**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Sahar Harati Date

Machine Learning Methods for Quantification of Depression Severity and Prediction of Recovery Trajectory using Longitudinal Video and Audio Data, with Applications to Deep Brain Stimulation Treatment Optimization

By

Sahar Harati
Doctor of Philosophy

Computer Science and Informatics

---

Shamim Nemati, Ph.D.
Advisor

---

Andrea Crowell, MD, Ph.D.
Committee Member

---

Helen Mayberg, MD, Ph.D.
Committee Member

---

Ashish Sharma, Ph.D.
Committee Member

Accepted:

---

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

---

Date

# Machine Learning Methods for Quantification of Depression Severity and Prediction of Recovery Trajectory using Longitudinal Video and Audio Data, with Applications to Deep Brain Stimulation Treatment Optimization

By

Sahar Harati
B.S., Sharif University of Technology, Iran, 2013
M.S., Emory University, GA, 2018

Advisor: Shamim Nemati, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2019

Abstract

Machine Learning Methods for Quantification of Depression Severity and Prediction of Recovery Trajectory using Longitudinal Video and Audio Data, with Applications to Deep Brain Stimulation Treatment Optimization
By Sahar Harati

Predictive analytics and computational phenotyping techniques have shown promising results in several areas of medicine, including automated classification of radiology imaging, survival analysis in cancer patients, and prediction of life-threatening events in hospitalized patients. In recent years, computational psychiatry has emerged as a field that combines multiple levels and types of data and computational modeling to improve understanding, prediction, and treatment of mental illness. Mental health patients often undergo a variety of non-invasive (e.g., cognitive counseling) and invasive (e.g., surgery) therapies before finding an effective treatment plan. Improved prediction of treatment response can shorten the duration of clinical trials and improve patient experience and outcomes. A key challenge of applying predictive modeling to this problem is that often, the effectiveness of a treatment regimen remains unknown for several weeks. In this thesis, we propose Machine Learning approaches to extracting audio-visual features for predicting the likely outcome of Deep Brain Stimulation (DBS) treatment several weeks in advance for patients suffering from major depressive disorder, a common psychiatric illness for which there are no objective, non-verbal, automated markers that can reliably track treatment response. We first explore the use of video analysis of facial expressivity in a cohort of severely depressed patients before and after DBS. We introduce a set of variability measurements to obtain unsupervised features from muted video recordings. We then leverage the link between short-term emotions and long-term depressed mood states and use a neural network model on the top of emotion-based audio features. The results show that unsupervised features extracted from these audio and video recordings, when incorporated in classification models, can discriminate different levels of depression severity during ongoing DBS treatment. Moreover, for the long term prediction and in the absence of immediate treatment-response feedback, we utilize a joint state-estimation and temporal difference learning approach to model both the trajectory of a patient's response and the delayed nature of feedbacks using deep neural networks. The results based on longitudinal recordings of patients with depression show that the learned state values are predictive of the long-term success of DBS treatments. Our findings suggest that Machine Learning models can discover objective biomarkers of depression and patient response to treatments, which have the potential to standardize treatment protocols and enhance the design of future clinical trials.

# Machine Learning Methods for Quantification of Depression Severity and Prediction of Recovery Trajectory using Longitudinal Video and Audio Data, with Applications to Deep Brain Stimulation Treatment Optimization

By

Sahar Harati
B.S., Sharif University of Technology, Iran, 2013
M.S., Emory University, GA, 2018

Advisor: Shamim Nemati, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2019

Acknowledgments

I would like to express my deepest appreciation and gratitude to my advisor Dr. Shamim Nemati for his invaluable insight, inspiration and support. I was always inspired by his in-depth knowledge of machine learning, cleverness, and visionary from which I learned a lot. My path through PhD would not be productive without my co-advisors Dr. Helen Mayberg and Dr. Andrea Crowell. They played a decisive role in advising me and supporting me with their relentless contributions.

Additionally, I would like to thank the other members of my Ph.D. dissertation committee: Dr. Ashish Sharma and Dr. Gari Clifford for their time to attend my thesis and proposal defense and their insightful suggestions and comments on my research.

My research achievements also benefit significantly from interacting with a marvelous group of faculties at Emory. I would like to acknowledge the assistance I received from Dr. Lee Cooper, Dr. Carlos Moreno, and Dr. Jun Kong, Dr. Kelly Bijanki, and Dr. Yijian Huang.

My journey through Ph.D. would not have been possible without the dedicated support form Lydia Denison, Sinad Quinn, Azadeh Tabaie, Fatemeh Amrollahi, Samaneh Nasiri, Safoora Yousefi, Fereshteh Razmi, and Renee Webb.

I feel deeply indebted to the selfless love and endless support of my parents. I cannot express in words my gratitude to them for their sacrifices for me. Their love is uncountable. I'm also grateful to my sister and brother, Sara and Mohammad, for their profound belief in my work. I would also like to express respect and gracious to my extended family members, grandparents, uncles and aunts, cousins, and my in laws for their love, trust, and support. And finally, last but by no means least, I'm always thankful to my husband, Mehrdad, for having this long journey with me, for the cheerful moments he made for us in these years, for never giving up and always encouraging me and for believing in me unconditionally. Thank you Mehrdad!

# Contents

**1 Introduction** **1**

1.1 Machine Learning and Applications in Mental Health . . . . . 1

1.1.1 Machine Learning Methods . . . . . 2

1.2 Major Depressive Disorder . . . . . 4

1.2.1 Assessment . . . . . 4

1.2.2 Treatment and Recovery . . . . . 5

1.3 Automatic Depression Assessment from Visual Cues and Audio . . . 7

1.4 Contributions . . . . . 9

1.5 Data . . . . . 10

1.5.1 Subjects and Clinical Assessment . . . . . 10

1.5.2 Video Collection . . . . . 11

**2 Visual Feature Extraction and Analysis** **13**

2.1 Introduction . . . . . 13

2.2 Preprocessing . . . . . 15

2.2.1 Face detection . . . . . 15

2.2.2 Normalization . . . . . 16

2.2.3 Face Alignment and Registration . . . . . 16

2.2.4 Downsampling . . . . . 17

2.3 Methods . . . . . 17

2.3.1 Multi-scale Entropy . . . . . 17
2.3.2 Switching Linear Dynamical Systems . . . . . 18
2.4 Results . . . . . 19
2.4.1 Visualization of Features . . . . . 19
2.5 Conclusion . . . . . 22

**3 Depression Classification via Visual Features 25**
3.1 Introduction . . . . . 25
3.2 Methods . . . . . 26
3.2.1 Evaluation Methods and Statistical Analysis . . . . . 26
3.2.2 Feature Selection . . . . . 28
3.3 Results . . . . . 29
3.4 Conclusion . . . . . 31

**4 Depression Classification via Audio Features 33**
4.1 Introduction . . . . . 33
4.2 Methods . . . . . 34
4.2.1 Preprocessing . . . . . 34
4.2.2 Basic Features . . . . . 35
4.2.3 Emotion Features . . . . . 35
4.2.4 Aggregation . . . . . 38
4.2.5 Prediction . . . . . 38
4.2.6 Baselines . . . . . 38
4.3 Results . . . . . 40
4.4 Conclusion . . . . . 43

**5 Treatment Outcome Prediction 44**
5.1 Introduction . . . . . 44
5.2 Methods . . . . . 46

5.2.1 Feature Extraction . . . . . 46
5.2.2 Temporal Difference Learning . . . . . 48
5.2.3 Baselines and Performance Measure . . . . . 52
5.3 Results . . . . . 53
5.4 Conclusion . . . . . 56

**6 Conclusion** **59**

**Appendix A Switching Linear Dynamical Systems** **64**
A.1 Modeling . . . . . 64
A.2 System Identification . . . . . 65
A.3 Experimental Setup . . . . . 66
A.4 Latent Dynamical Analysis . . . . . 66

**Appendix B Elastic Net Ordinal Logistic Regression** **69**

**Appendix C Audio Features** **72**

**Appendix D Data** **74**

**Bibliography** **79**

# List of Figures

2.1 The trajectory of HDRS for one DBS patient starting from one week pre-surgery. The highlighted area represents the transitional phase. . 15

2.2 Preprocessing steps: video in demonstration is an Advanced Motivational Interviewing sample for education uploaded by Dr. R.W. Watkins, which is publicly available on YouTube: `https://www.youtube.com/watch?v=3rSt4KIaN8I`. . . . . 16

2.3 Inferred modes for one subject during instructed alternating neutral and smile expressions. Lighter colors indicate higher probabilities. . 20

2.4 Inferred modes for another subject during spontaneous response during psychiatric interview. Lighter colors indicate higher probabilities. . . 21

2.5 Quantification of facial characteristics of three phases of recovery from MDD (depressed, transitional, and improved) using three different measures of variability. Results of applying MSE, absolute value of the eigenvalues of the latent dynamics matrix (A), and singular values of the observability matrix, are shown in columns (a), (b), and (c), respectively. Each row represents the variability features for one of the 3 representative subjects based on $5min$ video recordings during each phase of recovery. Note that a large value of MSE indicates improved variability, versus a smaller value of the other two metrics indicates improved co-variability across the face. . . . . 24

3.1 Confusion matrix for the prediction results using ROLR. Numbers represent number of videos and percentages are percentages of videos. The rows represent the predicted class and the columns represented the actual class. . . . . 31
3.2 Training and testing accuracy of predicting low, moderate, and high levels of depression severity, as a function of varying number of patients for training. . . . . 32

4.1 Preprocessing steps and network architecture of the stacked LSTM model 35
4.2 The extracted emotion representation of an interview for a patient in two phases. Top panel: while being depressed; Bottom panel: while improved . . . . . 41
4.3 ROC curve of different features . . . . . 42

5.1 The proposed model. SGLM modules estimate the latent state of the patients at different time steps while value networks predict the treatment outcome given the patient state. . . . . 51
5.2 Effect of feature set on performance . . . . . 54
5.3 Feature importance: The importance is calculated as the decrease in AUC after iteratively removing one feature at a time. Last seen HDRS (lsh) from the previous week, time (t), features in the audio feature set (a-$i$: $i^{th}$), features in the video feature set (v-$i$: $i^{th}$). . . . . 55
5.4 Trajectory of the estimated state value and HDRS for each subject. The weekly clinical scores (red circles; higher values indicate decline) are often noisy and may fluctuate from week to week. The proposed machine learning-based scores (blue diamonds; higher values indicate improvement) are less prone to weekly fluctuations and is able to predict the trajectory of a patient weeks in advance. . . . . 58

A.1 Graphical representation of the SLDS; Top) probability transitions between the three modes; Bottom) evolving SLDS over time . . . . . 67

# List of Tables

3.1 HDRS class distribution over patients . . . . . 27
3.2 Comparison of accuracy of the proposed method and the baselines . . 30
4.1 Performance Comparison . . . . . 42
5.1 AUC comparison when MSE is calculated only for forehead and eyes (Upper), nose and cheeks (Middle), mouth and chin (Lower), and for the whole face . . . . . 54
5.2 Comparison of AUC of the proposed method and the baselines . . . . 56
C.1 Audio features. *ANG: angry, EXC: excited, NEUT: neutral, var: variance, vrb: variability . . . . . 73
D.1 Chapter 2 . . . . . 75
D.2 Chapter 3 . . . . . 76
D.3 Chapter 4 . . . . . 77
D.4 Chapter 4 . . . . . 77
D.5 Chapter 5 . . . . . 78

# Chapter 1

# Introduction

## 1.1 Machine Learning and Applications in Mental Health

With the rapid advancements of technology, medical devices and wearable sensors are collecting a variety of signals and a vast amount of data in the field of psychology and mental health. Some examples are functional and structural MRIs, behavioral data, video and speech recordings, survey responses, and psychological assessments. Analyzing this volume of data is not feasible by traditional statistics with formal tests for group differences in small samples [1]. Machine learning that has emerged as a robust tool to analyze rapidly growing data has been recently used as a solution for analyzing mental health data. Bzdok *et al.* [1] have reviewed certain benefits and significances of choosing machine learning approaches over conventional statistical techniques in the field of psychiatry. Availability of such data enables machine learning, and data-driven methods join forces with traditional clinical analysis to bring new insights and techniques in helping people with a mental health condition during the course of treatment. On the other hand, the challenge of subjectivity in the measurements that are used for assessing, diagnosis, monitoring, and predicting mental health disorders,

has not yet been appropriately addressed. Thus, it's important to have an automated mechanism to identify, learn, and predict patterns in mental health data.

