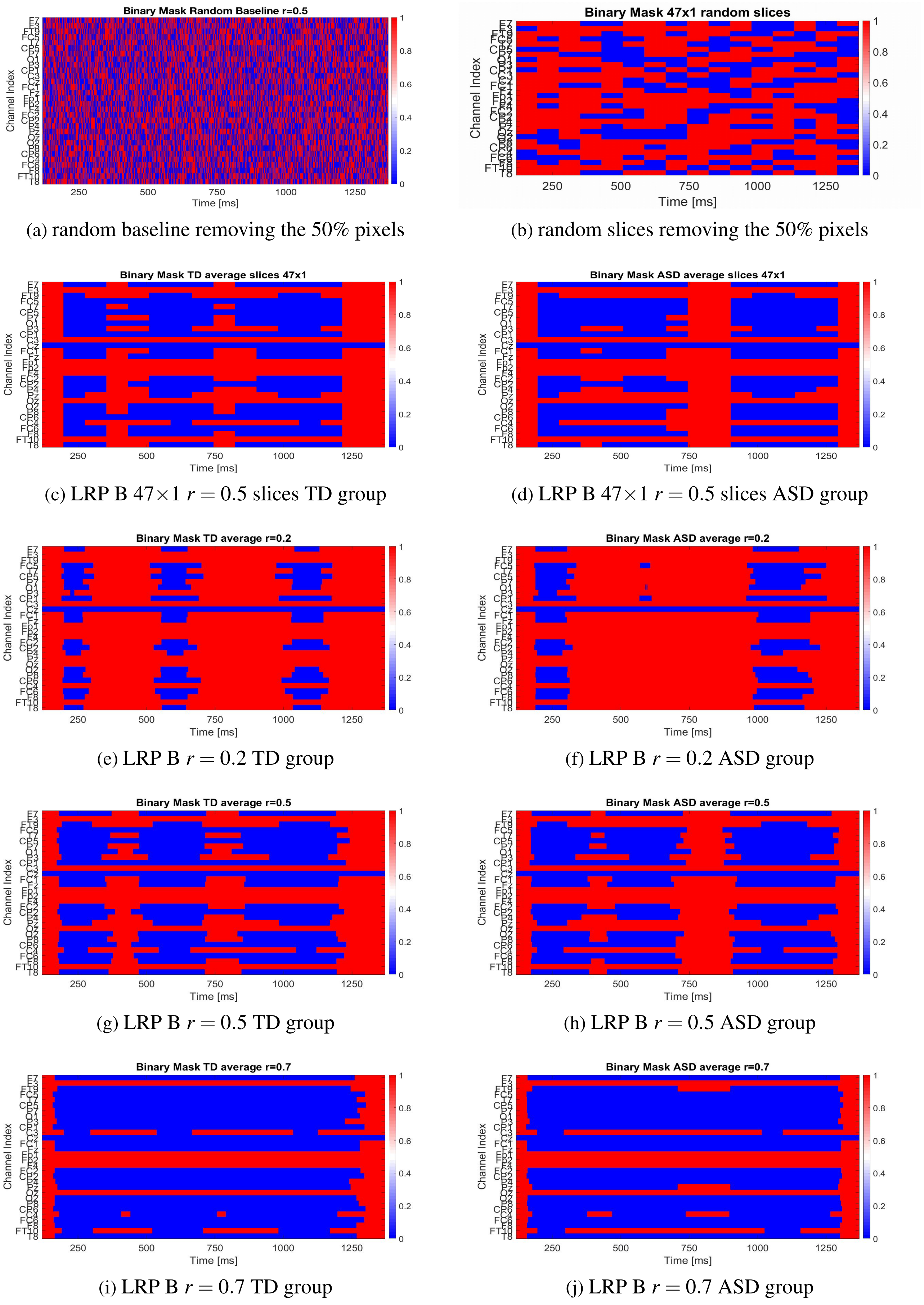
We can observe that changing the incidence of the relevance linearly the channel Cz shows the higher relevance between TD and ASD groups across the relevance-maps, and there is three *"relevant"* time spots in the TD group around 250, 550, and 1200ms after the onset. For the ASD group we observe two important spots on 250, and 1200ms being more pronounced in the late components as we observed in the ANOVA comparisons on the section above.

The first way we use to illustrate ROAR differences between the baselines and the saliency methods is using the barplots comparison of Figure 7.26. We select again LRP B preset as the most reliable method given our dataset on Sample #1 finding significant differences between the accuracies obtained from the random baselines, slices based on the method, and the method itself $p < 0.001$. The p-values for this comparison were reported after Bonferroni correction.

The comparison after corrections were: comparing TD and ASD FER performances we obtained the same results reported in Chapter 5 F(1,87)=4.69, with a greater p-value after correction p=0.0455.

The rest comparisons are pairing ASD groups performances only across the different modalities of Figure 7.26 such comparing FER and Deep ConvNet performances in ASD F(1,79)=12.78, p=0.000345, comparing ASD random baseline with LRP B slices F(1,79)=13.77, p=0.000112, comparing the ASD 47×1 random baseline with the LRP B slices F(1,79)=6.58, p=0.0113, comparing the random baseline with the LRP B $r = 0.5$ F(1,79)=20.34, p=2.45E-6, comparing the 47×1 slices performance with the LRP B $r = 0.2$ F(1,79)=17.88, p=0.0000345 having an opposite sign being the LRP B $r = 0.2$ performance higher than the 47×1 slices, comparing the same 47×1 slices accuracy with the LRP B $r = 0.7$ F(1,79)=34.89, p=1.67E-10.

A similar difference is observed between the random baseline and the LRP B $r = 0.5$ F(1,79)=9.99, p=0.000367, and the random baseline again in comparison the LRP B $r = 0.7$ F(1,79)=15.90,p=4.45E-8. These similar difference are observed for the other robust methods



(a) random baseline removing the 50% pixels

(b) random slices removing the 50% pixels

(c) LRP B 47×1 $r = 0.5$ slices TD group

(d) LRP B 47×1 $r = 0.5$ slices ASD group

(e) LRP B $r = 0.2$ TD group

(f) LRP B $r = 0.2$ ASD group

(g) LRP B $r = 0.5$ TD group

(h) LRP B $r = 0.5$ ASD group

(i) LRP B $r = 0.7$ TD group

(j) LRP B $r = 0.7$ ASD group

Fig. 7.25 Examples of binary-maps for random 47×1 slices baseline, and saliency method-based 47×1 slices baseline, and the saliency method-based relevance for the LRP-B preset

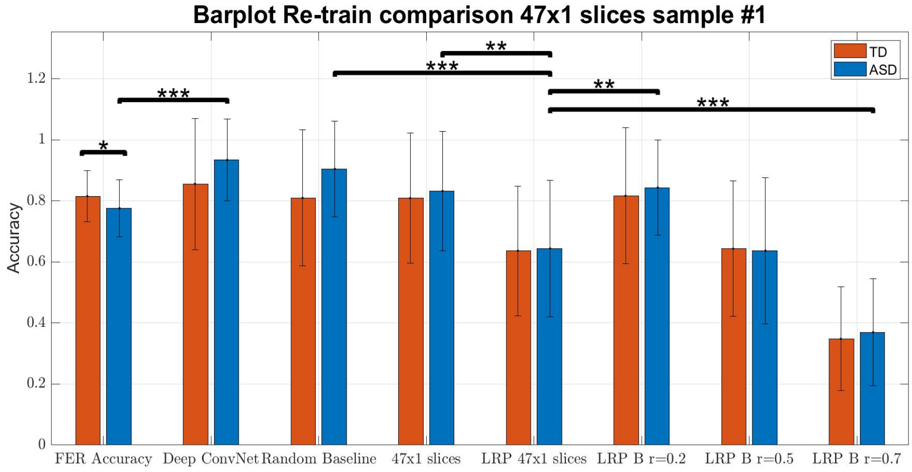


Fig. 7.26 LRP B preset barplots accuracies comparison between FER human accuracy, Deep ConvNet baseline, random baseline, 47×1 random slices, LRP B based slices for $r = 0.5$, and all the corresponding LRP B ROAR patterns $r = 0.2$, $r = 0.5$, and $r = 0.7$ shown in the x-axis. The values comparing bar accuracies with *** are significantly different $p < 0.0001$, and with ** $p < 0.001$

7.5.1 ROAR comparison TD/ASD

The second way to illustrate the accuracy changes using ROAR is observed in Figures. In these plots we show the accuracy in y-axis, and r or pixel removal rate in x-axis for SmoothGrad-Squared (go to Appendix B), PatternNet, Pattern-Attribution, and LRP B preset methods.

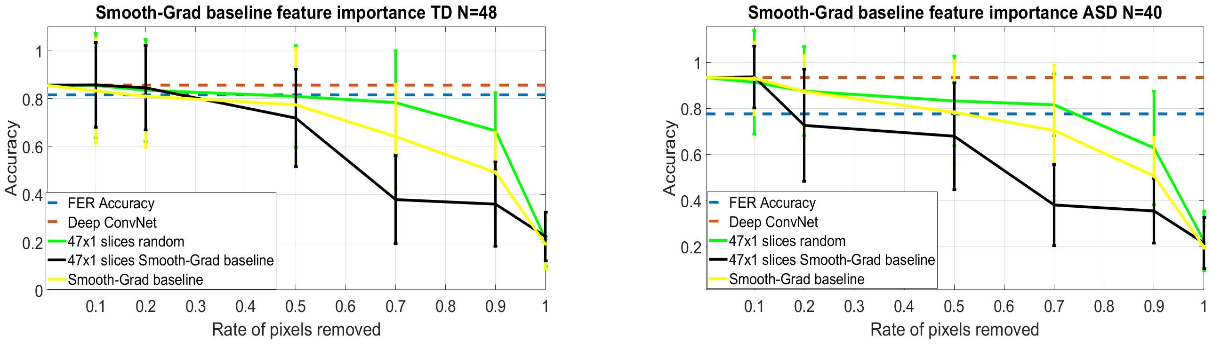
These results evaluate the current most reliable saliency maps on real EEG data and between two clinical groups such as TD and ASD. Being concordant with previous ML studies predicting a more quantifiable saliency methods reliability in comparison with random, or correlation baselines created the same relevance-maps.

To set a baseline for this comparison we use the plain Smooth-Grad method from iNNvestigate package (Appendix B) where we expect a more noisy relevance-map. We expect that the accuracy decreasing applying ROAR to this plain Smooth-Grad method won't be significant or the sign will be different in comparison with the slices and random baselines in comparison with the other methods.

In Figures 7.27a and 7.27b we show the variation across r values between 0.1 and 1 evaluating points in $r = 0.1$, $r = 0.2$, $r = 0.5$, $r = 0.7$, $r = 0.9$, and $r = 1$ for TD and ASD groups respectively. Evaluation baselines plotted in the black (slices baselines), and in green (random baseline) for the TD and ASD method the yellow line is in the middle of both baselines being significantly different in comparison with the random baseline for $r = 0.7$ F(1,95)=3.31,p=0.00155 in TD, and $r = 0.9$ F(1,95)=2.65,p=0.0224 in TD too, and for

$r = 0.7$ F(1,79)=4.01,p=0.000331 in ASD, and $r = 0.9$ F(1,79)=2.23,p<0.0338 in ASD too, and in comparison with the slices 47×1 for $r = 0.7$ F(1,95)=10.58,p=1.45e-6 in TD, and $r = 0.9$ F(1,95)=7.33,p=0.00023 in TD too, and for $r = 0.7$ F(1,79)=11.11,p=2.28E-7 in ASD, and $r = 0.9$ F(1,79)=7.45,p<0.000148 in ASD too.