As stated by Tom Mitchell, machine learning is to study and develop algorithms that allow computer programs to automatically improve through experience [2]. Conventional programs have their capability hard-coded and implanted by the programmer. All the cases need to be foreseen by the programmer because the program can not generalize to unseen data. In contrast, machine learning programs improve their performance through new experiences. They learn through examples. They seek general principles and common patterns underlying in data without explicit instructions [3, 4] and generalize them to new and unseen cases. Shatte *et al.* [5] have performed a systematic review on machine learning and big data applications for mental health. They identified different categories of mental health applications for machine learning techniques such as detection and diagnosis of mental health conditions including depression, Alzheimers disease, and schizophrenia, and prognosis and treatment of mental disorders. Detecting and monitoring depression and antidepressant treatment, particularly, have been of growing interest in recent years. Researchers have studied applications of machine learning in this area by exploiting a wide range of collected and stored data such as neuroimaging data (e.g., fMRI) [6, 7], sensor data (e.g., wearables, phone) [8, 9], and speech and video data [10, 11]. Audio and video data, in particular, have been shown to have high potential in detecting and assessing depression in a more automated and less subjective way.

### 1.1.1 Machine Learning Methods

Machine learning algorithms are broadly divided into three major categories. Supervised Learning, Unsupervised Learning, and Reinforcement Learning [1]. In the following each type will be explained separately.

[1]There are other related types of machine learning methods such as semi-supervised learning, deep learning, active learning, meta-learning, transfer learning, and zero-shot learning

**Supervised learning:** In supervised learning, a set of input data $X$ with known labels $Y$ is used to learn the function that maps the input to the output. This mapping is called a trained model that can predict the label for new data. Supervised learning can be thought of as a teacher supervising the learning process by providing the target $Y$ for each input $X$. The training works by the learning algorithm iteratively making predictions on the training data and being corrected by the teacher and learning stops when the algorithm performs sufficiently well. For example, in the case of video data, input $X$ can be a video recording or an image of a patient's face with $Y$ to be the depression score. After the training process, the mapping function will be able to predict depression score for new patients given their video recordings or images of their faces.

Supervised learning problems can be further grouped into regression, classification, and forecasting problems:

- Classification: Here the output variable is a category or a discrete value. For example depression severity level (low, medium, high).
- Regression: In this problem the output variable is a continuous value, such as depression score.
- Forecasting: In this type of learning the output variable is a prediction about the future according to the past and present data. It is mostly used to analyze trends. For example, given the changes in patients depression score over the past weeks, what would be their depression scores in the following weeks.

In chapters 3 and 4 we formulate a supervised classification problem and train a predictive model for the depression severity.

**Unsupervised Learning:** Unsupervised learning applies mathematical methods to identify the structure that is coherently available in data to provide new insights. In this type of learning there is no target variable $Y$. Examples of unsupervised

learning are categorizing patients into groups for targeted and specialized treatment (Clustering), or representing the high dimensional data such as video recordings in a lower-dimensional space (Dimensionality reduction)

**Reinforcement Learning:** This type of learning tries to optimize a sequence of decisions for the best possible outcome by forcing a reward for positive outcomes and a penalty for negative outcomes. For example, the decision of what type and what dosage of antidepressant medications should be administered at each visit over a course of treatment and therapy can be best modeled and identified by applying reinforcement learning techniques.

## 1.2 Major Depressive Disorder

Mental health disorders affect a notable portion of the population at any given time. According to the World Health Organization (WHO), depression (Major Depressive Disorder or clinical depression) alone is one of the leading causes of disability around the world, afflicting approximately 300 million people. Major Depressive Disorder impairs functioning at home, at work, and in relationships [12]. It affects how a person thinks, feels, and performs his/her daily activities such as sleeping, eating, and working. Diagnosis of MDD is characterized by a collection of symptoms and signs that last for at least two weeks [13].

### 1.2.1 Assessment

Evaluating the presence or severity of depressive symptoms is currently performed by a clinician with special training in mental health in a clinical interview setting and is usually supported by self-report scales.

Depression severity can be measured with the Hamilton Depression Rating Scale (HDRS) [14], a standardized measure and the current gold standard for measuring

treatment response in psychiatric studies. A score of 7 or less is within the normal range (or in clinical remission), and a score of 20 or higher is usually considered severely depressed. The original version contains 17 scored items (HDRS17) corresponding to 17 severity symptoms of depression experienced over the past week. Symptoms include, but are not limited to, depressed mood, suicidal ideation, and psychomotor retardation. Psychomotor slowing is a well-established symptom of MDD and contributes to the severity score in a number of depression rating scales, including the 17-item HDRS. While psychomotor slowing is one of the best clinical predictors of the melancholic subtype of MDD, it is defined using clinical observation rather than any standardized quantitative measurement. Nonetheless, it has relatively high discriminative power, as it is rarely endorsed in patients who do not meet criteria for a depressive episode. Psychomotor changes observed in depression include gross motor speed, head, face, trunk, and limb movements. Specific facial behaviors including eye contact, smiling, and eyebrow movement have been shown to distinguish depressed from non-depressed subjects. Such changes are relevant to biological hypotheses about abnormalities in the thalamocortical basal ganglia circuit, as well as psychological hypotheses about the way in which depression symptoms influence interpersonal interactions and predispose to social rejection in a way that reinforces depression [15, 16].

Self-reported measures, such as Beck Depression Inventory (BDI) [17] and PHQ-9 [18], are used as complementary to clinical assessments although they are also subject to response bias [19].

### 1.2.2 Treatment and Recovery

The most common treatments available for MDD are antidepressant medications and psychotherapy, which may take weeks to months to have a therapeutic effect. Many patients respond to these treatments, however according to Rush *et al.* more than

30% fail to reach complete and sustained remission [20]. Patients with MDD who have no response or poor response to multiple antidepressant treatments are diagnosed with treatment-resistant depression (TRD). These patients have more severe disability and a higher risk of relapse [21].

In the past three decades, new approaches have been discovered for TRD treatment. Among them Deep Brain Stimulation (DBS) of the subcallosal cingulate cortex has shown promising results [21]. DBS is a neurosurgical procedure in which two electrodes are implanted into a specific brain region referred to as Cg25 or Brodmann area 25 in each hemisphere; these electrodes are connected to a pulse generator that is implanted underneath the collarbone and controls stimulation and provides the power source for the DBS system.

As described by Crowell *et al*, patients with MDD who have participated in a DBS trial exhibit significant psychomotor slowing and blunting of facial expressivity [22] While some progressive improvement in depression is observed with DBS initiation, after several weeks of DBS, patients often experience a transitional phase distinguished by a return of subjective depressive symptoms, but with preserved emotional reactivity and relatively increased negative emotions and affect. Oftentimes this phase resolves and patients proceed toward subjective improvement and stabilized treatment response. Hence, this recovery course is non-linear, with transient subjective worsening interrupting the improvement trajectory. It is also the case that DBS parameters are sometimes adjusted during the course of treatment, suggesting that clinicians suspect depressive relapse and make treatment adjustments accordingly. Thus, worsening depression rating scores may represent a transient subjective response that does not require treatment changes, or a disease relapse that does require treatment changes.

A more objective biological marker of depression that could discriminate depression severity levels and distinguish recovery phases would guide treatment decisions.

## 1.3 Automatic Depression Assessment from Visual Cues and Audio

Depression is a clinical syndrome that is assumed to have a core underlying emotional state that defines the clinical illness. Many studies have attempted to automatically classify emotional states and to classify and predict depressive states and treatment outcome.

A few studies have focused on static facial appearance and have shown that information in facial cues plays an important role in distinguishing subjects who exhibit relatively high psychiatric symptoms from those with few or no symptoms [23, 24]. Increasingly, dynamic facial expression from video is being used in an attempt to predict or classify MDD, as reviewed recently by Pampouchidou *et al.* [25]. Cohn *et al.* [26] compared clinical diagnosis of major depression with automatically-measured facial actions and vocal prosody in patients undergoing treatment for depression. Using a support vector machine (SVM) classifier and logistic regression, they achieved an accuracy of 79% in detecting depression, defined as HDRS $\geq$ 15. Others have used facial expressivity or other features to predict depression *severity* by partitioning the HDRS into multiple classes for prediction purposes, with various levels of success. For instance, Pampouchidou *et al.* [27] achieved 55% accuracy, Ramasubbu *et al.* [28] reported 52-66% accuracy, Kacem *et al.* [29] achieved an accuracy of 66%, and Dibeklioulgu *et al.* [30] reached an accuracy of 66-84%.

Classification of facial movements has been studied extensively by psychologists, with the most common method being manual coding using the Facial Action Coding System [31]. Manual FACS coding is the gold standard for analysis of facial expressions, but is very time consuming and impractical for large datasets. Automated FACS analysis represents a supervised approach to facial expression analysis that has shown promise in structured and semi-structured interviews [26, 32, 33, 34, 35, 36].

These analytical methods focus on specific action units of facial movement associated with emotional expressions. Alternative approaches, such as convolutional neural networks (CNNs) [37], consider the full range of facial expression dynamics without the constraint of pre-defined (action unit) changes, including non-discrete or not consciously observable features of facial expressivity.

In recent years, automatically identifying and monitoring depression from behavioral signals has been extensively studied [38]. Hall *et al.* have shown that decreased verbal activity and monotonous and lifeless sounding speech can be an objective indicator of depression [39]. Moreover, according to Darby *et al.*, there is a perceptible change in the pitch, speaking rate, loudness, and articulation of depressed patients before and after treatment [40]. Deriving biomarkers of depression directly from a speech signal, both at the formant and spectral level, has been explored, and is shown to be useful for classifying presence or severity of depression [41, 42, 43]. However, since speech signals form a complex feature space, despite strong ability of classification methods such as Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), which have been successfully used for robustly classifying small and sparse datasets, relatively low accuracy was achieved by these studies.

To improve the depression severity classification from audio, other researchers invested in the connection between continuous affective measures and depression [44]. Vocal affect, *i.e.*, emotional expression of speech and its relationship to the overall mood of the patient, has been explored in the domain of affective computing and social signal processing [45]. In a successful effort, Stasak *et al.* [46] improved the accuracy of their automatic depression classification method by 5% by incorporating emotion ratings.

However, utilizing emotion information is not always straightforward in many depression classification tasks. The first step is to extract emotion from audio which needs sufficiently large annotated data to be used in model training. Unfortunately,

manually labeling audio from MDD patients is expensive and time consuming, and requires human supervision. Another concern in incorporating speech emotion is the human bias due to patients' speech content. Looking at the quality of speech signal (i.e. audio) without attention to its content is a challenge even for expert psychiatrists.

## 1.4 Contributions

In order to find novel quantitative biomarkers to supplement classical measurements of antidepressant response, we first use several metrics of variability to extract unsupervised features from weekly video recordings of patients before and throughout the first six months of DBS treatment for MDD [47]. Our goal is to quantify the effect of treatment on facial expressivity. A dynamic latent variable model is used to learn a low dimensional representation of factors that describe the relationship between high-dimensional pixels in each video frame and over time. Our work is similarly based on video feature extraction, but differs from previous works in that we utilize an unsupervised dynamical systems approach to extracting predictive features from video recordings of a highly homogeneous population of severely depressed patients undergoing DBS treatment who are followed on a weekly basis for several months. Our unsupervised approach does not assume that depressed state changes in facial expressivity are limited to discrete motor units in the face or even previously defined emotional expressions (*i.e.*, sad, happy, angry). By remaining open to a broader range of dynamic changes in expressivity, we may capture elements of facial expressivity that are missed by supervised approaches, which may come at some cost to interpretability.

To address the aforementioned challenge in using audio to assess depression and recovery, we propose a predictive model built on the top of emotion-based features.

Because acquiring emotion labels of MDD patients is a challenging task, we train an emotion recognition model on an *auxiliary* annotated dataset. Then, inspired by transfer learning [48], we utilize the trained model to extract emotion features of MDD patients to be used in the classification algorithm. In the last step of our approach, we feed the emotion features into a SVM classifier, which is especially robust in clinical settings where the number of samples is small. Our preliminary results show that our approach to extract emotion features using previously trained neural networks, when combined with SVM, can outperform alternative baselines

As our main contribution, we propose a Machine Learning approach to extracting audio-visual features for predicting the likely outcome of Deep Brain Stimulation (DBS) treatment several weeks in advance. In the absence of immediate treatment-response feedback, we utilize a joint state-estimation and temporal difference learning approach to model both the trajectory of a patients response and the delayed nature of feedbacks.

## 1.5 Data

### 1.5.1 Subjects and Clinical Assessment

Videos of subjects were collected as part of an ongoing DBS for Treatment Resistant Depression (TRD) study performed at Emory University[2]. Subjects in this study were evaluated weekly by study psychiatrists for eight months, starting before DBS surgery and throughout the first six months of chronic stimulation. Interviews continued with less frequency throughout a subject's participation in this long-term study. Twelve subjects were included in this analysis (ages 35-68) who were primarily Caucasian (with the exception of one African-American patient) and female (with the exception of two male patients). Ten subjects included in this analysis were considered

[2] *www.clinicaltrials.gov*, Identifier: *NCT00367003, NCT01984710*

treatment-responders after 6 months of stimulation. Clinical response was defined as a 50% decrease from pre-surgical baseline in the HDRS-17 [14], given at every time point included in this analysis.

**Interview Videos:** All interviews for all subjects were conducted by the same two psychiatrists for the duration of the subjects' participation in the study. Interviews typically started with very open ended questions (e.g. How are you?), asked the patient to describe events and feelings from the prior week, and were increasingly specific to an individual patients specific concerns/responses as the interview progressed. These unstructured clinical interviews were videotaped, capturing spontaneous conversation and unscripted responses to typical psychiatric assessment questions. Thus, these videos documented the evolution of DBS treatment and clinical response. The interview videos of 12 subjects were collected starting from 4 weeks before surgery and continuing up to seven months after surgery, for a total of 305 videos, roughly 25 videos per subject. Videos were typically about 30 minutes long.

**Smile Videos:** A collection of videos has also been captured as a control for any software that will be used to detect activation/movement of particular groups of facial muscles called action units. In this type of video, patients are asked to keep their baseline facial expression and then smile for 20-30 seconds then resume a neutral/baseline expression. The goal of recording this type of videos is to test a software with an obvious observable facial movement.

### 1.5.2 Video Collection

All videos were recorded using a Canon Vixia HF R600 digital video camera mounted on a tripod under conditions typical of a clinical psychiatric office. The subjects were facing generally in the direction of the camera seated in front of a plain white wall. The camera and chair where the subject was seated remained constant across videos. The interview was conducted in a standard psychiatric office room with the chair

approximately 5 feet from the camera.

All of the data gathering and analytic procedures were reviewed and approved by the Emory University Institutional Review Board (IRB). Collected dataset is not publicly available, since the corresponding agreement on confidentiality of the Protected Health Information (PHI) does not allow us to share the collected videos.

# Chapter 2

# Visual Feature Extraction and Analysis

## 2.1 Introduction

During the course of treatment patients usually undergo a series of phases. These knowledge of these phases and their dynamics could be a useful way to track and oversee the treatment process. Moreover, they are helpful in characterizing the effectiveness of a treatment and how to manage clinical resources. Being able to quantify their wellness and predicting what stage the patient subjects undergo will equip them with a personalized biomarker for the treatment process.

These phases were determined clinically by a study psychiatrist, and videos selected for analysis were those considered to be representative of each clinical phase. As described by Crowell *et al*, patients with MDD who have participated in a DBS trial exhibit significant psychomotor slowing and blunting of facial expressivity [22]. While some progressive improvement in depression is observed with DBS initiation, after several weeks of DBS, patients often experience a transitional phase distinguished based on a return of subjective depressive symptoms, but with preserved

emotional reactivity and fairly increased negative emotions and affect. Oftentimes this phase resolves and patients proceed on a way of subjective improvement and stabilize in remission from depression (figure 2.1). Hence, this recovery course is non-linear, with transient subjective worsening interrupting the improvement trajectory. It is also the case that DBS parameters are sometimes adjusted during the course of treatment, suggesting that clinicians suspect depressive relapse and make treatment adjustments accordingly. Thus, worsening depression rating scores may represent a transient subjective response that does not require treatment changes, or a disease relapse that does require treatment changes. A more objective biological marker that could differentiate between these two states would guide treatment decisions. Thus, in order to establish more refined clinical markers for depression severity, we focus here on classifying facial expressivity utilizing computational video analysis.

In order to find novel quantitative biomarkers to supplement classical measurements of antidepressant response, in this chapter, we first use several metrics of variability to extract unsupervised features from weekly video recordings of patients before and throughout the first six months of DBS treatment for MDD [47]. Our goal is to quantify the effect of treatment on facial expressivity. A dynamic latent variable model is used to learn a low dimensional representation of factors that describe the relationship between high-dimensional pixels in each video frame and over time.

For the purpose of this study clinical response is defined as a 50% decrease from pre-surgical baseline in the Hamilton Depression Rating Scale-17, given at every time point included in this analysis. Three clinical phases are defined for the purpose of this analysis: depressed, transitional (the putative "rough patch"), and improved.

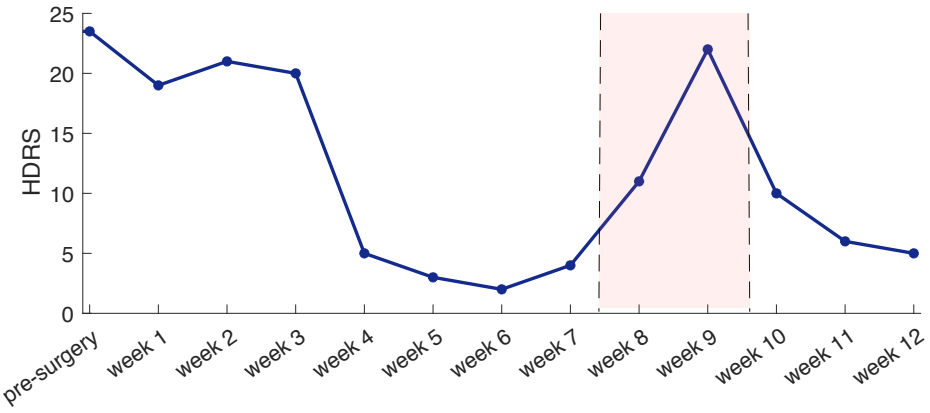


Figure 2.1: The trajectory of HDRS for one DBS patient starting from one week pre-surgery. The highlighted area represents the transitional phase.

## 2.2 Preprocessing

Figure 2.2 illustrates the preprocessing steps utilized in this thesis including 1) face detection, 2) contrast normalization, and 3) image registration, as described next.

### 2.2.1 Face detection

We use the Viola-Jones [49] and Kanade-Lucas-Tomasi (KLT) [50] algorithms implemented in the Computer Vision System Toolbox of MATLAB to automatically detect the patient's face in each video frame. We extract the area limited to the face, eliminating the hair and the background.

It's notable that failing to detect the face in a frame could happen when there was no face in the frame, for example when the patient covered her/his face or something blocked the camera (the average percentage of missing frames was 0.07% per video with standard deviation of 0.002).

In these situations, we simply ignore those frames and continue extracting faces.

### 2.2.2 Normalization

We applied histogram equalization [51] on each video frame to improve contrast. This allowed for areas of image with lower local contrast to achieve a higher contrast, and to improve details in frames that were under or over-exposed.

### 2.2.3 Face Alignment and Registration

To minimize the differences caused by different lighting conditions and variations in movement of the head, we utilize the *intensity-based automatic image registration* technique [51], implemented in `imregister` from MATLAB Image Processing Toolbox. For each video, we select the first frame with the whole face captured as the reference (fixed) image. We visually investigate the faces to make sure they are aligned after the automatic registration process, and make video clips from registered frame to make sure all pose variations and different zooming levels are eliminated.

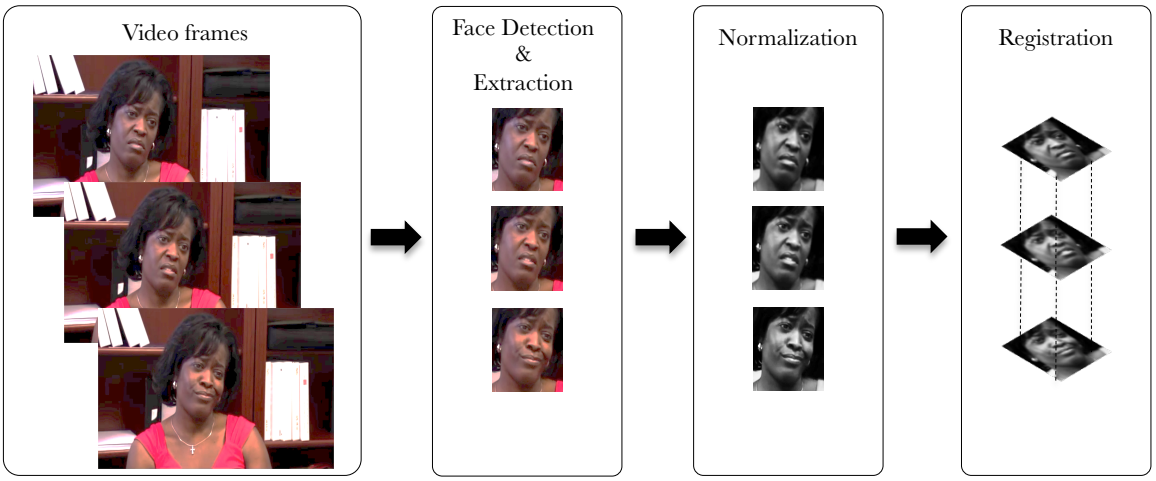


Figure 2.2: Preprocessing steps: video in demonstration is an Advanced Motivational Interviewing sample for education uploaded by Dr. R.W. Watkins, which is publicly available on YouTube: `https://www.youtube.com/watch?v=3rSt4KIaN8I`.

### 2.2.4 Downsampling

Videos are captured at a rate of 30 frames per second and are downsampled to 3 frames per second to reduce computational cost without compromising detection of facial expressions. Evolution of facial expressions occurs in much coarser resolution of 0.5 to 4 seconds [52].

## 2.3 Methods

### 2.3.1 Multi-scale Entropy

Entropy is a measure of the unpredictability or randomness of a sequence of numbers (the higher the more unpredictable). Assume a scenario in which the outcome of the sensor is relatively unpredictable, and actually performing the sensing in the next frame and learning the results gives some new information; this is a way of saying that the entropy of that pixel is large. In contrast, consider another scenario where we measure the value of the pixel in the next frame after the first one. Since the result of the former is already known, the outcome of the later can be predicted well and the results should not contain much new information; in this case we say that the entropy of the pixel value is small. *Sample entropy* is a an approximate entropy, used broadly for measuring the complexity of a time-series. For a given *embedding dimension* $m$ and *tolerance* $r$, sample entropy is defined as the negative logarithm of the probability that if two sets of simultaneous data points of length $m$ have distance smaller than $r$, then two sets of simultaneous data points of length $m + 1$ also have distance smaller than $r$. Multi-scale Entropy (MSE) is defined as the entropy of a time series at different time-scales [53]. In practice, it's calculated by repeatedly applying sample entropy to a time series after coarse-graining by a factor $\delta$.

We extract MSE from the preprocessed video frames by calculating the sample

entropy of pixels along different scales ($\delta = \{1, \ldots, 12\}$), corresponding to time-scales of 1/3 second to 4 seconds. Considering an embedding dimension of $m = 2$, this corresponds to templates of length up to 8 seconds. The MSE coefficients resulted from this procedure capture variability across a single pixel and are thus a monovariate measure of variability. For the purpose of feature engineering, we averaged the MSE coefficients across all pixels, resulting in a set of 12 features.