In Figure 7.27 the dashed red line show the FER accuracy quota for the Sample #1 participant quantification, and in blue the Deep ConvNet quota obtained using the entire feature-set without pixels removing. These latter values are reported in Tables 5.2 and 5.3.



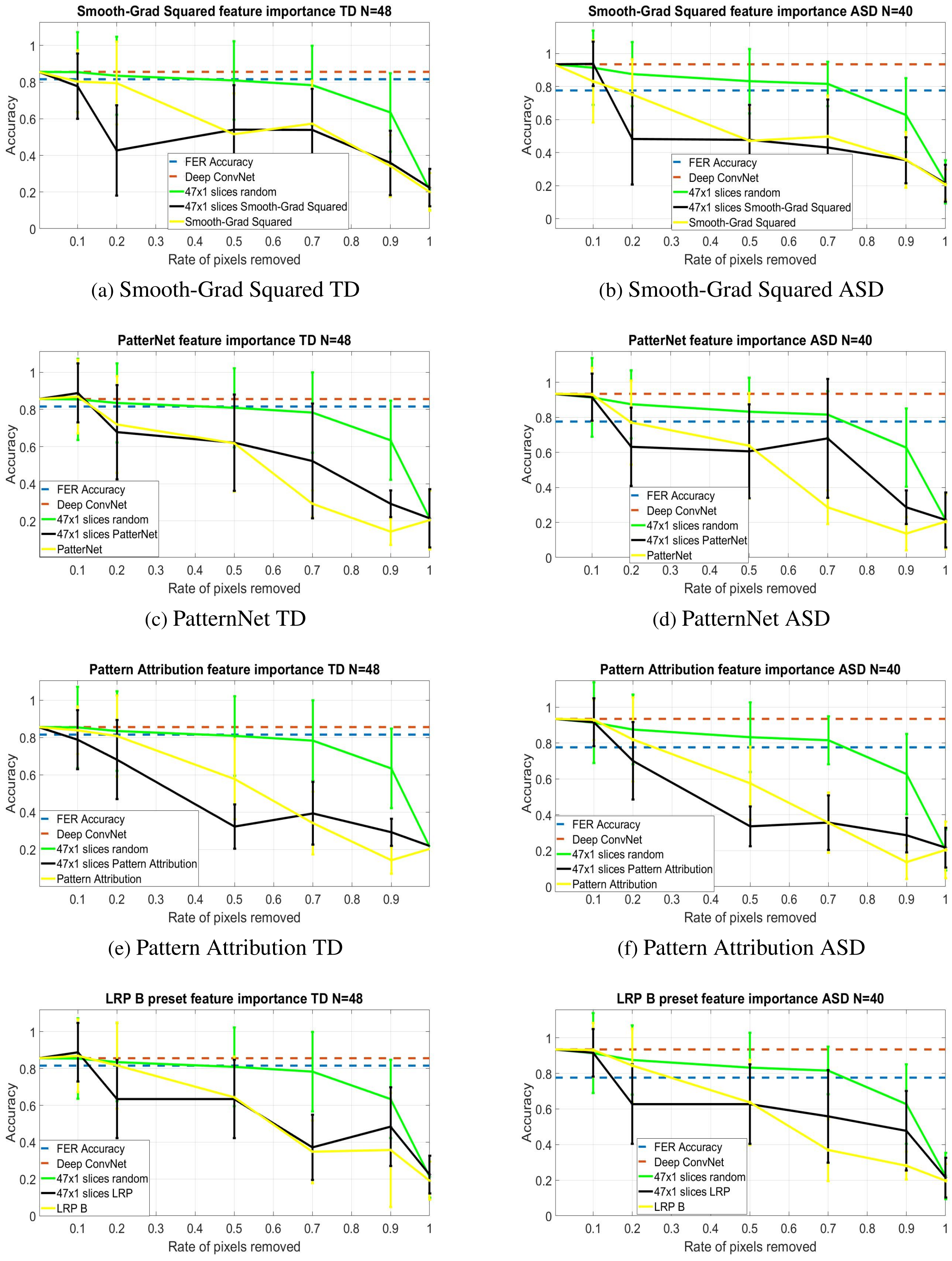
(a) Smooth-Grad baseline TD

(b) Smooth-Grad baseline ASD

Fig. 7.27 Accuracies comparison between the 47×1 random baseline slices, 47×1 slices weighting for the Smooth-Grad baseline and TD/ASD groups.

To evaluate ROAR in the most robust methods Figure 7.28 is showing a common decrease accuracy for the saliency method itself in terms of the percentage of pixels removed from input set. We found signifcant differences comparing the accuracy decreasing observed by SmoothGrad-Squared, PatternNet, Pattern-Attribution, and LRP B preset methods with the decreasing obtained from random and LRP 47×1 baselines with the method performance decreasing after $r = 0.5$ being the yellow line always down in terms of accuracies comparing it with baselines (black line, green line).

However, we found the significant differences across multiple different r values across the different robust saliency methods proposed here. We can observe an interesting performance pattern for some methods where the slices baseline is decreasing more accuracy than the method itself in a significant way. This effect is attributed to single channel correlations propagated across the trained Deep ConvNet (Chandaka, Chatterjee, and Munshi, 2009).



(a) Smooth-Grad Squared TD

(b) Smooth-Grad Squared ASD

(c) PatternNet TD

(d) PatternNet ASD

(e) Pattern Attribution TD

(f) Pattern Attribution ASD

(g) LRP B preset TD

(h) LRP B preset ASD

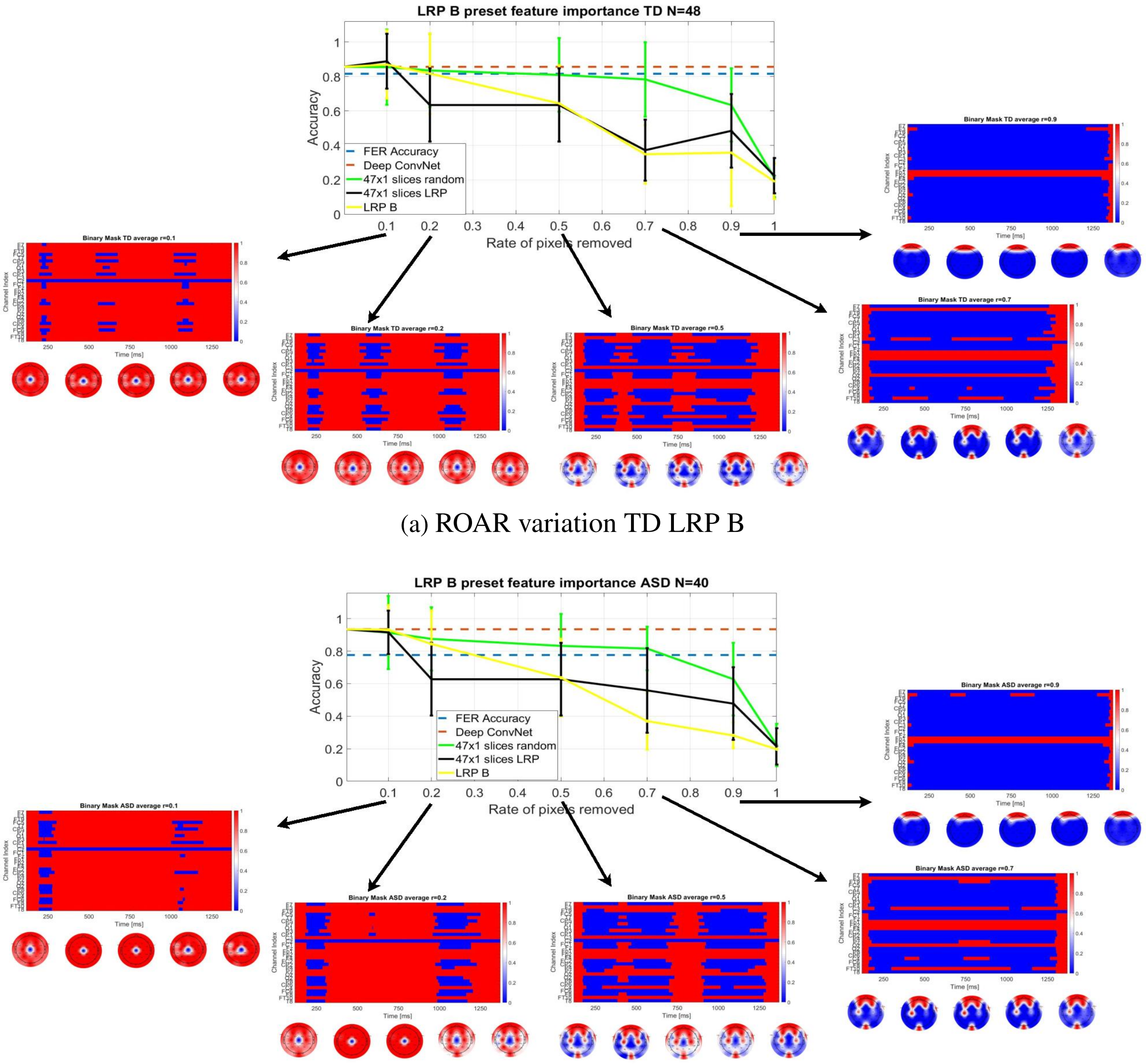
Fig. 7.28 Accuracies comparison results between the 47×1 random baseline slices, 47×1 slices weighting for the Smooth-Grad baseline and TD/ASD groups. Figures 7.28a and 7.28b, 7.28c and 7.28d, 7.28e and 7.28f, and 7.28g and 7.28h show the plots for ROAR performance evaluation for Smooth-Grad Squared, PatternNet, PatternAttribution, and LRP B preset respectively.

We found significant performance differences where the random slices baseline is showing a profuse accuracy decreasing for $r = 0.2$ for instance for Smooth-Grad Squared. For TD group we found F(1,95)=10.99, p=0.0000274, and for ASD group F(1,79)=12.45, p=0.0000345 being the slice baseline always lower than the method itself. However, we did not find significant differences for TD group F(1,95)<2.58, p>0.1893, and ASD group F(1,79)<2.88, p>0.1910 for all the values of r different than 0.2. Even without significant differences between the slices baseline and the method itself the performance decrease following the same patter of the slices baseline and lower than the random baseline as we expected.

On the other hand, for PatternNet method we found only significant differences between the slices baseline and the method itself in $r = 0.7$ where the method shows a lower accuracy for TD F(1,95)=13.38, p=0.0001178, and for ASD F(1,79)=20.45, p=2.77E-7, and for $r = 0.9$ TD F(1,95)=3.42, p=0.0267, and ASD F(1,79)=2.99, p=0.0321. For the rest r values we did not find significant differences F(1,95)<1.56, p>0.1034 for TD groups, and for ASD group F(1,79)<3.26, p>0.0551, between the slices baseline and the method accuracies being the method accuracies always lower than the baseline for $r \geq 0.5$.