### 2.3.2 Switching Linear Dynamical Systems

Dynamical systems are typically used to describe the evolution of a multivariate time series. The state of a dynamical system is defined as a set of observed or latent factors, that summarize all the information required to predict the future evolution of the system. In a Linear Dynamical System (LDS), a set of linear equations describes the evolution of the system. However, most real-world time series exhibit nonlinear dynamics. This includes head movements and face occlusion events in video recordings of DBS patients, which results in a nonlinear evolution of pixel values. Nonlinear dynamical systems are generally intractable and computationally expensive to model. One approach to describing such complex dynamics is through a Switching Linear Dynamical System (SLDS) that describes the evolution of a nonlinear dynamical system as the superposition of simpler linear systems. The key parameters of a linear dynamical system are the state transition matrix $\mathbf{A}$ and the observation matrix $\mathbf{C}$, describing the evolution of the system state (here, a lower dimensional latent representation of the observed pixel values) and the relationship between the state and the observations (here, the pixel values within a frame), respectively (See Appendix A for more information). In the context of a SLDS model, we have a set of $J$ such model parameters. Assuming that the most persistent mode corresponds to the period that the patient is still, we extract the following dynamical features from the matrices $\mathbf{A}$ and $\mathbf{C}$ of such mode.

The top 10 largest eigenvalues (in magnitude) of **A** are used as a representation of the dynamical properties of the latent state, with eigenvalues with higher absolute real values (closer to 1) indicating smoother state transitions and more similarity across subsequent video frames, while smaller values indicate significant variations. In essence, these features capture the dynamics of the lower dimensional latent representations of the observed pixel values. Additionally, we extract the following 5 features: 1) the slope of eigenvalues attenuation curve (see Fig. 2.5(b)), 2) the difference between the first and the tenth eigenvalue, 3) the difference between the first and the fifth eigenvalues, 4) sum of all 10 eigenvalues, and 5) the product of all 10 eigenvalues.

The Singular Values (SVs) of the Observability matrix [54] (constructed from both **A** and **C** matrices) are another comprehensive measure of variability that capture the characteristics of the output subspaces (corresponding to the observed pixel-level dynamics) defined by the dynamical system. Similar to the analysis of dynamical properties, we use the most dominant mode of the inferred model for each video to construct the Observability matrix and extract its 10 most significant SVs. In addition, we extract 5 more features: 1) the slope of SVs attenuation curve (see Fig. 2.5(c)), 2) the gap between the first and the tenth SV, 3) the gap between the first and the fifth SV, 4) sum of all top 10 SVs, and 5) product of all top 10 SVs. (For more information about the SLDS approach and the corresponding features see Appendix A.)

## 2.4 Results

### 2.4.1 Visualization of Features

Figure 2.3 shows the switching dynamics in a video recording corresponding to 20 seconds of baseline facial expressions followed by another 25 seconds of smiling, ending

in a period of free-style recording. The inference algorithm accurately separates the neutral expressions from the smiling phase. The last 15 seconds of this video includes a typical conversation. As this example illustrates, the SLDS method is capable of segmenting the video recording into distinct dynamical modes of facial expression. Figure 2.4 shows a spontaneous recording during a psychiatric interview for another patient. The models are learned separately on each video and therefore there is no direct correspondence between the modes in figure 2.3 versus figure 2.4. While the SLDS models are fit to each video individually and therefore the meaning of a mode is video-specific, the method allows us to find the longest contiguous segments of the same dynamical mode within each video, which we then use for extracting predictive features.

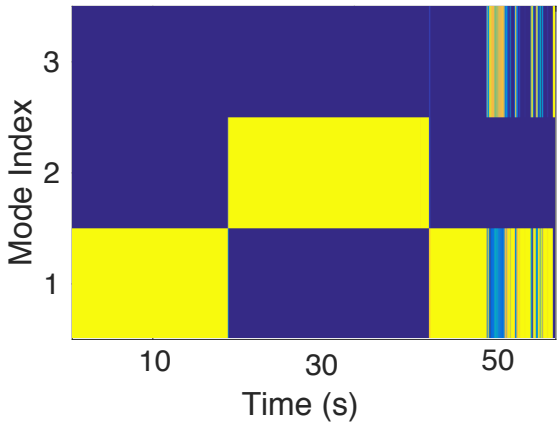


Figure 2.3: Inferred modes for one subject during instructed alternating neutral and smile expressions. Lighter colors indicate higher probabilities.

Clinical response was defined as a 50% decrease from pre-surgical baseline in the HDRS, given at every time point included in this analysis. Three clinical phases are defined for the purpose of this analysis: depressed, transitional (the putative "rough patch"), and improved. These phases were determined clinically by a study psychiatrist, and videos selected for analysis were those considered to be representative of each clinical phase.

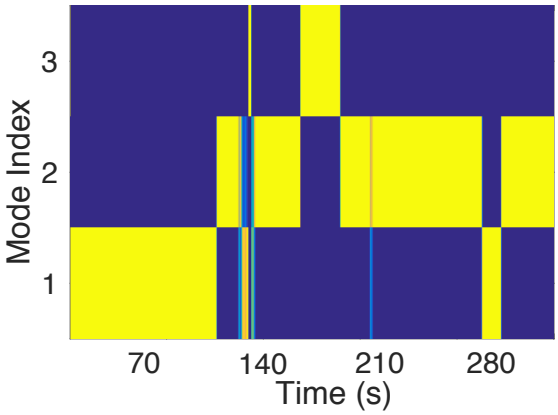


Figure 2.4: Inferred modes for another subject during spontaneous response during psychiatric interview. Lighter colors indicate higher probabilities.

### Entropy Analysis

Figure 2.5 column (a) shows the average sample entropy of pixels along different scales for 7 sample patients in 3 clinical states. To save space we don't replicate these figures for all the 12 patients, however, we observed similar patterns across all patients. Given that entropy separates different clinical states in almost all subjects, it's useful in characterizing these states. The fact that entropy being usually higher in the improved state compared to other two states (transitional and depressed) is in line with clinical observations, which report a flattened affect and limited reactivity in the depressed state. Conversely, in the improved state, higher entropy suggests more emotional reactivity and/or higher expressiveness across multiple time-scales. In most cases, the transitional state lies somewhere between these two states. Sometimes closer to depressed and sometimes closer to the improved.

### Dynamical Analysis

Figure 2.5 column b demonstrates the spectral properties of the same three patients in depressed, improved, and transitional phases. We took the 10 largest eigenvalues (in absolute value) of state evolution matrix ($\mathbf{A}$) in the most dominant mode and draw them in the figure.

Here we observe a few interesting patterns. For example, for all subjects, there is a notable discrimination between the improved and depressed phases. Furthermore, the eigenvalues of the improved state are usually further away from 1 (corresponding to flat dynamics) compared to the corresponding ones in the depressed state. This is expected due to loss of emotional expressiveness or reactivity in depressed states. Accordingly, the behavior of the transitional phase is sometimes more similar to depressed and sometimes more similar to improved states.

**Observability Analysis**

Figure 2.5-c demonstrates the spectral properties of observability by drawing 10 significant SVs of the observability matrix. It's interesting to note that the depressed, improved, and transitional phases can be distinguished using these features. In the improved phase, SVs usually have a lower magnitude compared to two other phases (depressed and transitional) . This pattern is consistent across all patients. Interestingly, the transitional phase often lies somewhere in between the other two (improved and depressed) in the most of the cases.

## 2.5 Conclusion

In summary, we have proposed three unsupervised features to analyze the clinical phases of recovery from MDD. Experimental validation on video recordings of 7 subjects confirms their power to discriminate between different clinical phases. Multiscale Entropy is a very simple yet effective feature that can be computed without much computational burden. Dynamical analysis is effective when the subject facial dynamic is itself subject to change due to movement and variations in pose. Observability features, on the other hand, are more consistent compared to the first two. It seems that by increasing complexity the features can capture the discrimination more

consistently, albeit with a computational cost.

These early results suggest that computational and quantitative approaches may lead to biomarkers for clinical state changes during treatment of depression. These techniques may be useful for both monitoring outcomes and justifying treatment interventions. These interventions could include changing DBS parameters, making a medication adjustment, or initiating psychotherapy.

One limitation of the data presented here is that our unsupervised features are used in an exploratory setting. A more systematic way that, for example, can effectively weight all the features according to their discriminatory power and classify the phases or subjects may be of more practical value. Applying these techniques to all of a subject's videos over time may show evidence of transition points in the recovery process that precede or predict subjective mood improvement. Further, while the features presented here can reliably discriminate between clinical phases in each subject, the pattern distinguishing the phases varies to some extent between subjects. This may be related to the reliance on each subject's dominant mode for analysis, which may be different across videos and across subjects. Development of a predictive marker based on video analysis would have to address this source of variability.

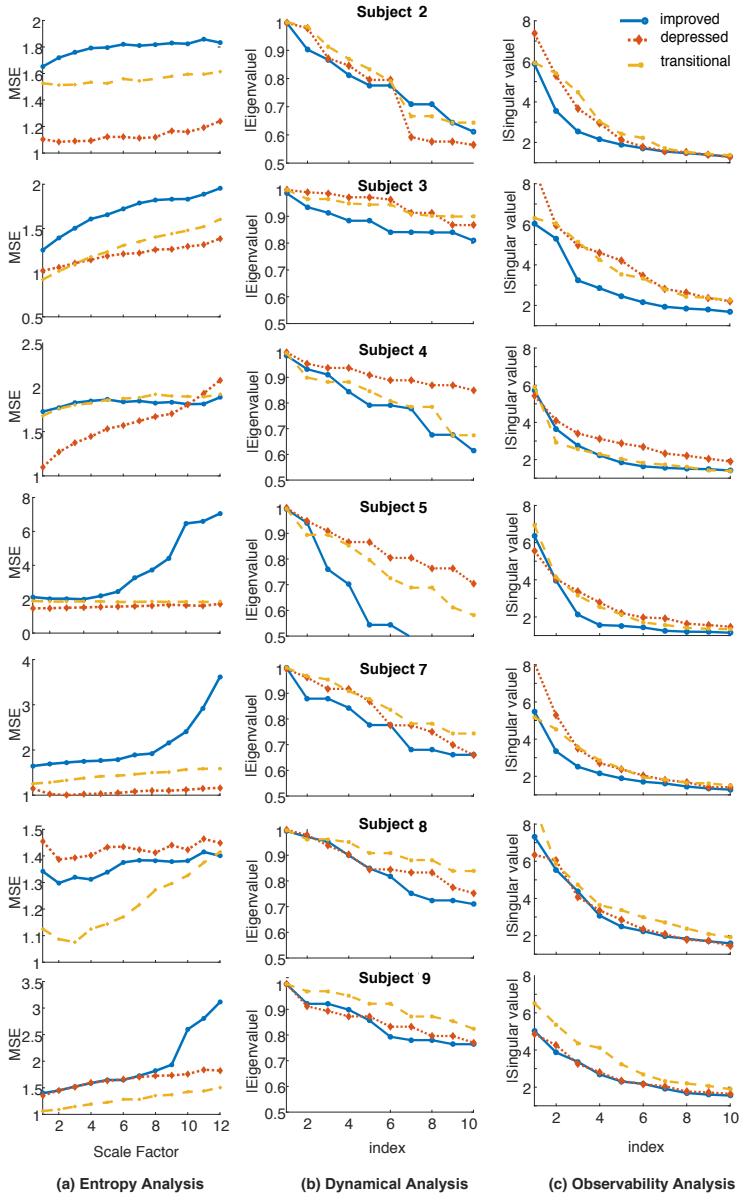


Figure 2.5: Quantification of facial characteristics of three phases of recovery from MDD (depressed, transitional, and improved) using three different measures of variability. Results of applying MSE, absolute value of the eigenvalues of the latent dynamics matrix (A), and singular values of the observability matrix, are shown in columns (a), (b), and (c), respectively. Each row represents the variability features for one of the 3 representative subjects based on $5min$ video recordings during each phase of recovery. Note that a large value of MSE indicates improved variability, versus a smaller value of the other two metrics indicates improved co-variability across the face.