As for the Pattern Attribution method we found significant differences for $r = 0.2$,F(1,95)=8.35, p=0.00327 for TD group, and for ASD group F(1,79)=7.91, p=0.00899, again associating the persistence presence of the channels correlations with the propagation of ω^T for the estimator optimization described in the sections above. We found another significant difference where the slice baseline is showing a lower performance than the method itself for $r = 0.5$, for TD group F(1,95)=10.12, p=0.000224, and for ASD group F(1,79)=9.88, p=0.003367. For the rest of r values we did not found any significant difference having F(1,95)<1.78,p>0.1256 for TD group, and F(1,79)<1.99,p>0.2212 for ASD group.

Evaluating the LRP B preset method we found again significant differences for $r = 0.2$ supporting again the single channel correlation per trial as we mentioned above. The difference here is as well as in previous methods lower for the slices baseline in comparison with the method itself. We selected LRP B preset instead of LRP A preset because more significance difference on the topo-maps analysis above. For TD group in $r = 0.2$ we have F(1,95)=3.56, p=0.00214, and for ASD group F(1,79)=3.66, p=0.00203. The LRP B accuracy decreasing has a different pattern in comparison with the other methods and we only observe a significant difference for the ASD group in $r = 0.7$ F(1,79)=3.91, p=0.000156. For other r values we did not find any significant difference between the slices baseline and the saliency method for the rest of r values and for TD F(1,95)<2.88, p>0.1036, and for ASD F(1,79)<1.88, p>0.265.



(a) ROAR variation TD LRP B

(b) ROAR variation ASD LRP B

Fig. 7.29 ROAR removal-rate r variation is shown in both plots here having the plotted lines in the upper part, and with arrows we are pointing the binary mask variation depending on the r values on the x-axis. Figure 7.29a and 7.29b shows the corresponding variation summary for the TD and ASD groups of Sample #1

As a third option to show the ROAR method dynamics applied to the neural data from Sample #1 for successful emotion decoding we use the same plot from Figures 7.28g and 7.28h and we attach the binary mask variation on the bottom. We can see these new plots in Figures 7.29a and 7.29a. In these plots we can see a significant different averaged binary-mask pattern for LRP B preset especially for $r = 0.7$, and $r = 0.9$ where the three important timing spots at 250, 550, and 1200ms are more evident in TD groups, and two timing spots at 250 and 1200ms after stimulus onset, measuring it in the last topo-maps on [750-1250]ms, and [1000-1500]ms, **F(1,87)>5.57,p<0.0466** after correction.

Chapter 8

Conclusions

In this dissertation we evaluate multiple ML pipelines for EEG-based emotion decoding based on DEAP dataset, semantic classification using EEG signals, and HR and IBI signal estimation using PPG sensors on realistic environments. These evaluations help us to identify important pipeline parameters and signal treatment techniques to support robust class decoding using biosignals as a first objective of this dissertation.

From the initial evaluation we can support the possibility of a successful emotion decoding from Controls and Autism individuals neural activity using our proposed emotion decoding pipeline based on 3 conv-pool blocks Deep ConvNet.

For all the samples described in this study including three participant samples from 3 different age ranges the Deep ConvNet's accuracies and other related metrics such as Precision and Recall outperform FER human corresponding metrics **We obtained a successful discrimination and generalization of the neural activity for multiple emotion decoding using neural activity.**

The statistical disentanglement observed between Deep ConvNet accuracies, FER accuracies, and the ADOS-CS scores suggest a diverse and isolated numerical representation from the trained Deep ConvNet correlated not only with the classifier, but with the entire pipeline. The conjunction of Prep+ADJUST automatic artifact removal, and the ZCA whitening normalization increase the separability of emotion classes. The Deep ConvNet metrics are providing information for a completely different numerical model defined by our proposed pipeline in comparison with the behavorial models obtained from the ASD groups.

The feature importance results suggest that each trial correctly decoded by the pipeline shows activation patterns attributing high importance to almost the complete channel array, and some particular time spots differentiating between TD and ASD groups. **These results can re-define the current state-of-the-art emotion decoding pipeline supporting our Deep**

ConvNet classifier as a correct and perceptual pipeline being able to overcome the behavioral and neural emotion appraisal deficits observed in ASD groups.

Along this dissertation we can see FER human accuracy is negatively correlated with the ADOS-CS scores across the three samples. This is expected as a golden rule, and the saliency maps used for feature-importance representation are an initial intuition to constitute how neural activity is affecting the classification rates.

The saliency methods used in this dissertation are sensitive to input perturbations. Different initial image focal points, and different Deep ConvNet activation functions affect current gradient-based saliency maps. Therefore, to overcome these limitations is necessary a further exploration to include in other multiple clinical trials. Nevertheless, multiple clinical approaches have used LRP saliency maps for motor imagery, error potentials, and sleep stages classification (Andreotti, Phan, and De Vos, 2018; Palazzo et al., 2018; Sturm et al., 2016; Torres and Stepanov, 2017; Torres et al., 2018).

This study can be considered the first in analyzing emotion decoding sensitivity using saliency maps for clinical trials including Control and Autism individuals. **The usage of these saliency maps opens a new path for the implementation of more detailed saliency methods in the future, and a broad understanding of the current saliency maps applied to EEG-based emotion decoding.**

After evaluating the performances, the statistical correlation between machine and human parameters, and the feature-importance results we can support the usage of Deep ConvNet classifiers with an adequate artifact rejection and numerical normalization to successfully decode emotion using neural activity. The training process is transparent between TD and ASD groups showing better results comparing Deep ConvNet performances with FER human performances, and showing time spot early/late differences between TD and ASD saliency-maps.

The statistical disentanglement found between Deep ConvNet performances can be related with an abrupt numerical differences included by our proposed emotion decoding pipeline, and a no statistical relationship found between the resulting saliency maps and ADOS-CS spectrum showing no statistical correlation too.

This pipeline can be considered a strong candidate for ASD assisted intervention, for behavioral clinical measures, and online emotion decoding for ASD and TD groups indistinguishably. A good size training dataset for emotion decoding can be used to implement multiple online EEG-based emotion decoding experiments in the future.

This classifier can be personalized easily evaluating the performances with a LOTO cross-validation per subject modality including a training and test set for each participant. The Deep ConvNet can be also trained including the frequency domain being able to predict other outcome measures such as ASD early diagnosis, ADOS-CSS, and social skills outcome measures included in ASD clinical trials as a future work.

Bibliography

Abadi, Martín et al. (2016). “Tensorflow: a system for large-scale machine learning.” In: *OSDI*. Vol. 16, pp. 265–283.

Abadi, Mojtaba Khomami et al. (2015). “DECAF: MEG-based multimodal database for decoding affective physiological responses”. In: *IEEE Transactions on Affective Computing* 6.3, pp. 209–222.

Acharya, U Rajendra et al. (2011). “Application of recurrence quantification analysis for the automated identification of epileptic EEG signals”. In: *International journal of neural systems* 21.03, pp. 199–211.

Adams, Andra and Peter Robinson (2011). “An android head for social-emotional intervention for children with autism spectrum conditions”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 183–190.

Adolphs, Ralph, Lonnie Sears, and Joseph Piven (2001). “Abnormal processing of social information from faces in autism”. In: *Journal of cognitive neuroscience* 13.2, pp. 232–240.

Aftanas, Ljubomir I et al. (1997). “Non-linear analysis of emotion EEG: calculation of Kolmogorov entropy and the principal Lyapunov exponent”. In: *Neuroscience letters* 226.1, pp. 13–16.

Alber, Maximilian et al. (2019). “iNNvestigate neural networks!” In: *Journal of Machine Learning Research* 20.93, pp. 1–8.

Almeida, Pedro R et al. (2016). “Perceived arousal of facial expressions of emotion modulates the N170, regardless of emotional category: time domain and time–frequency dynamics”. In: *International Journal of Psychophysiology* 99, pp. 48–56.

Altman, Yair M (2014). *Accelerating MATLAB Performance: 1001 tips to speed up MATLAB programs*. Chapman and Hall/CRC.

Ameis, Stephanie H and Marco Catani (2015). “Altered white matter connectivity as a neural substrate for social impairment in Autism Spectrum Disorder”. In: *Cortex* 62, pp. 158–181.

Andreotti, Fernando, Huy Phan, and Maarten De Vos (2018). "Visualising convolutional neural network decisions in automatic sleep scoring". In: *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, pp. 70–81.

Andreotti, Fernando et al. (2018). "Multichannel sleep stage classification and transfer learning using convolutional neural networks". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 171–174.

Asthana, Stuti, Dinesh Goyal, and Amitkant Pandit (2017). "Analysis on Multiple Hidden Layer Complexity of BPNN". In: *International Journal of Applied Engineering Research* 12.14, pp. 4723–4728.

Athavale, Yashodhan and Sridhar Krishnan (2017). "Biosignal monitoring using wearables: Observations and opportunities". In: *Biomedical Signal Processing and Control* 38, pp. 22–33.

Bach, Sebastian et al. (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7, e0130140.

Bal, Elgiz et al. (2010). "Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state". In: *Journal of autism and developmental disorders* 40.3, pp. 358–370.

Banerjee, Buddhananda and Biswabrata Pradhan (2018). "Kolmogorov–Smirnov test for life test data with hybrid censoring". In: *Communications in Statistics-Theory and Methods* 47.11, pp. 2590–2604.

Barak, Boaz and Guoping Feng (2016). "Neurobiology of social behavior abnormalities in autism and Williams syndrome". In: *Nature neuroscience* 19.5, p. 647.

Baron-Cohen, Simon (2016). "Autism and the Empathizing–Systemizing (ES) theory". In: *Developmental social cognitive neuroscience*. Psychology Press, pp. 139–152.

Baron-Cohen, Simon et al. (1999). "Social intelligence in the normal and autistic brain: an fMRI study". In: *European journal of neuroscience* 11.6, pp. 1891–1898.

Baron-Cohen, Simon et al. (2001). "The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians". In: *Journal of autism and developmental disorders* 31.1, pp. 5–17.

Baveye, Yoann et al. (2017). "Affective video content analysis: A multidisciplinary insight". In: *IEEE Transactions on Affective Computing* 9.4, pp. 396–409.

Bayer, Mareike et al. (2019). "EEG-fMRI reveals rapid representation of personal relevance of faces in social cognition and reward-related brain regions". In: *bioRxiv*, p. 585133.

Beam, Andrew L and Isaac S Kohane (2018). “Big data and machine learning in health care”. In: *Jama* 319.13, pp. 1317–1318.

Bengio, Yoshua et al. (2009). “Learning deep architectures for AI”. In: *Foundations and trends® in Machine Learning* 2.1, pp. 1–127.

Benning, Stephen D et al. (2016). “Late positive potential ERP responses to social and nonsocial stimuli in youth with autism spectrum disorder”. In: *Journal of autism and developmental disorders* 46.9, pp. 3068–3077.

Berggren, Steve et al. (2018). “Emotion recognition training in autism spectrum disorder: A systematic review of challenges related to generalizability”. In: *Developmental neurorehabilitation* 21.3, pp. 141–154.

Bigdely-Shamlo, Nima et al. (2015). “The PREP pipeline: standardized preprocessing for large-scale EEG analysis”. In: *Frontiers in neuroinformatics* 9, p. 16.