# Chapter 3

# Depression Classification via Visual Features

## 3.1 Introduction

In this chapter, we use the extracted features introduced in chapter 2 to predict depression severity, corresponding to low, moderate, and high severity according to the HDRS. Effective antidepressant response is defined as a 50% improvement in this score and a score of less than 8 indicates remission from MDD. We formulate a classification problem to identify the severity of depression using the extracted features. The naive multinomial regression approach to classification ignores the natural ordering of depression severity levels. Therefore, we propose to utilize an ordinal modeling of severity classes. Classification methods are typically sensitive to noise and tend to overfit when the number of training examples compared to the number of features is small. Therefore, we augment our ordinal regression modeling to incorporate regularization to make the classification less prone to overfitting. To this end, we develop an elastic net [55] ordinal regression method to classify the severity of depression into three ordinal classes and substantiate the effectiveness of

the proposed method on this dataset. To summarize, the major contributions of this chapter are:

- Formulation of response to treatment as an ordinal classification problem, with labels defined as low, moderate and high depression ratings, and development of an ordinal regression method regularized via L1-L2 regularization (also known as an Elastic Net).
- Evaluation of the performance of the proposed recovery prediction method on 12 DBS patients, and comparison to baseline methods (e.g. multinomial logistic regression).

## 3.2 Methods

### 3.2.1 Evaluation Methods and Statistical Analysis

#### Prediction Problem Formulation

For the purpose of severity classification, we consider three classes of depression severity (according to their clinical utility), namely low, moderate, and high, corresponding to the HDRS score of 7 or less, 8 to 14, and greater than or equal to 15, respectively. A number of previous studies also have considered similar 3-class categories [56, 57, 28]. Moreover, clinical teams commonly utilize these categories to describe a patient's illness severity and make treatment decisions accordingly. Since all subjects scored in the *high* severity class in the weeks before DBS surgery, the *moderate* and *low* depression scores can also be thought of as partial and full response to treatment, respectively. Table 3.1 summarizes the number of videos in each of these three classes, for each patient. A total of 305 videos are classified according to these severity categories. The number of videos in each class is almost the same, allowing for a balanced supervised classification approach. It's important to emphasize that the

Table 3.1: HDRS class distribution over patients

| **Patient** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **total** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **low*** | 9 | 11 | 8 | 14 | 1 | 15 | 3 | 0 | 14 | 5 | 12 | 11 | 103 |
| **mid*** | 9 | 12 | 12 | 2 | 17 | 5 | 11 | 4 | 4 | 7 | 7 | 16 | 106 |
| **high*** | 5 | 3 | 4 | 14 | 12 | 9 | 18 | 13 | 4 | 2 | 5 | 7 | 96 |
| # videos | 23 | 26 | 24 | 30 | 30 | 29 | 32 | 17 | 22 | 14 | 24 | 34 | 305 |

*__low__:HDRS=$[0-8)$, *__mid__:HDRS=$[8-15)$, *__high__:HDRS=$[15-29]$

extracted features do not contain any patient identifiers, i.e. there is no hint in the features that videos (e.g. video 1 and 2 of patient 1) come from a single patient. We took this strategy to have the model generalize over patients and predict the depression phase via only facial and timing features.

All statistical evaluation methods are performed subject-wise [58]. Accuracy is reported via leave-one-subject-out, such that in each run, videos of only one patient are used for test and those of the rest are used for training. In other words, training and test sets do not share patients. This is in contrast to leave-one-video-out that uses videos from the same patient in both train and test set, which may lead to an unrealistically high accuracy. The results in the following subsections are reported by finding the mean and standard deviation of classification accuracy over 12 possible test patients and all corresponding video recordings.

### Ordinal Regression

The HDRS classification problem consists of predicting a hidden class label $y \in \mathcal{Y}$ based on features $x \in \mathcal{X}$ using a classifier $h : \mathcal{X} \to \mathcal{Y}$. Here $\mathcal{Y}$ is the depression severity levels $\{low, moderate, high\}$. Any classification method can be utilized to learn function $h$ [59]. However, to properly account for the ordinal nature of the class labels, we utilize an ordinal regression framework. Ordinal regression is different from conventional regression such that it fits both a coefficient vector and a set of thresholds

that allows for separating the output of the regression model into distinct classes. In this chapter, we propose a Regularized Ordinal Linear Regression (ROLR) method to predict the severity classes (*low*, *moderate*, *high*) given the dynamical features extracted from the patients' videos. To make the method robust with respect to noise and overfiting we add L1 and L2 losses and formulate it as an elastic net regularization (see Appendix B for more information and a link to the ROLR source code).

Multinomial Logistic Regression (MLR) is one of the most basic yet effective and popular method for multi-class classification tasks [60]. MLR can also be regularized to reduce overfitting. Therefore, we utilize an L1-L2 Regularized Multinomial Logistic Regression (RMLR; also known as an elastic net) model for benchmarking purposes. Finally, we utilize a simple Ordinal Logistic Regression (OLR) model to assess the utility of regularization as a means to improve model generalizability.

The Matlab implementation codes can be found online[1]. Similar to our work, the authors in [61, 62] have proposed ordinal regression techniques for the case of extremely high dimensional data specialized to be used in genomics implemented in R. Of note, the elastic net coefficients were learned in a subject-naive way via leave-one-out cross validation.

Our ROLR method is now ready to predict the severity classes (*low*, *moderate*, *high*) given the dynamical features extracted from the patients' videos.

### 3.2.2 Feature Selection

Our feature extraction methods resulted in 42 features, which with the addition of the video session week number (explicitly accounting for the potential effect of time on recovery), yielded our final set of 43 features. To avoid overfitting, we utilized a Bayesian Optimization (BO) approach to features selection [63]. BO is a highly competitive global optimization method for hyperparameter optimization [64]. We used

[1] https://github.com/Saharati90/DBS_Project

the internal (within the training set) cross-validated accuracy cost as the objective function for BO, with a binary vector of size 43 as the hyperparameter (with 1 indicating presence of a feature, and 0 indicating exclusion of a feature). Next, Bayesian Optimization was used to find the optimal hyperparameter values that resulted in optimizing the objective function. In contrast to the method of greedy step-wise variable selection and elimination, the BO method is not sensitive to the ordering of features, and can be used to perform feature selection concurrently with optimization of other model hyperparameters [63].

## 3.3 Results

Our feature selection method yields 25 out of the 43 features discussed in the previous section as the most predictive features. We next report the performance of the ROLR method against other competing methods using these 25 features. We report average accuracy (and interquartiles) using bootstrap sampling.

#### Overall classification

As reported in Table 3.2, we observe that the proposed ROLR method outperforms the other methods in average accuracy, although our estimation of model accuracy is imprecise due to our small sample size of 12 patients (The p-values for ROLR against MLR, RMLR, and OLR, using *exact permutation test*, are equal to 0.13, 0.12, and 0.40 respectively).

These results suggest that utilizing an ordinal model (ROLR versus RMLR) and performing regularization (ROLR versus OLR) is likely to result in a better performing model. The video recording week number was observed to be an important feature, as the accuracy of the ROLR model without this feature is 48% in contrast to the 51% accuracy with all features.

Table 3.2: Comparison of accuracy of the proposed method and the baselines

| **Classification Method** | **Average(train)** | **Range(train)** | **Average(test)** | **Range(test)** |
|---|---|---|---|---|
| Multi. Logistic Regress. (MLR) | 0.64 | [0.63,0.65] | 0.44 | [0.34,0.54] |
| Regul. Multi. Logistic. Regress. (RMLR) | 0.59 | [0.57,0.61] | 0.45 | [0.35,0.55] |
| Ordinal Logistic Regress. (OLR) | 0.57 | [0.56,0.58] | 0.49 | [0.39,0.59] |
| Regul. Ordinal Logistic Regress. (ROLR) | 0.56 | [0.54,0.58] | **0.51** | [0.41,0.61] |

### Sensitivity Analysis

Figure 3.1 shows the confusion matrix where the diagonal cells show the number of cases that were correctly classified for each class. The off-diagonal cells show the number of cases that were misclassified. The bottom right cell shows the total percentage of correctly predicted cases (in green) and the total percentage of incorrectly predicted cases (in red). White cells at the bottom of the matrix show sensitivity or recall of each target class (in green). For example, the number in row 3 and column 2 represents that 14 videos with actual moderate HDRS are incorrectly classified as high HDRS and this corresponds to 4.6 % of all 305 videos. Similarly, the number in row 2 and column 3 represents that 38 of the high HDRS videos are incorrectly classified as moderate HDRS and this corresponds to 12.5% of all 305 videos.

Cells in the forth column of confusion matrices represent the precision or positive predictive value (PPV) (in green). The higher precision of class "High" shows there is less tendency of cases to be misclassified into this class rather than the other two classes. Furthermore, cells with highest misclassification belong to either "Moderate" row or "Moderate" column, which suggests that there might not be a clear distinction between "Moderate" and "Low", and "Moderate" and "High".

### Prediction Performance

Figure 3.2 shows the prediction accuracy varying by the number of training data. The test accuracy is reported over videos of one patient which is held out separately from the training patients. The corresponding number of training patients are selected

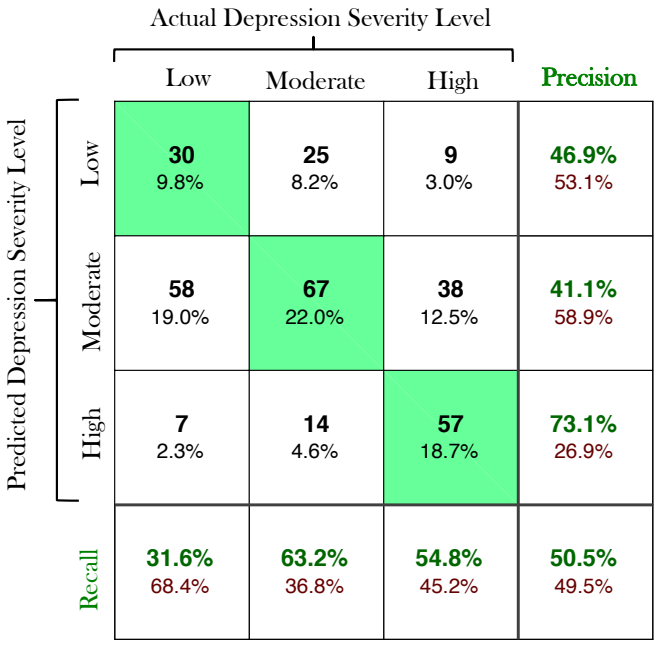


Figure 3.1: Confusion matrix for the prediction results using ROLR. Numbers represent number of videos and percentages are percentages of videos. The rows represent the predicted class and the columns represented the actual class.

randomly from the remaining 11 patients uniformly and the results are averaged over the 10 such random draws. This procedure is repeated for any of the 12 patients as test patient and averaged. As shown in the figure, the accuracy over test data has an increasing overall trend. As expected, the accuracy over the training data is decreased constantly and matches the accuracy of test data as the number of training patients increases. Note that the accuracy on training data is not a value of interest as it does not reflect the generalization or learning capability. To elaborate, an algorithm that memorizes all training data can achieve 100% accuracy on training set while no generalizable learning has taken place.

## 3.4 Conclusion

In this chapter, we leveraged the extracted features to build predictive models that can generalize from limited observed labeled data, *i.e.* patient videos and associ-

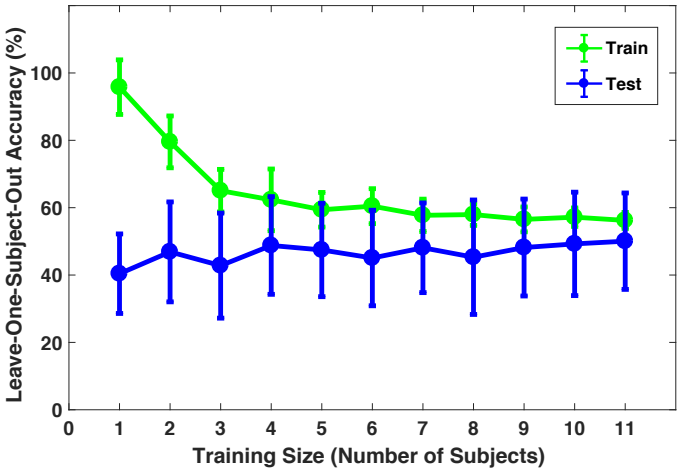


Figure 3.2: Training and testing accuracy of predicting low, moderate, and high levels of depression severity, as a function of varying number of patients for training.

ated depression severity score, and predict the depression severity level of unseen new patients. The ROLR method outperformed the alternatives in prediction accuracy. This shows the significance of ordinal modeling of the severity classes and importance of regularizing the space of feasible solutions. The former holds because the severity levels are actually ordered classes of depression rating scores, while the latter is effective because of the small number of data points compared to the dimensionality of the feature space.

Our classification model results suggest there may not be a clear distinction between moderate and low severity depression classes, nor between moderate and high severity depression classes. Taken together, a more parsimonious explanation is that there are only two distinct states (depressed, improved) marked by high or low HDRS scores. The apparent third transitional phase, initially described clinically, may not be a true independent state, but rather a period of oscillation between sick and well presenting an unstable treatment response.

# Chapter 4

# Depression Classification via Audio Features

## 4.1 Introduction

To improve the depression severity classification from audio, other researchers invested in the connection between continuous affective measures and depression [44]. Vocal affect, *i.e.*, emotional expression of speech and its relationship to the overall mood of the patient, has been explored in the domain of affective computing and social signal processing [45]. In a successful effort, Stasak *et al.* [46] improved the accuracy of their automatic depression classification method by 5% by incorporating emotion ratings.

However, utilizing emotion information is not always straightforward in many depression classification tasks. The first step is to extract emotion from audio which needs sufficiently large annotated data to be used in model training. Unfortunately, manually labeling audios from MDD patients is expensive and time consuming, and requires human supervision. Another concern in incorporating speech emotion is the human bias due to patients' speech content. Looking at the quality of speech signal (i.e. audio) without attention to its content is a challenge even for expert

psychiatrists.

To address the aforementioned challenge, we propose to train an emotion recognition model on an *auxiliary* annotated dataset. Then, inspired by transfer learning [48], we utilize the trained model to extract emotion features of MDD patients. For the emotion recognition model, we propose to take a deep learning approach [65] that can effectively learn the dynamics in audio signals.

Deep neural networks have been previously applied to recognize emotion from speech. For example, Trigeorgis *et al.* [66] performed representation learning for end-to-end speech emotion recognition. Lee *et al.* proposed a Recurrent Neural Network (RNN) based speech emotion recognition framework with an efficient learning approach, in which the label of each frame is modeled as a sequence of random variables [67]. In the last step of our approach, we feed the emotion features into an SVM classifier, which is especially robust in clinical settings where the number of samples is small.

## 4.2 Methods

### 4.2.1 Preprocessing

We first extract audio signals from video recordings. Then, we use `SpeakerDiarization` [68] method implemented in Matlab `AudioAnalysis` [69] library to first, partition the extracted audio to utterances and then select those that only contain the voice of the patient. It is noteworthy that in this study utterances refer to characteristics of the sound or speech quality not its content. Therefore, each interview session is partitioned into consecutive utterances which form our feature space with binary labels of "depressed" or "improved".

### 4.2.2 Basic Features

We extract short-term features (audio frames) using `pyAudioAnalysis` library. The feature vector extracted from each 0.2 second frame (time window) of each utterance consists of time-domain variables, (e.g. energy and entropy), and frequency-domain ones (e.g. Spectral Entropy and MFCCs), resulting in a vector of size 34 per frame per utterance. The complete list is explained in [70].

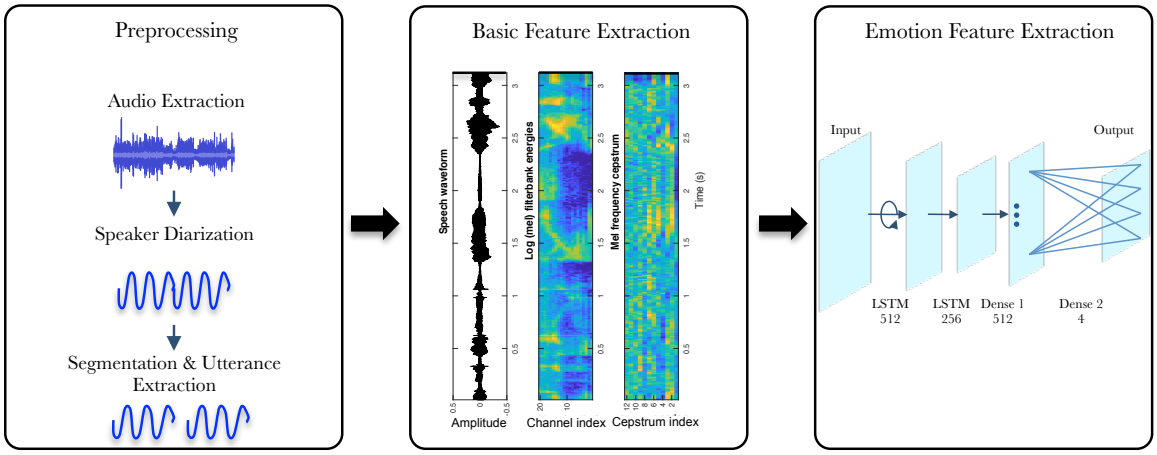


Figure 4.1: Preprocessing steps and network architecture of the stacked LSTM model

### 4.2.3 Emotion Features

We propose to use the utterances and apply an emotion recognition model to extract a low dimensional representation of the whole utterance. Due to the lack of training data we first train a deep neural network model on an auxiliary emotion dataset.

Interactive emotional dyadic motion capture (IEMOCAP) [71] dataset is one of the most commonly used datasets in emotion classification, collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California. This dataset contains 12 hours of audio-visual data recorded from five male and five female actors. Sessions of the dataset are manually segmented into utterances and each utterance is annotated by humans. We consider only the utterances that more than one annotator have agreed on the same emotion over the emotion classes

of: *angry*, *happy+excited*, *sad*, and *neutral*. These classes represent the majority of the emotion categories in this database. We combine excitement and happiness utterances to create the excitement category, as in [71]. The final dataset contains 4886 utterances (1083 angry, 1041 excited, 1678 neutral, 1084 sad). We use this dataset to train a replicated model proposed in previous studies that have been thoroughly reviewed by Swain *et al.* [72] for an emotion recognition task and use this model to predict emotions on the depression dataset.

Different advanced machine learning models including Deep Neural Networks (DNNs) [73], Long Short-Term Memory (LSTM) Networks [74], and Recurrent Neural Networks (RNNs) [75, 76], have been competing to achieve higher accuracy in predicting emotions from speech. Many of them have used IEMOCAP dataset to indicate the superiority of their performance. They have reported $52\% - 59\%$ unweighted accuracy for classifying mentioned emotions using speech.

For this thesis, we replicate the LSTM network architecture that has been used in other studies and has been shown to have a comparable performance to the more advanced ones but less complicated implementation.

By looking at the training data our neural network learns how an utterance is transformed into a useful representation for the classification task. Then, we feed the clinical audio utterances into the trained network in order to extract similar conceptual features for prediction.

In our architecture shown in Fig. 4.1, we use a stacked LSTM network [65] that has two hidden LSTM layers and each layer contains multiple memory cells. The added LSTM layers learn higher-level temporal representations. The first LSTM layer provides a sequence output rather than a single value to the second one.

Furthermore, the hidden state output of the second LTSM is carried to a fully connected layer (with *softmax* activation) to predict the probability of each emotion. After training the network on the labeled dataset, we feed the unlabeled utterances

of the DBS patient interviews to the network and use the output of the *softmax* layer to get the probability of emotions. We use these probabilities to create a new feature set of size 4 (corresponding to proportion of being angry, excited, neutral, and sad). The LSTM model is as follows

$$h_t = \mathbf{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{4.1}$$

$$y_t = W_{hy}h_t + b_y \tag{4.2}$$

where the $W$'s denote weight matrices, the $b$'s denote bias vectors and $\mathbf{H}$ is the recurrent hidden layer function implemented as in follows,

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4.3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.4}$$

$$c_t = f_t c_{t-1} + i_t a_t \tag{4.5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.6}$$

$$h_t = o_t \theta(c_t) \tag{4.7}$$

Here, $\sigma$ is the logistic sigmoid function, and $i$, $f$, $o$, $a$ and $c$ are respectively the input gate, forget gate, output gate, and cell state vectors, and all of which are the same size as the hidden vector $h$. $W_{ci}$, $W_{cf}$, $W_{co}$ are diagonal weight matrices for peephole connections. $\tau$ and $\theta$ are the cell input and cell output non-linear activation functions; $tanh$ in our architecture.

The loss function is categorical cross entropy and the optimizer is root square RMSprop. The batch size is set to 320 and number of epochs is 25.

### 4.2.4 Aggregation

Using our proposed neural network we reduce each utterance to 4 emotions. Then, per emotion, we extract 6 statistics over all utterances in each interview session. The statistics include *Minimum, Maximum, Mean, Variance, Skewness, and Kurtosis*. Therefore, we end up with a 24 dimensional representation as a distillation of emotion features of an interview.

### 4.2.5 Prediction

We apply the well known Support Vector Machines (SVM) method using the `libsvm` library [77] in order to classify videos. SVM is a discriminative classifier formally defined by a separating hyperplane that represents the largest separation, or margin, between the two classes.

### 4.2.6 Baselines

Here we introduce alternative ways to extract features and build predictors that are used for comparison.

#### Features

We have used 2 alternative feature sets.

- Basic Features: To show the effectiveness of the emotion features we experimented with the basic features as well.
- SLDS Features: A switching linear dynamic system (SLDS) describes the dynamics of a complex physical process by switching between a set of linear dynamic systems (LDS) whose characteristic are determined by latent transition matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(J)}$. SLDS has been previously utilized for analyzing the recovery of DBS patients from facial dynamics [47]. Using SLDS, we learn a low

dimensional set of dynamic factors that explain the observed covariance across the basic features within each utterance of the video clip and across time. The SLDS is evolving over time as

$$x_t = A^{(s_t)} x_{t-1} + v_t$$

for latent space and

$$y_t = C^{(s_t)} x_t + w_t$$

for observations, where $x_t$ is latent state, $v_t$ and $w_t$ are zero mean Gaussian noise processes and $y_t$ is the basic features. The details can be found in [47]. As a high-level feature set, we use the statistics of the state sequence $x_t$ of each interview session for classification. Similar to the proposed model, the same six aggregated statistics are extracted form these features are computed over all utterances of each video, leading to a $D \times 6$ dimensional feature set $D = 20$ is the state variable dimension.

**Predictors**

We have used three alternative prediction algorithms to substantiate our decision to use SVM as predictor. The first one is a simple non-sequential model just to show the benefit of a robust model like SVM. The next two models are sequence classification models that can operate over time-series and dynamical data.

- KNN: A simple k-nearest neighbor model is trained via the data with $k$ chosen by a 10-fold cross validation.
- HMM-GMM: This a sequential model comprised of a Hidden Markov Model with Gaussian mixture emissions. The HMM is a generative probabilistic model,

in which a sequence of observable $y_t$ is generated by a sequence of internal hidden states $x_t$. The transitions between hidden states are assumed to have the form of a Markov chain. Given just the observed training data, we build one ergodic HMM model for each of the depressed and improved class with 4 hidden states and we estimate the parameters for each model[1]. Then, given the parameters and observed data, we calculate the likelihood of the test data and assign the sequence to a class with higher probability.

- RNN: We also, train a Gated Recurrent Unit (GRU)-based recurrent neural network. We use the final hidden state of the RNN as input to a hidden (fully connected) layer with *Sigmoid* activation to predict the probability of recovery phase. We used `Keras`, a standard Python library, for implementing RNN [78].

## 4.3 Results

Here, we study the proposed approach experimentally and compare its effectiveness in classifying depression severity.

- Fig. 4.2 shows the extracted emotion features for a patient in a depressed state (4 weeks before DBS surgery) and non-depressed state (1 year after surgery). The proportion of each of the 4 affective states during the interview is plotted. In the depressed state sadness dominates, while the excited and neutral states become prevalent after treatment.

- Then we study the combination of the proposed emotion-based feature with SVM and compare it with other mixtures in Table 4.1. Reported performance measures are Area Under the Curve (AUC) and Positive Predictive Value (PPV) on test data using leave-one-patient-out cross validation. *I.e.*, in each run,

[1]refer to http://hmmlearn.readthedocs.io for details.