Binder, Alexander et al. (2016). “Layer-wise relevance propagation for deep neural network architectures”. In: *Information Science and Applications (ICISA) 2016*. Springer, pp. 913–922.

Black, Melissa H et al. (2017). “Mechanisms of facial emotion recognition in autism spectrum disorders: insights from eye tracking and electroencephalography”. In: *Neuroscience & Biobehavioral Reviews* 80, pp. 488–515.

Bland, J Martin and Douglas G Altman (2002). “Validating scales and indexes”. In: *Bmj* 324.7337, pp. 606–607.

Blankertz, Benjamin et al. (2004). “The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials”. In: *IEEE transactions on biomedical engineering* 51.6, pp. 1044–1051.

Blankertz, Benjamin et al. (2011). “Single-trial analysis and classification of ERP components—a tutorial”. In: *NeuroImage* 56.2, pp. 814–825.

Bölte, Sven, Martin Holtmann, and Fritz Poustka (2008). “The Social Communication Questionnaire (SCQ) as a screener for autism spectrum disorders: additional evidence and cross-cultural validity.” In:

Bölte, Sven et al. (2006). “Facial affect recognition training in autism: can we animate the fusiform gyrus?” In: *Behavioral neuroscience* 120.1, p. 211.

Bos, Danny Oude et al. (2006). “EEG-based emotion recognition”. In: *The Influence of Visual and Auditory Stimuli* 56.3, pp. 1–17.

Bosl, William J, Helen Tager-Flusberg, and Charles A Nelson (2018). “EEG analytics for early detection of autism spectrum disorder: a data-driven approach”. In: *Scientific reports* 8.1, p. 6828.

Bosl, William et al. (2011). “EEG complexity as a biomarker for autism spectrum disorder risk”. In: *BMC medicine* 9.1, p. 18.

Brooks, Brian L, Elisabeth MS Sherman, and Esther Strauss (2009). “NEPSY-II: A developmental neuropsychological assessment”. In: *Child Neuropsychology* 16.1, pp. 80–101.

Bzdok, Danilo and Andreas Meyer-Lindenberg (2018). “Machine learning for precision psychiatry: opportunities and challenges”. In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3.3, pp. 223–230.

Caronna, Elizabeth B, Jeff M Milunsky, and Helen Tager-Flusberg (2008). “Autism spectrum disorders: clinical and research frontiers”. In: *Archives of Disease in Childhood* 93.6, pp. 518–523.

Castelhano, João et al. (2018). “Stimulus dependent neural oscillatory patterns show reliable statistical identification of autism spectrum disorder in a face perceptual decision task”. In: *Clinical Neurophysiology* 129.5, pp. 981–989.

Chakrabarti, Bhismadev and Simon Baron-Cohen (2011). “Variation in the human cannabinoid receptor CNR1 gene modulates gaze duration for happy faces”. In: *Molecular Autism* 2.1, p. 10.

Chandaka, Suryannarayana, Amitava Chatterjee, and Sugata Munshi (2009). “Cross-correlation aided support vector machine classifier for classification of EEG signals”. In: *Expert Systems with Applications* 36.2, pp. 1329–1336.

Chandler, Susie et al. (2007). “Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 46.10, pp. 1324–1332.

Chattopadhay, Aditya et al. (2018). “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 839–847.

Choi, Euisun and Chulhee Lee (2003). “Feature extraction based on the Bhattacharyya distance”. In: *Pattern Recognition* 36.8, pp. 1703–1709.

Choudhury, Anirban Dutta et al. (2014). “Heartsense: Estimating heart rate from smartphone photoplethysmogram using adaptive filter and interpolation”. In: *International Internet of Things Summit*. Springer, pp. 203–209.

Clarkson, Tessa et al. (2019). “T54. Youth With ASD Symptoms Have Reduced LPP as Learn About Unpredictable Peers During Social Interactions”. In: *Biological Psychiatry* 85.10, S149–S150.

Coates, Adam and Andrew Y Ng (2011). “Selecting receptive fields in deep networks”. In: *Advances in Neural Information Processing Systems*, pp. 2528–2536.

— (2012). “Learning feature representations with k-means”. In: *Neural networks: Tricks of the trade*. Springer, pp. 561–580.

Constantino, John N (2013). *Social responsiveness scale*. Springer.

Cotter, Shane F et al. (2005). “Sparse solutions to linear inverse problems with multiple measurement vectors”. In: *IEEE Transactions on Signal Processing* 53.7, pp. 2477–2488.

Courellis, Hristos S et al. (2019). “Using a Novel Approach to Assess Dynamic Cortical Connectivity Changes Following Neurofeedback Training in Children on the Autism Spectrum”. In: *Neurotechnology and Brain Stimulation in Pediatric Psychiatric and Neurodevelopmental Disorders*. Elsevier, pp. 253–276.

Courville, Troy and Bruce Thompson (2001). “Use of structure coefficients in published multiple regression articles: β is not enough”. In: *Educational and Psychological Measurement* 61.2, pp. 229–248.

Cui, Yang et al. (2015). “Non-contact time varying heart rate monitoring in exercise by video camera”. In: *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, pp. 1–5.

Dawson, Geraldine and Raphael Bernier (2007). “Development of social brain circuitry in autism”. In: *Human behavior, learning, and the developing brain: Atypical development*, pp. 28–56.

Dawson, Geraldine, Sara Jane Webb, and James McPartland (2005). “Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies”. In: *Developmental neuropsychology* 27.3, pp. 403–424.

Dawson, Geraldine et al. (2002). “Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development”. In: *Child development* 73.3, pp. 700–717.

Debener, Stefan et al. (2012). “How about taking a low-cost, small, and wireless EEG for a walk?” In: *Psychophysiology* 49.11, pp. 1617–1621.

Delorme, Arnaud and Scott Makeig (2004). “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis”. In: *Journal of neuroscience methods* 134.1, pp. 9–21.

Delorme, Arnaud, Scott Makeig, and TJ Sejnowski (2001). “Automatic artifact rejection for EEG data using high-order statistics and independent component analysis”. In: *Proceedings of the third international ICA conference*, pp. 9–12.

Dennis, Tracy A and Greg Hajcak (2009). “The late positive potential: a neurophysiological marker for emotion regulation in children”. In: *Journal of Child Psychology and Psychiatry* 50.11, pp. 1373–1383.

Dinov, Martin and Robert Leech (2017). “Modeling uncertainties in eeg microstates: analysis of real and imagined motor movements using probabilistic clustering-driven training of probabilistic neural networks”. In: *Frontiers in human neuroscience* 11, p. 534.

Duan, Ruo-Nan, Jia-Yi Zhu, and Bao-Liang Lu (2013). “Differential entropy feature for EEG-based emotion classification”. In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, pp. 81–84.

Duin, RPW (2000). “Prtools version 3.0: A matlab toolbox for pattern recognition”. In: *Proc. of SPIE*. Citeseer.

Dumoulin, Vincent and Francesco Visin (2016). “A guide to convolution arithmetic for deep learning”. In: *arXiv preprint arXiv:1603.07285*.

Eldridge, Justin et al. (2014). “Robust features for the automatic identification of autism spectrum disorder in children”. In: *Journal of neurodevelopmental disorders* 6.1, p. 12.

Euser, Anne M, Friedo W Dekker, and Saskia le Cessie (2008). “A practical approach to Bland-Altman plots and variation coefficients for log transformed variables”. In: *Journal of clinical epidemiology* 61.10, pp. 978–982.

Fan, Jing et al. (2015). “A Step towards EEG-based brain computer interface for autism intervention”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 3767–3770.

Fan, Jing et al. (2017a). “EEG analysis of facial affect recognition process of individuals with ASD performance prediction leveraging social context”. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, pp. 38–43.

Fan, Jing et al. (2017b). “EEG-based affect and workload recognition in a virtual driving environment for ASD intervention”. In: *IEEE Transactions on Biomedical Engineering* 65.1, pp. 43–51.

Ferri, Jamie, Anna Weinberg, and Greg Hajcak (2012). “I see people: The presence of human faces impacts the processing of complex emotional stimuli”. In: *Social neuroscience* 7.4, pp. 436–443.

Fletcher-Watson, Sue and Francesca Happé (2019). *Autism: a new introduction to psychological theory and current debate*. Routledge.

Fletcher-Watson, Sue et al. (2014). “Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD)”. In: *Cochrane Database of Systematic Reviews* 3.

Foti, Dan, Greg Hajcak, and Joseph Dien (2009). “Differentiating neural responses to emotional pictures: evidence from temporal-spatial PCA”. In: *Psychophysiology* 46.3, pp. 521–530.

França, FMG et al. (2014). “Advances in weightless neural systems”. In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 497–504.

Frantzidis, Christos A et al. (2010). “Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli”. In: *IEEE Transactions on Information Technology in Biomedicine* 14.3, pp. 589–597.

Friederici, Angela D and Wolf Singer (2015). “Grounding language processing on basic neurophysiological principles”. In: *Trends in cognitive sciences* 19.6, pp. 329–338.

Gao, Yongbin, Hyo Jong Lee, and Raja Majid Mehmood (2015). “Deep learninig of EEG signals for emotion recognition”. In: *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, pp. 1–5.

Gerber, Andrew J et al. (2008). “An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces”. In: *Neuropsychologia* 46.8, pp. 2129–2139.

Ghosh, Arindam, Morena Danieli, and Giuseppe Riccardi (2015). “Annotation and prediction of stress and workload from physiological and inertial signals”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 1621–1624.

Ghosh, Arindam et al. (2015). “Detection of essential hypertension with physiological signals from wearable devices”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 8095–8098.

Giavarina, Davide (2015). “Understanding bland altman analysis”. In: *Biochemia medica: Biochemia medica* 25.2, pp. 141–151.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323.

Golan, Ofer, Simon Baron-Cohen, and Jacqueline Hill (2006). “The Cambridge mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger syndrome”. In: *Journal of autism and developmental disorders* 36.2, pp. 169–183.

Golan, Ofer et al. (2010). “Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces”. In: *Journal of autism and developmental disorders* 40.3, pp. 269–279.

Gómez-Herrero, Germán et al. (2006). “Automatic removal of ocular artifacts in the EEG without an EOG reference channel”. In: *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*. IEEE, pp. 130–133.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Gorodnitsky, Irina F and Bhaskar D Rao (1997). “Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm”. In: *IEEE Transactions on signal processing* 45.3, pp. 600–616.

Goroshin, Ross et al. (2015). “Unsupervised learning of spatiotemporally coherent metrics”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 4086–4093.

Gotham, Katherine, Andrew Pickles, and Catherine Lord (2009). “Standardizing ADOS scores for a measure of severity in autism spectrum disorders”. In: *Journal of autism and developmental disorders* 39.5, pp. 693–705.

— (2012). “Trajectories of autism severity in children using standardized ADOS scores”. In: *Pediatrics* 130.5, e1278–e1284.

Gotham, Katherine et al. (2007). “The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity”. In: *Journal of autism and developmental disorders* 37.4, p. 613.

Govindarajan, Usha and Narasimhan Kumaravelu (2019). “A Review of Electroencephalogram Signal as Clinical Decision Support System”. In: *Systematic Reviews in Pharmacy* 10.1, pp. 49–54.

Gresham, Frank M et al. (2011). “Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels.” In: *School Psychology Quarterly* 26.1, p. 27.

Grossard, Charline et al. (2017). “Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD)”. In: *Computers & Education* 113, pp. 195–211.

Han, Hyonyoung and Jung Kim (2012). “Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method”. In: *Computers in biology and medicine* 42.4, pp. 387–393.

Handouzi, Wahida et al. (2014). “Objective model assessment for short-term anxiety recognition from blood volume pulse signal”. In: *Biomedical Signal Processing and Control* 14, pp. 217–227.

Harms, Madeline B, Alex Martin, and Gregory L Wallace (2010). “Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies”. In: *Neuropsychology review* 20.3, pp. 290–322.

Heunis, T et al. (2018). “Recurrence quantification analysis of resting state EEG signals in autism spectrum disorder–a systematic methodological exploration of technical and demographic confounders in the search for biomarkers”. In: *BMC medicine* 16.1, p. 101.

Hobson, R Peter, J Ouston, and Antony Lee (1988). “Emotion recognition in autism: Coordinating faces and voices”. In: *Psychological medicine* 18.4, pp. 911–923.

Hooker, Sara et al. (2018). “Evaluating feature importance estimates”. In: *arXiv preprint arXiv:1806.10758*.

Hopkins, Ingrid Maria et al. (2011). “Avatar assistant: improving social skills in students with an ASD through a computer-based intervention”. In: *Journal of autism and developmental disorders* 41.11, pp. 1543–1555.

Huang, Houjing et al. (2018). “Adversarially Occluded Samples for Person Re-Identification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5098–5107.

Hyvarinen, Aapo (1999). “Fast and robust fixed-point algorithms for independent component analysis”. In: *IEEE transactions on Neural Networks* 10.3, pp. 626–634.

Hyvärinen, Aapo and Erkki Oja (2000). “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5, pp. 411–430.

Ingalhalikar, Madhura et al. (2014). “Creating multimodal predictors using missing data: Classifying and subtyping autism spectrum disorder”. In: *Journal of neuroscience methods* 235, pp. 1–9.

Iqbal, Sajid et al. (2016). “Application of intelligent agents in health-care”. In: *Artificial Intelligence Review* 46.1, pp. 83–112.

Iyer, Ganesh et al. (2018). “Geometric consistency for self-supervised end-to-end visual odometry”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 267–275.

Jamal, Wasifa et al. (2014). “Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates”. In: *Journal of neural engineering* 11.4, p. 046019.

Jamal, Wasifa et al. (2015). “On the existence of synchrostates in multichannel EEG signals during face-perception tasks”. In: *Biomedical Physics & Engineering Express* 1.1, p. 015002.

Jenke, Robert, Angelika Peer, and Martin Buss (2014). “Feature extraction and selection for emotion recognition from EEG”. In: *IEEE Transactions on Affective Computing* 5.3, pp. 327–339.

Jeon, Hyeon-Ae and Angela D Friederici (2015). “Degree of automaticity and the prefrontal cortex”. In: *Trends in cognitive sciences* 19.5, pp. 244–250.

Ji, Yuzhu et al. (2019). “Atypical N170 lateralization of face and word recognition in Chinese children with autism spectrum disorder”. In: *Journal of Neurolinguistics* 52, p. 100858.

Jirayucharoensak, Suwicha, Setha Pan-Ngum, and Pasin Israsena (2014). “EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation”. In: *The Scientific World Journal* 2014.

Jones, Emily JH, Geraldine Dawson, and SJ Webb (2018). “Sensory hypersensitivity predicts enhanced attention capture by faces in the early development of ASD”. In: *Developmental cognitive neuroscience* 29, pp. 11–20.

Kang, Jae-Hwan et al. (2015). “Modulation of alpha oscillations in the human EEG with facial preference”. In: *PloS one* 10.9, e0138153.

Kapishnikov, Andrei et al. (2019). “Segment Integrated Gradients: Better attributions through regions”. In: *arXiv preprint arXiv:1906.02825*.

Keehn, Brandon et al. (2015). “Atypical hemispheric specialization for faces in infants at risk for autism spectrum disorder”. In: *Autism Research* 8.2, pp. 187–198.

Kindermans, Pieter-Jan et al. (2017a). “Learning how to explain neural networks: Patternnet and patternattribution”. In: *arXiv preprint arXiv:1705.05598*.

Kindermans, Pieter-Jan et al. (2017b). “Patternnet and patternlrp–improving the interpretability of neural networks”. In: *stat* 1050, p. 16.

Kindermans, Pieter-Jan et al. (2017c). “The (un) reliability of saliency methods”. In: *arXiv preprint arXiv:1711.00867*.

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Koelstra, Sander and Ioannis Patras (2013). “Fusion of facial expressions and EEG for implicit affective tagging”. In: *Image and Vision Computing* 31.2, pp. 164–174.

Koelstra, Sander et al. (2010). “Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos”. In: *International Conference on Brain Informatics*. Springer, pp. 89–100.

Koelstra, Sander et al. (2011). “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE transactions on affective computing* 3.1, pp. 18–31.

Kothe, Christian A and Scott Makeig (2011). “Estimation of task workload from EEG data: new and current tools and perspectives”. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 6547–6551.

Krishna, D Hari, IA Pasha, and T Satya Savithri (2016). “Classification of EEG motor imagery multi class signals based on cross correlation”. In: *Procedia Computer Science* 85, pp. 490–495.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*, pp. 1097–1105.

Lacava, Paul G et al. (2007). “Using assistive technology to teach emotion recognition to students with Asperger syndrome: A pilot study”. In: *Remedial and Special Education* 28.3, pp. 174–181.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *nature* 521.7553, p. 436.

Lee, Clara SC et al. (2018). “The Effectiveness of Technology-Based Intervention in Improving Emotion Recognition Through Facial Expression in People with Autism Spectrum Disorder: a Systematic Review”. In: *Review Journal of Autism and Developmental Disorders* 5.2, pp. 91–104.

Lee, Junyeon (2014). “Motion artifacts reduction from PPG using cyclic moving average filter”. In: *Technology and Health Care* 22.3, pp. 409–417.

Lerner, Matthew D, Tiffany L Hutchins, and Patricia A Prelock (2011). “Brief report: Preliminary evaluation of the theory of mind inventory and its relationship to measures of social skills”. In: *Journal of autism and developmental disorders* 41.4, pp. 512–517.

Lerner, Matthew D, James C McPartland, and James P Morris (2013). “Multimodal emotion processing in autism spectrum disorders: an event-related potential study”. In: *Developmental cognitive neuroscience* 3, pp. 11–21.

Leung, Rachel C et al. (2018). “Young adults with autism spectrum disorder show early atypical neural activity during emotional face processing”. In: *Frontiers in human neuroscience* 12, p. 57.

Leventon, Jacqueline S, Jennifer S Stevens, and Patricia J Bauer (2014). “Development in the neurophysiology of emotion processing and memory in school-age children”. In: *Developmental cognitive neuroscience* 10, pp. 21–33.

Li, He et al. (2018). “Cross-Subject Emotion Recognition Using Deep Adaptation Networks”. In: *International Conference on Neural Information Processing*. Springer, pp. 403–413.

Li, Jinpeng, Zhaoxiang Zhang, and Huiguang He (2016). “Implementation of eeg emotion recognition system based on hierarchical convolutional neural networks”. In: *International Conference on Brain Inspired Cognitive Systems*. Springer, pp. 22–33.

Lischke, Alexander et al. (2017). “Inter-individual differences in heart rate variability are associated with inter-individual differences in mind-reading”. In: *Scientific reports* 7.1, p. 11557.

Liu, Wei, Wei-Long Zheng, and Bao-Liang Lu (2016). “Emotion recognition using multimodal deep learning”. In: *International conference on neural information processing*. Springer, pp. 521–529.

Long, Mingsheng et al. (2015). “Learning transferable features with deep adaptation networks”. In: *arXiv preprint arXiv:1502.02791*.

Lopata, Christopher et al. (2010). “RCT of a manualized social treatment for high-functioning autism spectrum disorders”. In: *Journal of autism and developmental disorders* 40.11, pp. 1297–1310.

Lord, Catherine et al. (2000). “Autism spectrum disorders”. In: *Neuron* 28.2, pp. 355–363.

Lotte, Fabien et al. (2007). “A review of classification algorithms for EEG-based brain–computer interfaces”. In: *Journal of neural engineering* 4.2, R1.

Lotte, Fabien et al. (2018). “A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update”. In: *Journal of neural engineering* 15.3, p. 031005.

Luck, Steven J and Emily S Kappenman (2011). *The Oxford handbook of event-related potential components*. Oxford university press.

Luckhardt, Christina et al. (2017). “Neural correlates of explicit versus implicit facial emotion processing in ASD”. In: *Journal of autism and developmental disorders* 47.7, pp. 1944–1955.

Lundström, Sebastian et al. (2015). “Autism phenotype versus registered diagnosis in Swedish children: prevalence trends over 10 years in general population samples”. In: *bmj* 350, h1961.

Luyster, Rhiannon J et al. (2014). “Neural measures of social attention across the first years of life: Characterizing typical development and markers of autism risk”. In: *Developmental Cognitive Neuroscience* 8, pp. 131–143.

Lyall, Kristen et al. (2017). “The changing epidemiology of autism spectrum disorders”. In: *Annual review of public health* 38, pp. 81–102.

Lynn, Andrew C et al. (2018). “Functional connectivity differences in autism during face and car recognition: underconnectivity and atypical age-related changes”. In: *Developmental science* 21.1, e12508.

Maffei, Antonio, Chiara Spironelli, and Alessandro Angrilli (2019). “Affective and cortical EEG gamma responses to emotional movies in women with high vs low traits of empathy”. In: *Neuropsychologia*, p. 107175.

Mahdi, Soheil et al. (2018). “An international qualitative study of functioning in autism spectrum disorder using the World Health Organization international classification of functioning, disability and health framework”. In: *Autism Research* 11.3, pp. 463–475.

Marino, Flavia et al. (2019). “Outcomes of a Robot-Assisted Social-Emotional Understanding Intervention for Young Children with Autism Spectrum Disorders”. In: *Journal of autism and developmental disorders*, pp. 1–15.

Maris, Eric and Robert Oostenveld (2007). “Nonparametric statistical testing of EEG-and MEG-data”. In: *Journal of neuroscience methods* 164.1, pp. 177–190.

Martinez, Hector P, Yoshua Bengio, and Georgios N Yannakakis (2013). “Learning deep physiological models of affect”. In: *IEEE Computational Intelligence Magazine* 8.2, pp. 20–33.

Mayor Torres, Juan Manuel et al. (2018). “EEG-based Single trial Classification Emotion Recognition: A Comparative Analysis in Individuals with and without Autism Spectrum Disorder”. In: *International Society for Autism Research, INSAR 2018* 85.10, S149–S150.

McDuff, Daniel, Sarah Gontarek, and Rosalind W Picard (2014). “Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera”. In: *IEEE Transactions on Biomedical Engineering* 61.12, pp. 2948–2954.

McPartland, James et al. (2004). “Event-related brain potentials reveal anomalies in temporal processing of faces in autism spectrum disorder”. In: *Journal of Child Psychology and Psychiatry* 45.7, pp. 1235–1245.