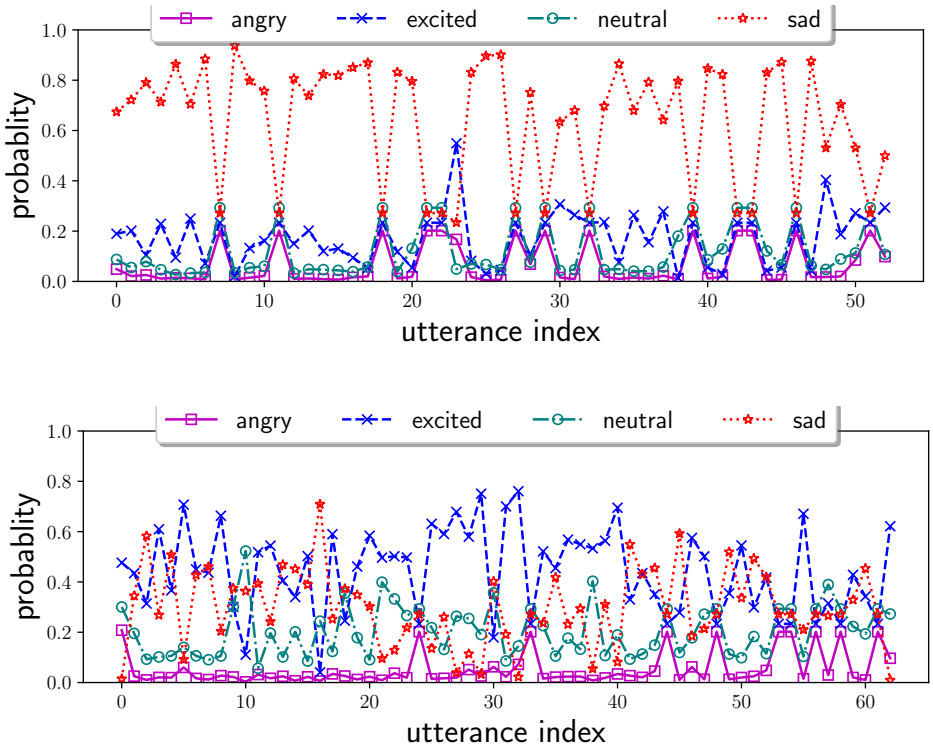


Figure 4.2: The extracted emotion representation of an interview for a patient in two phases. Top panel: while being depressed; Bottom panel: while improved

videos of one patient are used for the test and those of the rest are used for the training. This way we are confident that the performance measure is over unseen data and representative of the generalization ability of the model.

The first 3 rows show the performance when we fix the predictor (SVM) and vary the features (basic, SLDS, and emotions). It is already apparent that basic features which are composed of short-term frames have lower representative ability compared to emotion features. Moreover, SLDS, which extracts features by viewing the whole interview as dynamical systems, fails to successfully represent the audio in different phases.

- Next, we study the effectiveness of SVM compared to a conventional simple predictor (KNN), a flexible sequential model (GMM-HMM), and a recurrent neural network model. We fix the feature set (Emotion) among these classifiers

Table 4.1: Performance Comparison

| Model | | Performance | |
|---|---|---|---|
| **Feature** | **Predictor** | **AUC** | **PPV** |
| Basic | SVM | 0.64 | 0.57 |
| SLDS | SVM | 0.73 | 0.59 |
| Emotion | SVM | **0.80** | **0.70** |
| Emotion | KNN | 0.62 | 0.5 |
| Emotion | GMM-HMM | 0.57 | 0.5 |
| Emotion | RNN-GRU | 0.64 | 0.58 |

and observe that too-complex models are not effective on small datasets unless we use knowledge transfer from available labeled datasets. Furthermore, too-simple models like KNN are also unable to effectively capture the underlying concept. The model with SVM as classifier on the emotion features significantly (*i.e.* with p-value $< 0.05$) outperforms others. Finally, we plot the Receiver Operating characteristic (ROC) curve that allows one to study the sensitivity and specificity of the proposed feature set. The emotion feature consistently outperforms the other two (figure 4.3).

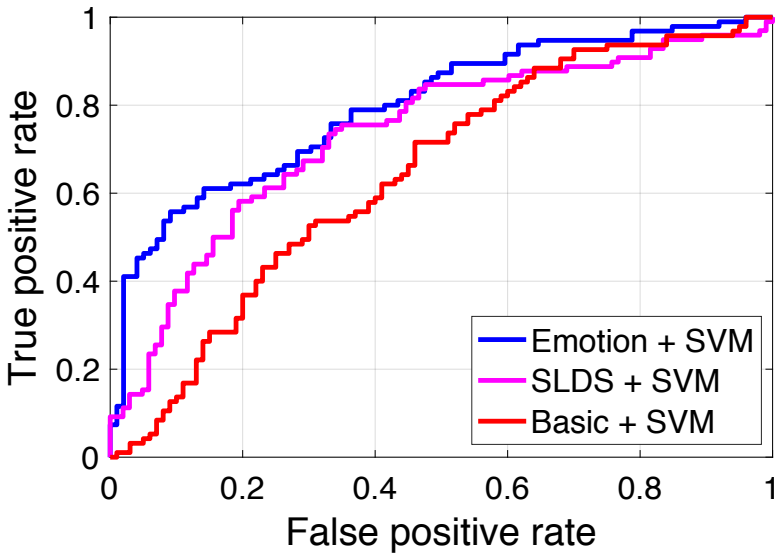


Figure 4.3: ROC curve of different features

Our preliminary results show that our approach to extract emotion features us-

ing previously trained neural networks, when combined with SVM, can outperform alternative baselines.

## 4.4 Conclusion

In this chapter, we tackled a depression severity classification problem by accounting/adjusting for short-term emotions when assessing long-term changes in mood or depression state. Our proposed model benefits from both complexity of recurrent neural networks in extracting higher level representations and the robustness of SVM in handling small and sparse training data.

# Chapter 5

# Treatment Outcome Prediction

## 5.1 Introduction

In this chapter we propose a machine learning method that first leverages a combination of both audio and visual features to achieve improved prediction accuracy, second models the sequential nature of treatment and assessment/feedback delay over the course of time by utilizing the framework of temporal difference (TD) learning in the field of Reinforcement Learning, third uses state estimation to infer the hidden state (or latent representation) of the patient over time, thus exploiting the temporal information embedded in longitudinal patient recordings, and fourth takes advantage of a deep neural network structure for the state-estimation and outcome prediction (via value iteration) that is trained end-to-end using gradient descent optimization.

Recently, there has been increasing interest in quantifying and predicting depression and treatment outcomes from both video and audio recordings. Biomarkers of depression from speech signals are shown to be useful for classifying presence or severity of depression [41, 42, 43]. For example, Darby *et al.* [40] reported a quantifiable change in the pitch, speaking rate, loudness, and articulation of depressed patients before and after treatment. Harati *et al.* [79] used emotion-related features from

audio recordings of TRD patients to train a deep neural network capable of predicting the treatment outcomes. Moreover, facial expression features derived from video recordings has been shown to be a good predictor of depression and recovery. For instance, Cohn *et al.* [26] used a support vector machine (SVM) classifier to measure spontaneous facial expressions in a small group of subjects. Others have used facial expressivity to predict depression severity either empirically [80] or using accepted clinical classification of severity: Pampouchidou *et al* [27] achieved 55% accuracy , Ramasubbu *et al* [28] reported 52-66% accuracy , Anis *et al* [29] achieved an accuracy of 66%, and Dibeklioulu *et al* [30] reached an accuracy of 66-84%.

The key shortcoming of the existing methods is that their utilized learning labels are based on short-term feedback (either subjective or clinical assessments), which may not correspond to the long-term trajectory of the patient and its outcome.

For this chapter, a total number of 14 videos are selected for each subject, each about 30 minutes long. Due to some missing weekly videos for all subjects (either due to missed acquisition or unprocessable recordings), we restricted analyses to a common dataset of 14 videos per subject covering the full 7 months (1 month pre-surgery and 6 months post-surgery) for each patient.

According to Crowell *et al.* [22], stable clinical treatment response to DBS is typically not achieved until at least 12 weeks of chronic stimulation. So, two clinical phases are considered here: *depressed* and *improved*. Treatment response for the purpose of this outcome prediction model is defined as 30% decrease from the pre-surgical baseline HDRS, resulting in nine improved and three depressed subjects. *Temporal Credit Assignment* refers to the problem of determining how the ultimate success (or failure) of a sequence of treatments is attributable to the various intermediate clinical states of the patient. We demonstrate that temporal patterns in the data captured by the proposed joint state-estimation and TD-learning framework are useful by showing that credit assignment via back-propagation allows us to train the model without im-

mediate feedback. We learn from the accumulated rewards rather than instantaneous HDRS value, which is a self-reported integration of what happened over the previous week only, thus noisy.

The proposed framework for predicting long-term success of a trial from quantifiable audio/visual features is novel due to its utilization of a TD-learning method known as *Value Iteration* to estimate the long-term accumulated reward associated with a patient state, which is indicative of a patient's long-term recovery trajectory.

## 5.2 Methods

### 5.2.1 Feature Extraction

We use both audio and visual features to test the hypothesis that fusion of multimodal data can improve prediction accuracy.

For the visual features, we leverage facial features described in chapter 2 for MDD subjects [47] that are shown to be effective in distinguishing the recovery phases of DBS patients during treatment. Briefly,

- The images are put through face detection, contrast normalization, and image registration and alignment.
- Three types of dynamical features are extracted using MSE and SLDS approach.
  - First, MSE measures the randomness or unpredictability exists in a sequence of patient's facial expression. We use scales from 1 to 12 to get 12 features, and calculated the average entropy across all the video pixels.
  - Second, an SLDS is fit to the data, which has the advantage of being multivariate and thus capable of extracting correlated activity of facial muscle groups. To capture the dynamical behavior of the video recordings of facial expression, 15 eigenvalue features (or spectral properties) of the

state transition matrix from the most dominant dynamical mode are used as another variability feature set.

- Third, 15 observability features are used for a comprehensive coverage of dynamical behavior of facial expression. These lead to an overall feature set size of 42. For more details please refer to our previous chapters or published works [47].

For the audio features we use the same technique described in chapter 4 summarized as follows.

- First, audio signals are extracted from video recordings.
- Then, from each 0.2-second frame of each utterance, time-domain variables (e.g. energy and entropy) and frequency-domain variables (e.g. Spectral Entropy and Mel-Frequency Cepstral Coefficients (MFCCs)) are extracted, resulting in a vector of size 34 per frame per utterance.
- Then, on these raw features, a Long Short Term Memory (LSTM)-based emotion recognition neural network is applied to get a 4-dimensional representation corresponding to emotions: angry, happy, sad, and neutral.
- Finally, per emotion, seven statistics over all utterances in each interview session are computed. The statistics include *Minimum, Maximum, Mean, Variance, Skewness, Kurtosis, and Variability* leading to a 28-dimensional representation as a distillation of emotion features of an interview. These audio features have proven to be effective for studying depressed subjects [79].

In summary, we extract 28 audio and 42 facial expression-related features per video recording. 'Time since start of the trial' and the 'HDRS from the preceding week' constituted two additional features, resulting in a total of 72 features per video recording.

### 5.2.2 Temporal Difference Learning

We develop a TD-learning approach to predict treatment outcome, known as Value Iteration [81]. Given the multivariate time series of features described in the previous section, a Switching Generalized Linear Model (SGLM) [82] is utilized to identify patient-specific clinical states, which is then fed into the value iteration network to assess the long-term value of a given clinical state. The overall model is optimized end-to-end as described next.