Mehmood, Raja Majid, Ruoyu Du, and Hyo Jong Lee (2017). “Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors”. In: *Ieee Access* 5, pp. 14797–14806.

Mehmood, Raja Majid and Hyo Jong Lee (2015). “ERP analysis of emotional stimuli from brain EEG signal”. In: *International conference on biomedical engineering and science*, p. 5.

— (2016). “A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns”. In: *Computers & Electrical Engineering* 53, pp. 444–457.

Meulman, Nienke et al. (2014). “An ERP study on L2 syntax processing: When do learners fail?” In: *Frontiers in psychology* 5, p. 1072.

Mognon, Andrea et al. (2011). “ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features”. In: *Psychophysiology* 48.2, pp. 229–240.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). “Methods for interpreting and understanding deep neural networks”. In: *Digital Signal Processing* 73, pp. 1–15.

Mora, Niccolò, Ilaria De Munari, and Paolo Ciampolini (2015). “Improving BCI usability as HCI in ambient assisted living system control”. In: *International Conference on Augmented Cognition*. Springer, pp. 293–303.

Mühl, Christian et al. (2014). “A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges”. In: *Brain-Computer Interfaces* 1.2, pp. 66–84.

Mullen, Tim et al. (2013). “Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG”. In: *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, pp. 2184–2187.

Murias, Michael et al. (2018). “Validation of eye-tracking measures of social attention as a potential biomarker for autism clinical trials”. In: *Autism Research* 11.1, pp. 166–174.

Murphy, Brian et al. (2011). “EEG decoding of semantic category reveals distributed representations for single concepts”. In: *Brain and language* 117.1, pp. 12–22.

Nakisa, Bahareh et al. (2018). “Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors”. In: *Expert Systems with Applications* 93, pp. 143–155.

Naumann, Sandra et al. (2018). “Neurophysiological correlates of holistic face processing in adolescents with and without autism spectrum disorder”. In: *Journal of neurodevelopmental disorders* 10.1, p. 27.

Ng, Rowena, Kimberley Heinrich, and Elise K Hodges (2019). “Brief Report: Neuropsychological Testing and Informant-Ratings of Children with Autism Spectrum Disorder, Attention-Deficit/Hyperactivity Disorder, or Comorbid Diagnosis”. In: *Journal of autism and developmental disorders*, pp. 1–8.

Ngiam, Jiquan et al. (2011). “Multimodal deep learning”. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696.

Nik, H Saberi and F Soleymani (2013). “A Taylor-type numerical method for solving nonlinear ordinary differential equations”. In: *Alexandria Engineering Journal* 52.3, pp. 543–550.

Nowicki, S (2000). “Manual for the receptive tests of the Diagnostic Analysis of Nonverbal Accuracy 2”. In: *Atlanta, GA: Department of Psychology, Emory University*.

O’Leary, Heather M et al. (2017a). “Classification of respiratory disturbances in Rett Syndrome patients using Restricted Boltzmann Machine”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 442–445.

O'Leary, Heather M et al. (2017b). "Multimodal Hand Stereotypies Detection in Rett Syndrome Treatment using Deep Belief Neural Networks". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).*

Obermaier, Bernhard et al. (2001). "Hidden Markov models for online classification of single trial EEG data". In: *Pattern recognition letters* 22.12, pp. 1299–1309.

Palazzo, Simone et al. (2018). "Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features". In: *arXiv preprint arXiv:1810.10974*.

Palm, Rasmus Berg (2012). "Prediction as a candidate for learning deep hierarchical models of data". In: *Technical University of Denmark* 5.

Pan, Sinno Jialin et al. (2010). "Domain adaptation via transfer component analysis". In: *IEEE Transactions on Neural Networks* 22.2, pp. 199–210.

Parcollet, Titouan et al. (2018). "Quaternion convolutional neural networks for end-to-end automatic speech recognition". In: *arXiv preprint arXiv:1806.07789*.

Park, Seong Ho and Kyunghwa Han (2018). "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction". In: *Radiology* 286.3, pp. 800–809.

Paslawski, Teresa (2005). "The Clinical Evaluation of Language Fundamentals, (CELF-4) A Review". In: *Canadian Journal of School Psychology* 20.1-2, pp. 129–134.

Perry, Anat et al. (2015). "Interpersonal distance and social anxiety in autistic spectrum disorders: A behavioral and ERP study". In: *Social neuroscience* 10.4, pp. 354–365.

Petrantonakis, Panagiotis C and Leontios J Hadjileontiadis (2011). "A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition". In: *IEEE Transactions on Information Technology in Biomedicine* 15.5, pp. 737–746.

Pfurtscheller, Gert et al. (2006). "Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks". In: *NeuroImage* 31.1, pp. 153–159.

Pistorius, T et al. (2013). "Early detection of risk of autism spectrum disorder based on recurrence quantification analysis of electroencephalographic signals". In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, pp. 198–201.

Potter, Kristin et al. (2009). "Ensemble-vis: A framework for the statistical visualization of ensemble data". In: *2009 IEEE International Conference on Data Mining Workshops*. IEEE, pp. 233–240.

Rice, Linda Marie et al. (2015). “Computer-assisted face processing instruction improves emotion recognition, mentalizing, and social skills in students with ASD”. In: *Journal of autism and developmental disorders* 45.7, pp. 2176–2186.

Russo-Ponsaran, Nicole et al. (2018). “Virtual environment for social information processing: Assessment of children with and without autism spectrum disorders”. In: *Autism Research* 11.2, pp. 305–317.

Rutter, Michael et al. (2007). *SCQ: Social Communication Questionnaire: Manuale*. Giunti OS.

Ryan, Christian and Caitríona Ní Charragáin (2010). “Teaching emotion recognition skills to children with autism”. In: *Journal of Autism and Developmental Disorders* 40.12, pp. 1505–1511.

Safar, Kristina et al. (2018). “Increased functional connectivity during emotional face processing in children with autism spectrum disorder”. In: *Frontiers in human neuroscience* 12, p. 408.

Samaha, Jason et al. (2015). “Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction”. In: *Proceedings of the National Academy of Sciences* 112.27, pp. 8439–8444.

Sangineto, Enver et al. (2014). “We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 357–366.

Schirrmeister, Robin Tibor et al. (2017). “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Human brain mapping* 38.11, pp. 5391–5420.

Schuller, Björn et al. (2011). “Avec 2011–the first international audio/visual emotion challenge”. In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 415–424.

Schupp, Harald T et al. (2003). “Emotional facilitation of sensory processing in the visual cortex”. In: *Psychological science* 14.1, pp. 7–13.

Schupp, Harald T et al. (2004). “The facilitated processing of threatening faces: an ERP analysis.” In: *Emotion* 4.2, p. 189.

Selvaraju, Ramprasaath R et al. (2016). “Grad-CAM: Why did you say that?” In: *arXiv preprint arXiv:1611.07450*.

Selvaraju, Ramprasaath R et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626.

Senju, Atsushi et al. (2009). “Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome”. In: *Science* 325.5942, pp. 883–885.

Sharma, Manish et al. (2017). “An automatic detection of focal EEG signals using new class of time–frequency localized orthogonal wavelet filter banks”. In: *Knowledge-Based Systems* 118, pp. 217–227.

Shimazaki, Takunori and Shinsuke Hara (2015). “Breathing motion artifact cancellation in PPG-based heart rate sensing”. In: *2015 9th International Symposium on Medical Information and Communication Technology (ISMICT)*. IEEE, pp. 200–203.

Shimazaki, Takunori et al. (2014). “Cancellation of motion artifact induced by exercise for ppg-based heart rate sensing”. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 3216–3219.

Silver, Miriam and Peter Oakes (2001). “Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others”. In: *Autism* 5.3, pp. 299–316.

Simanova, Irina et al. (2010). “Identifying object categories from event-related EEG: toward decoding of conceptual representations”. In: *PloS one* 5.12, e14465.

Simoes, Marco et al. (2018). “A Novel Biomarker of Compensatory Recruitment of Face Emotional Imagery Networks in Autism Spectrum Disorder”. In: *Frontiers in neuroscience* 12, p. 791.

Singh, Manoj Kumar et al. (2015). “System modeling and signal processing of microwave Doppler radar for cardiopulmonary sensing”. In: *2015 International Conference on Signal Processing and Communication (ICSC)*. IEEE, pp. 227–232.

Smilkov, Daniel et al. (2017). “Smoothgrad: removing noise by adding noise”. In: *arXiv preprint arXiv:1706.03825*.

Soleymani, Mohammad et al. (2011). “A multimodal database for affect recognition and implicit tagging”. In: *IEEE Transactions on Affective Computing* 3.1, pp. 42–55.

Soleymani, Mohammad et al. (2015). “Analysis of EEG signals and facial expressions for continuous emotion detection”. In: *IEEE Transactions on Affective Computing* 7.1, pp. 17–28.

Solomon, Marjorie, Beth L Goodlin-Jones, and Thomas F Anders (2004). “A social adjustment enhancement intervention for high functioning autism, Asperger’s syndrome, and pervasive developmental disorder NOS”. In: *Journal of autism and developmental disorders* 34.6, pp. 649–668.

Somol, Petr, Jana Novovičová, and Pavel Pudil (2006). “Flexible-hybrid sequential floating search in statistical feature selection”. In: *Joint IAPR International Workshops on Sta-*

tistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, pp. 632–639.

Spampinato, Concetto et al. (2017). “Deep learning human mind for automated visual classification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6809–6817.

Stavropoulos, Katherine KM et al. (2018). “Autistic traits modulate conscious and nonconscious face perception”. In: *Social neuroscience* 13.1, pp. 40–51.

Stone, Wendy L, Caitlin R McMahon, and Lynnette M Henderson (2008). “Use of the Screening Tool for Autism in Two-Year-Olds (STAT) for children under 24 months: an exploratory study”. In: *Autism* 12.5, pp. 557–573.

Sturm, Irene et al. (2016). “Interpretable deep neural networks for single-trial EEG classification”. In: *Journal of neuroscience methods* 274, pp. 141–145.

Subasi, Abdulhamit and M Ismail Gursoy (2010). “EEG signal classification using PCA, ICA, LDA and support vector machines”. In: *Expert systems with applications* 37.12, pp. 8659–8666.

Sun, Biao and Zhilin Zhang (2015). “Photoplethysmography-based heart rate monitoring using asymmetric least squares spectrum subtraction and bayesian decision theory”. In: *IEEE Sensors Journal* 15.12, pp. 7161–7168.

Szepessy, Anders (1989). “An existence result for scalar conservation laws using measure valued solutions.” In: *Communications in Partial Differential Equations* 14.10, pp. 1329–1350.

Tanaka, James W et al. (2010). “Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let’s Face It! program”. In: *Journal of Child Psychology and Psychiatry* 51.8, pp. 944–952.

Thomeer, Marcus L et al. (2011). “Open-trial pilot of Mind Reading and in vivo rehearsal for children with HFASD”. In: *Focus on Autism and Other Developmental Disabilities* 26.3, pp. 153–161.

Thomeer, Marcus L et al. (2012). “Randomized clinical trial replication of a psychosocial treatment for children with high-functioning autism spectrum disorders”. In: *Psychology in the Schools* 49.10, pp. 942–954.

Thomeer, Marcus L et al. (2015). “Randomized controlled trial of mind reading and in vivo rehearsal for high-functioning children with ASD”. In: *Journal of autism and developmental disorders* 45.7, pp. 2115–2127.

Tian, Y-I, Takeo Kanade, and Jeffrey F Cohn (2001). “Recognizing action units for facial expression analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 23.2, pp. 97–115.

Torres, Juan M Mayor and Evgeny A Stepanov (2017). “Enhanced face/audio emotion recognition: video and instance level classification using ConvNets and restricted Boltzmann Machines”. In: *Proceedings of the International Conference on Web Intelligence*. ACM, pp. 939–946.

Torres, Juan M Mayor et al. (July 2018). “Enhanced Error Decoding from Error-Related Potentials using Convolutional Neural Networks”. In: *40th International Engineering in Medicine and Biology Conference, EMBC 2018*.

Torres, Juan M Mayor et al. (2019). “Distinct but Effective Neural Networks for Facial Emotion Recognition in Individuals with Autism: A Deep Learning Approach”. In: *International Society for Autism Research (INSAR) 2019 Annual Meeting*.

Torres, Juan Manuel Mayor (2013). “EEG signals classification using linear and non-linear discriminant methods”. In: *El Hombre y la Máquina* 41, pp. 71–80.

Torres, Juan Manuel Mayor, Evgeny A Stepanov, and Giuseppe Riccardi (2016). “Eeg semantic decoding using deep neural networks”. In: *Rovereto Workshop on Concepts, Actions, and Objects (CAOS)*.

Torres, Juan Manuel Mayor et al. (2016). “Heal-T: An efficient PPG-based heart-rate and IBI estimation method during physical exercise”. In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1438–1442.

Vaid, Swati, Preeti Singh, and Chamandeep Kaur (2015). “EEG signal analysis for BCI interface: A review”. In: *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. IEEE, pp. 143–147.

Vernetti, Angelina et al. (2018). “Simulating interaction: Using gaze-contingent eye-tracking to measure the reward value of social signals in toddlers with and without autism”. In: *Developmental cognitive neuroscience* 29, pp. 21–29.

Vettori, Sofie et al. (2019). “Reduced neural sensitivity to rapid individual face discrimination in autism spectrum disorder”. In: *NeuroImage: Clinical* 21, p. 101613.

Volker, Martin A et al. (2010). “BASC-2 PRS profiles for students with high-functioning autism spectrum disorders”. In: *Journal of autism and developmental disorders* 40.2, pp. 188–199.

WHO (2014). *Helping people with developmental disorders: meeting report: autism spectrum disorders and other developmental disorders: from raising awareness to building capacity: World Health Organization, Geneva, Switzerland 16-18 September 2013: easy read*. Tech. rep. World Health Organization.

Waddington, Francesca et al. (2018). “An emotion recognition subtyping approach to studying the heterogeneity and comorbidity of autism spectrum disorders and attention-deficit/hyperactivity disorder”. In: *Journal of neurodevelopmental disorders* 10.1, p. 31.

Wang, Jun et al. (2013). “Resting state EEG abnormalities in autism spectrum disorders”. In: *Journal of neurodevelopmental disorders* 5.1, p. 24.

Wang, Lijun et al. (2015). “Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing”. In: *Scientific reports* 5, p. 8708.

Wang, Xiao-Wei, Dan Nie, and Bao-Liang Lu (2014). “Emotional state classification from EEG data using machine learning approach”. In: *Neurocomputing* 129, pp. 94–106.

Webb, Sara J et al. (2006). “ERP evidence of atypical face processing in young children with autism”. In: *Journal of autism and developmental disorders* 36.7, p. 881.

Webb, Sara Jane, Emily Neuhaus, and Susan Faja (2017). “Face perception and learning in autism spectrum disorders”. In: *The Quarterly Journal of Experimental Psychology* 70.5, pp. 970–986.

Webb, Sara Jane et al. (2011). “Developmental change in the ERP responses to familiar faces in toddlers with autism spectrum disorders versus typical development”. In: *Child development* 82.6, pp. 1868–1886.

Weitz, Katharina et al. (2018). “Towards Explaining Deep Learning Networks to Distinguish Facial Expressions of Pain and Emotions”. In: *Forum Bildverarbeitung 2018*. KIT Scientific Publishing, p. 197.

Wendt, Frank R et al. (2019). “The effect of the genetic liability to autism spectrum disorder on emotion recognition in young unaffected probands from a population-based cohort”. In: *medRxiv*, p. 19001230.

White, Susan Williams, Kathleen Keonig, and Lawrence Scahill (2007). “Social skills development in children with autism spectrum disorders: A review of the intervention research”. In: *Journal of autism and developmental disorders* 37.10, pp. 1858–1868.

Whyte, Elisabeth M, Joshua M Smyth, and K Suzanne Scherf (2015). “Designing serious game interventions for individuals with autism”. In: *Journal of autism and developmental disorders* 45.12, pp. 3820–3831.

Wijnhoven, Lieke AMW et al. (2018). “Prevalence and risk factors of anxiety in a clinical Dutch sample of children with an autism spectrum disorder”. In: *Frontiers in psychiatry* 9, p. 50.

Winkler, Irene, Stefan Haufe, and Michael Tangermann (2011). “Automatic classification of artifactual ICA-components for artifact removal in EEG signals”. In: *Behavioral and Brain Functions* 7.1, p. 30.

Winkler, Irene et al. (2015). “On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 4101–4105.

Young, Robyn L and Miriam Posselt (2012). “Using the transporters DVD as a learning tool for children with autism spectrum disorders (ASD)”. In: *Journal of autism and developmental disorders* 42.6, pp. 984–991.

Zeiler, Matthew D and Rob Fergus (2014). “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer, pp. 818–833.

Zeiler, Matthew D, Graham W Taylor, Rob Fergus, et al. (2011). “Adaptive deconvolutional networks for mid and high level feature learning.” In: *ICCV*. Vol. 1. 2, p. 6.

Zhang, Quan-shi and Song-Chun Zhu (2018). “Visual interpretability for deep learning: a survey”. In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 27–39.

Zhang, Tong et al. (2018). “Spatial–temporal recurrent neural network for emotion recognition”. In: *IEEE transactions on cybernetics* 49.3, pp. 839–847.

Zhang, Yong, Xiaomin Ji, and Suhua Zhang (2016). “An approach to EEG-based emotion recognition using combined feature extraction method”. In: *Neuroscience letters* 633, pp. 152–157.

Zhang, Zhilin (2015). “Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction”. In: *IEEE transactions on biomedical engineering* 62.8, pp. 1902–1910.

Zhang, Zhilin, Zhouyue Pi, and Benyuan Liu (2014). “TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise”. In: *IEEE Transactions on biomedical engineering* 62.2, pp. 522–531.

Zheng, Wei-Long and Bao-Liang Lu (2015). “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks”. In: *IEEE Transactions on Autonomous Mental Development* 7.3, pp. 162–175.

Zheng, Wei-Long et al. (2015). “Transfer components between subjects for EEG-based emotion recognition”. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 917–922.

Zheng, Wei-Long et al. (2018). “Emotionmeter: A multimodal framework for recognizing human emotions”. In: *IEEE transactions on cybernetics* 49.3, pp. 1110–1122.

Appendix A

Statistical Generalized Linear Model (GLM) for variable interaction

For the linear regressions executed in the current dissertation, we applied a basic linear regression model described by Equation A.1, where x is the independent variable that can be the FER performances, the ADOS-CS values, and y the dependent variable or Deep ConvNet accuracies.

For our specific case G is the binary indexes for both groups TD and ASD, and b is the y-axis intercept for the complete regression. When the analyzed samples only include TD or ASD groups β_2 and β_3 are suppressed from the regression analysis.

The β_s are defined as the linear regression factors calculated from the (x_i,y_i) subject pair being them the individual variables to correlate behavioral or ConvNet-based. For the analyses explained on Chapter 6 x-axis changes for the ADOS-CSS, then x will be the ADOS-CSS, and y FER and Deep ConvNet accuracies finding similar parameter regressions reported on Tables 6.5, 6.6, and 6.7.

$$y = \beta_1 x + \beta_2 G + \beta_3 Gx + b \tag{A.1}$$

Each β value has a p-value related showing the level of relevance of this β value into the corresponding regression. The R-pearson value reported on Tables 6.5, 6.6, and 6.7 is proportional to the slope calculate from the regression and it is more significant meanwhile more pairs (x_i,y_i) support the linear model inclination. This type of regression is not very sensitive to outliers so the values calculating from it are more formally considered as descriptors of the two variables relationship (Courville and Thompson, 2001).

Appendix B

Some Saliency Methods

In this appendix we will report the LRP Taylor-type constraint (Montavon, Samek, and Müller, 2018) and some graphical results for the rest of the most important iNNvestigate (Alber et al., 2019) package methods such as Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline. We won't describe the complete models for each of these extra saliency methods. To refer to the complete models of each saliency maps mentioned above see Chapter 7.

To complemente the LRP-z and LRP-ε methods the LRP Deep Taylor is defined. From these models as we mentioned in Chapter 7 the distractor can not be detected based on propagating reconstruction.

Therefore to control the relevance propagation even more we formulate a derivative linear model can that can additively affect the distractor propagation. To isolate the level of signal propagated from the Deep ConvNet we modify the direction of the signal s analyzing the attribution/relevance-based from LRP using Taylor constraint root value denoted by x_0.

To transform the relevance dimensionality following that premise we can define the relevance of the top layer following the Taylor-type decomposition without a residual term and differentiating the output function $f(x)$ in Equation B.1.

$$R_d^1 = (x - x_0)_d \frac{\partial f}{\partial x_d} x_0 \tag{B.1}$$

The roots of x_0 are calculated with the nearest neighbors approximation having the learned output decision function $f(x)$. Subsequently, the last step of the back-propagated LRP is to normalize the final relevance map R_d^1 between $[-1, 1]$. We can generalize the Equation B.1

using the parameters learnt in the network and follow the Equation B.2.

$$R_d^l = \sum_j \frac{\partial R_d^{l+1}}{\partial \omega_T}(\omega_T x + b_j) \tag{B.2}$$

We can denote the LRP Deep Taylor method applying the Equation B.2 but without any $\alpha\beta$ numerical balance as we describe in Chapter 7.

Figures B.1, B.2, B.3, B.4, and B.5 show the relevance maps for the Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods saliency methods, and for Average, Happy, Sad, Angry, Fear classes respectively. As we can see in these results the relevance maps differs considerably across the saliency methods even comparing them statistically with the methods reported in Chapter 7.

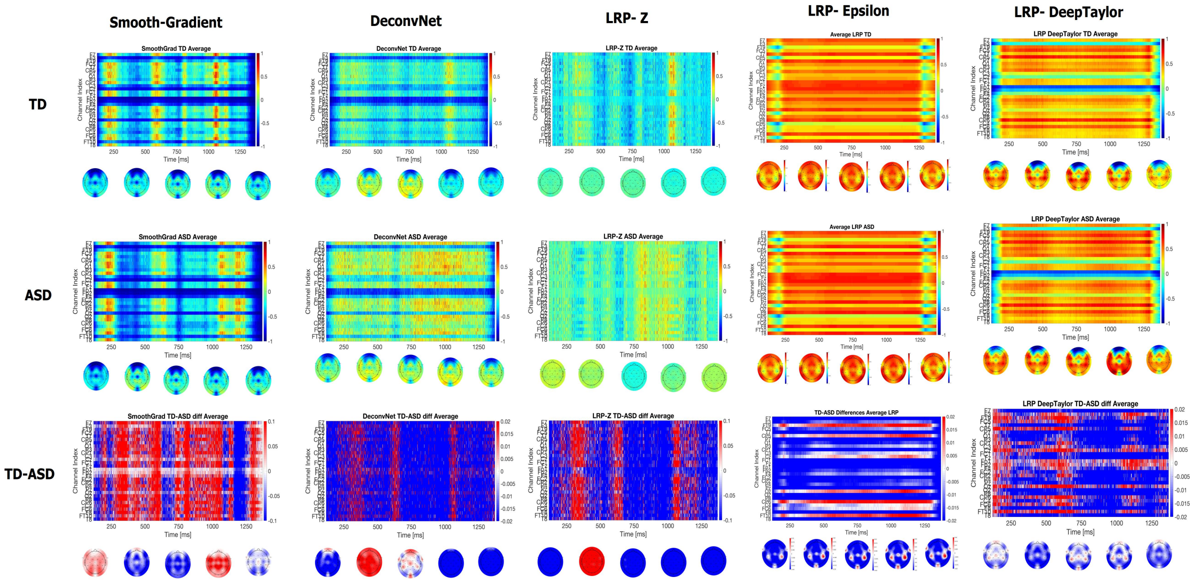


Fig. B.1 Average relevance maps for Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods evaluated on class Average. The methods name are denoted in the columns and the groups TD, ASD, and TD-ASD are denoted in the rows.

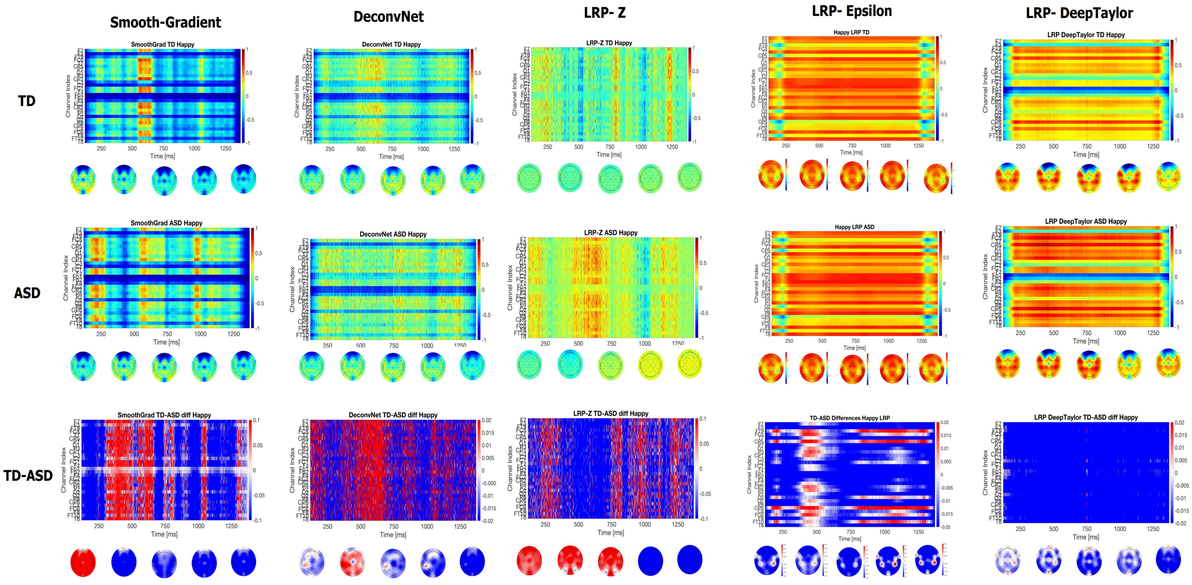


Fig. B.2 Average relevance maps for Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods evaluated on class Happy. The methods name are denoted in the columns and the groups TD, ASD, and TD-ASD are denoted in the rows.

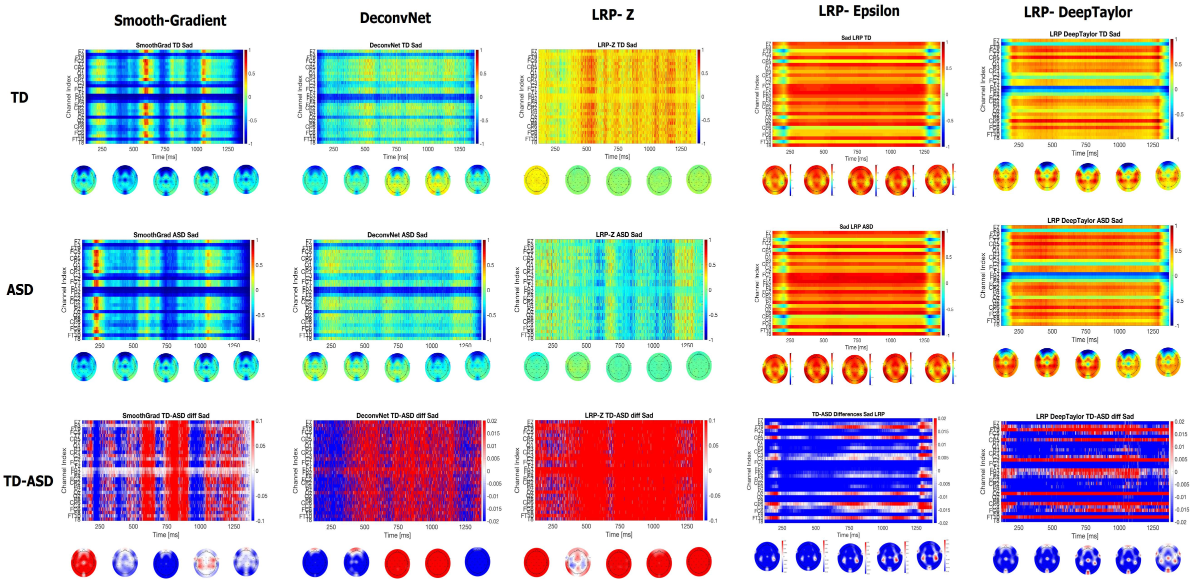


Fig. B.3 Average relevance maps for Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods evaluated on class Sad. The methods name are denoted in the columns and the groups TD, ASD, and TD-ASD are denoted in the rows.

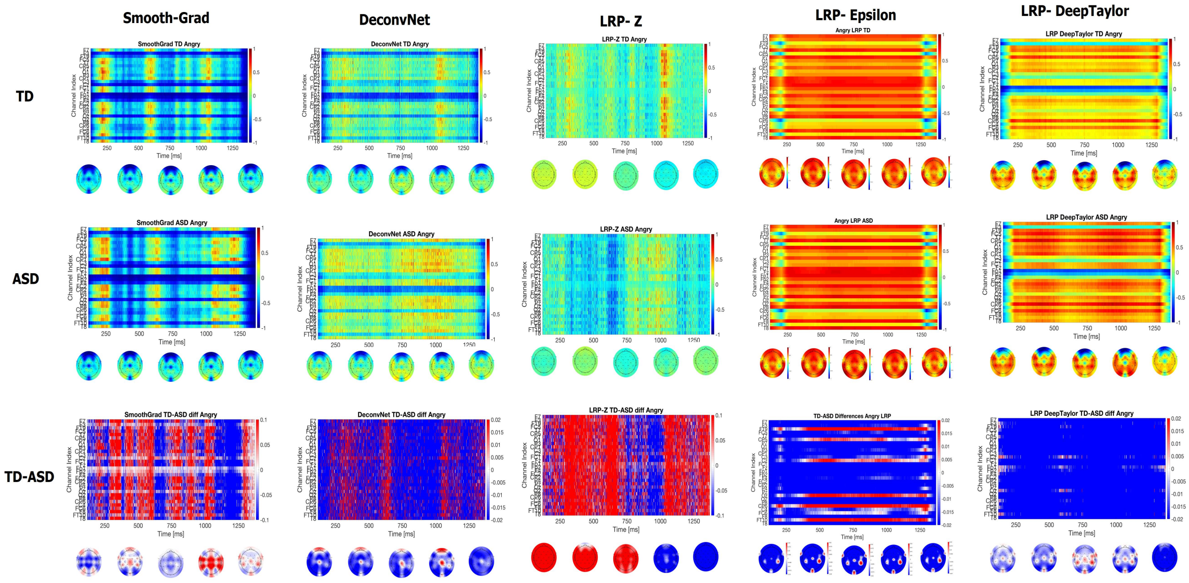


Fig. B.4 Average relevance maps for Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods evaluated on class Angry. The methods name are denoted in the columns and the groups TD, ASD, and TD-ASD are denoted in the rows.

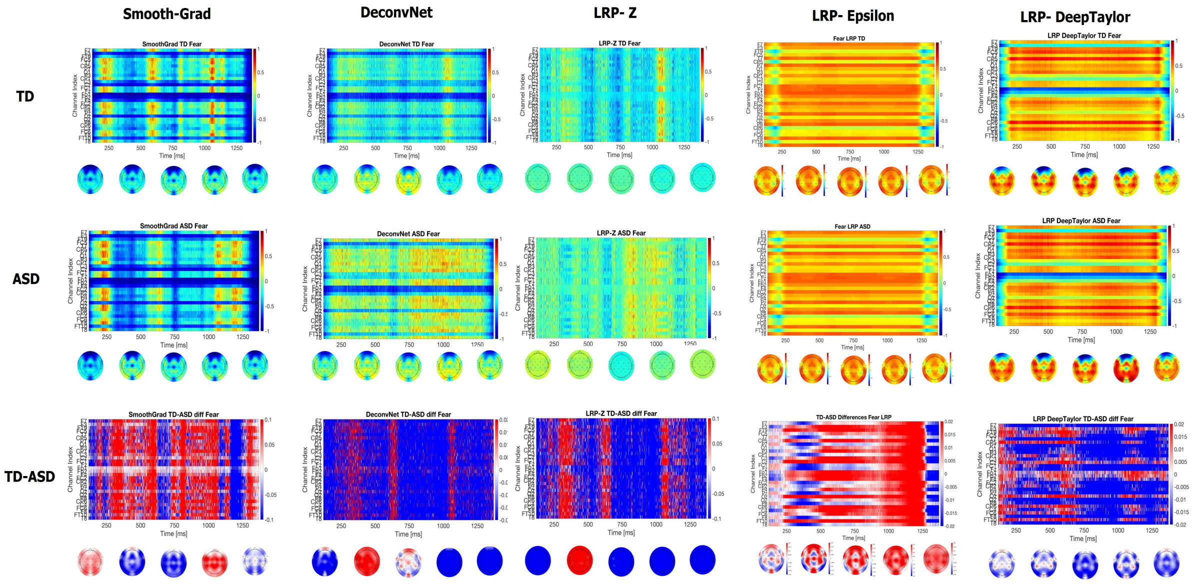


Fig. B.5 Average relevance maps for Deconvolution, LRP-z, LRP-ε, LRP Deep Taylor and Smooth-Grad baseline methods evaluated on class Fear. The methods name are denoted in the columns and the groups TD, ASD, and TD-ASD are denoted in the rows.