#### State-Estimation

In order to track the treatment process, we first identify the *state* ($s_t$) that the patient is in at any given point in time, which encodes all the useful information from the past required to predict the future state of the patient. We choose a supervised approach to hidden state-estimation (known as the SGLM model) under the assumption of Markovianity and a linear state transition model [82]. In the top layer, there are $J$ possible hidden states (or *modes*), and the likelihood function of states takes the form of a softmax classifier with parameter $\alpha$; mapping the observations to the likelihood of the $J$ latent states. The network uses a forward pass over the time series data to predict the latent states using the $J \times J$ transition matrix $Z$ and the supervised likelihood model. To further elaborate, consider the posterior probability of the latent state at time $t$ given the set of observations up to that time is given by

$$P\left(s_t = j \middle| \{\mathbf{x}_{1:t}\}\right) = \frac{1}{C} \cdot P_\alpha\left(x_t | s_t = j\right) \cdot \sum_{i=1}^{J} Z(i,j) \cdot P\left(s_{t-1} = i \middle| \{\mathbf{x}_{1:t-1}\}\right), \quad (5.1)$$

where, $P\left(s_t = j \middle| \{\mathbf{x}_{1:t}\}\right)$ denotes the probability that the latent state $s$ at time $t$ is equal to $j$ given the observations $\mathbf{x}_{1:t}$, $P_\alpha\left(x_t | s_t = j\right)$ is the likelihood function (the computed probability of the observation $x_t$ given the latent state $s_t$ is $j$) parameterized by $\alpha$, and $C$ is a normalizing factor. The set $\{\alpha, Z\}$ consists of model parameters to

be learned using training data. In our supervised setting, the likelihood function is a softmax classifier that is trained along with the rest of the value iteration network.

### Value Iteration

After decoding a patient's mode or latent state ($s_t$) using the SGLM network, we use the inferred latent state along with other available data to build a predictive model of the outcome of the treatment. Given a patient in state $s$, this outcome is called the *value* of the state or the long-term reward associated with the state, where a positive reward corresponds to an improved HDRS score and vice versa. We leverage three sources of information at each time step $t$ to model the value function:

- observations ($x_t$), including image and audio features of the patient's interview video;
- covariates ($c$), comprised of constant features of patient during the treatment (including age, gender, and body mass index or BMI); and
- inferred state ($s_t$), which is the hidden state deriving patient's treatment dynamics.

Let

$$y_t = [x_t, s_t, c],$$

then, $V(y_t)$ is the expected value of the patient treatment, corresponding to the observations, hidden state and covariates at time $t$. In other words,

$$V(y_t) = E[\sum_{i=t}^{T} r_i],$$

where, $T$ is total number of treatment steps and $r_i$ is the instantaneous reward or wellness of the patient at step $i$. From this definition it's clear that $V(y_t)$ corresponds to accumulated reward or long-term return. In our case, $r_i$ is the HDRS in the

corresponding intermediate steps $i < T$. Moreover, we set $r_T = 1$ if the patient is treated and $r_T = 0$ otherwise. In this study, we use a neural network to model the value function parameterized with $\beta$. The value iteration algorithm tries to find the best value network satisfying

$$V_\beta(y_t) = r_t + \gamma V_\beta(y_{t+1}). \tag{5.2}$$

The 0.95 quantile of the expected return $V_\beta(y)$ in weeks 8-11 is then used as our prediction of treatment outcome at the end of the 14th week, and is used to calculate the prediction accuracy.

**Optimization**

Our neural network model uses a forward pass over the time series data to predict the latent states using the transition matrix $Z$ and the supervised likelihood model parameterized by $\alpha$. Learning of the model parameters is achieved by unrolling the model into a neural network and training the resulting network to find a set of states and parameters that gives the best value function parameterized by $\beta$. Training is done end-to-end similar to deep reinforcement learning models.

Defining

$$\Theta = \{Z, \alpha, \beta\}$$

as the parameter set, our SGLM-RL network aims to minimize following loss function:

$$\mathcal{L}(\Theta) = \sum_{t=1}^{T-1} (V_{\beta^{new}}(y_t) - (r_t + V_{\beta^{old}}(y_{t+1})))^2 \ , \tag{5.3}$$

where the dependence $y_t$ on $Z$ and $\alpha$ are omitted for brevity, and the $\beta^{new}$ and $\beta^{old}$ correspond to the updated and the previous values of the network parameters. The

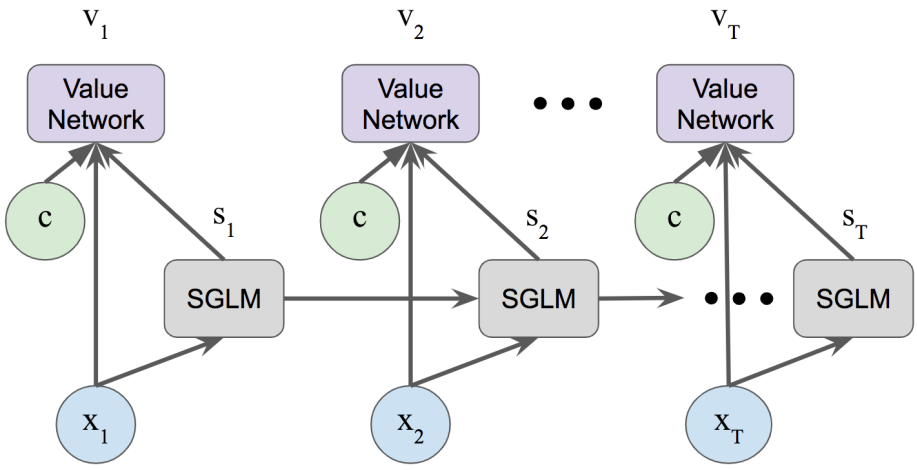


Figure 5.1: The proposed model. SGLM modules estimate the latent state of the patients at different time steps while value networks predict the treatment outcome given the patient state.

overall network parameters can be jointly optimized via gradient descent:

$$\Theta^{new} = \Theta^{old} + \eta \nabla_{\Theta} \mathcal{L}(\Theta), \tag{5.4}$$

where, $\eta$ is the learning rate. With each pass through the observational data, not only will the model learn to better predict the outcome given the patient state, but also the SGLM model learns to better predict the hidden state of the patient at each time point. The overall architecture of the proposed model is depicted in Figure 5.1. Due to the relatively small sample size, we utilize a simple model that includes a 7 state markov model for state estimation (we test $5-10$ states using grid search on a single fold and select 7 states, although the model is not sensitive to this parameter choice) and a single hidden layer neural network for value function approximation with (7 states + 5 covariates) 12 input to 15 hidden states, to 1 output. These parameters are fixed across all subsequent folds to avoid overfitting. Therefore, all models (across all folds) have the same hyperparameters. The only remaining parameter is the regularization constant (*lambda*) that is also selected using grid search ($1e-5$ to 0.1, with optimal value of $1e-4$).

### 5.2.3 Baselines and Performance Measure

We compare our proposed algorithms with the following baselines. To better show the effectiveness of our model we use both temporal (sequential) models and non-temporal (classic) Machine Learning algorithms.

For the baseline temporal models we use the same features fed into our model:

- **LSTM** [83]: This is a recurrent neural network consisting Long Short-Term Memory (LSTM) units which are composed of a cell, an input gate, an output gate and a forget gate. These cells provide an effective way to attend to the right historical data. Comparison with this shows how state-estimation helps improve prediction.

- **Value iteration with LSTM**: This is similar to the proposed approach but the SGLM network is replaced by an LSTM and it's trained end-to-end. This comparison shows how effective our SGLM is compared to the state-of-the-art recurrent modeling method, i.e. LSTM.

For non-temporal methods, we unroll the features over time and form a larger representation.

- **SVM** [84]: Support Vector classifier with linear kernel and LASSO regularization trained via stochastic gradient descent.

- **Decision Tree** [4]: This is a decision tree with Gini's diversity index as split criterion with pruning.

- **Ensemble Learner** [85]: This method is an ensemble method trained via adaptive LogitBoost (Adaptive Logistic Regression) over 100 learning cycles where the weak learners are decision trees. The learning rate for shrinkage of the LogitBoost is set to 1.

For hyperparameter optimization and evaluation purposes cross-validation is typically used, however, Parker *et al.* [86] have shown that when considering the Area Under the Curve (AUC) in small-sample studies, many commonly used cross-validation schemes suffer from significant negative bias. Following Airola *et al.*[87] we use leave-pair-out cross-validation as an approach that provides an almost unbiased estimate of the expected AUC performance. We report the performance of our model based on pooled AUC from a 66-fold leave-pair-out cross-validation study, based on training the model on $N - 2$ patients and testing on the remaining 2, and repeating this process 66 times (or 12 choose 2). All scores were placed in a bucket to calculate the pooled AUC. According to Airola *et al.*[87], this approach leads to a robust measure when the sample size is small.

## 5.3 Results

First we report the effect of different features on the performance of the proposed method. Our hypothesis is that combining vocal, facial, HDRS, and time features provide the best performance. Figure 5.2 demonstrates the Receiver Operating Characteristic (ROC) curves for the full and the individual feature sets. It's apparent that using all the features together outperforms using each of them individually. Using more features leads to a better representation of patient's state and its trajectory over over time, which in turn results in a stronger predictive model.

In order to demonstrate the significance of each feature we iteratively remove a single feature from our feature set and measure the accuracy of the model. As it's shown in figure 5.3, including each of the features that are selected by the feature selection method is is necessary for the model to perform accurately. Moreover, besides HDRS and time, the combination of both audio (e.g., a-03) and video features (e.g., v-35) contribute to achieving higher performance. More details about audio

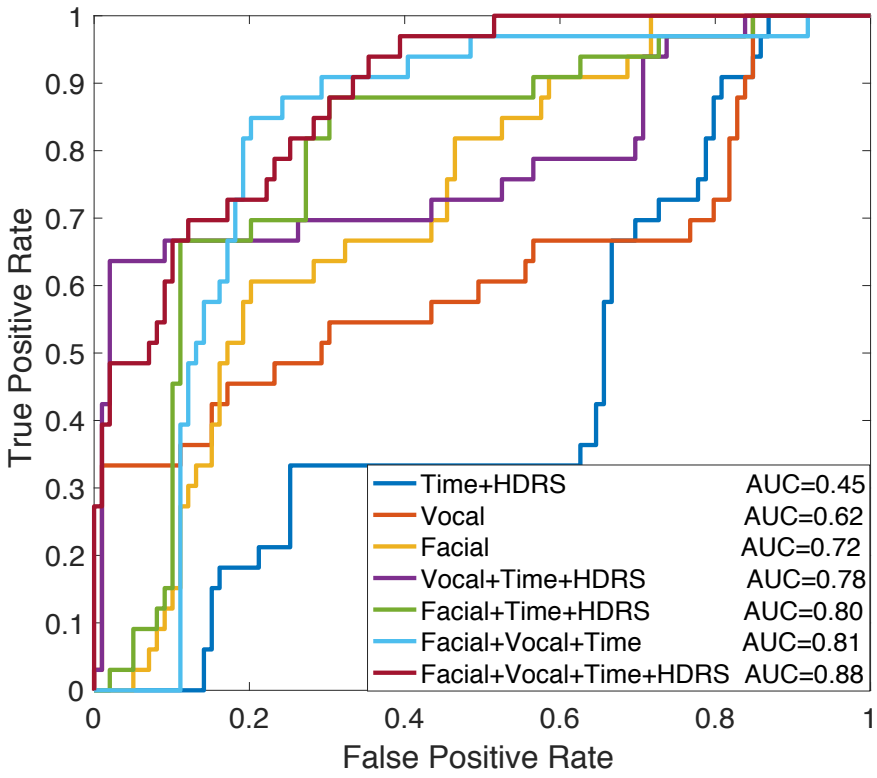


Figure 5.2: Effect of feature set on performance

features can be found in table C.1.

Second, we show that using only a part of the face is not sufficient for our facial variability analysis. More specifically, we partition each face into three areas, *i.e.* upper part that includes forehead, eyebrows and eyes, middle part that includes nose and cheeks, and lower part that covers mouth and chin. Then each time we replace the 12 features corresponding to the MSE of the whole face with the MSE of each part.

Table 5.1: AUC comparison when MSE is calculated only for forehead and eyes (Upper), nose and cheeks (Middle), mouth and chin (Lower), and for the whole face

| **MSE Features** | **Pooled AUC** | **PPV** |
|---|---|---|
| Upper | 0.72 | 0.80 |
| Middle | 0.75 | 0.79 |
| Lower | 0.74 | 0.80 |
| Whole | 0.88 | 0.89 |

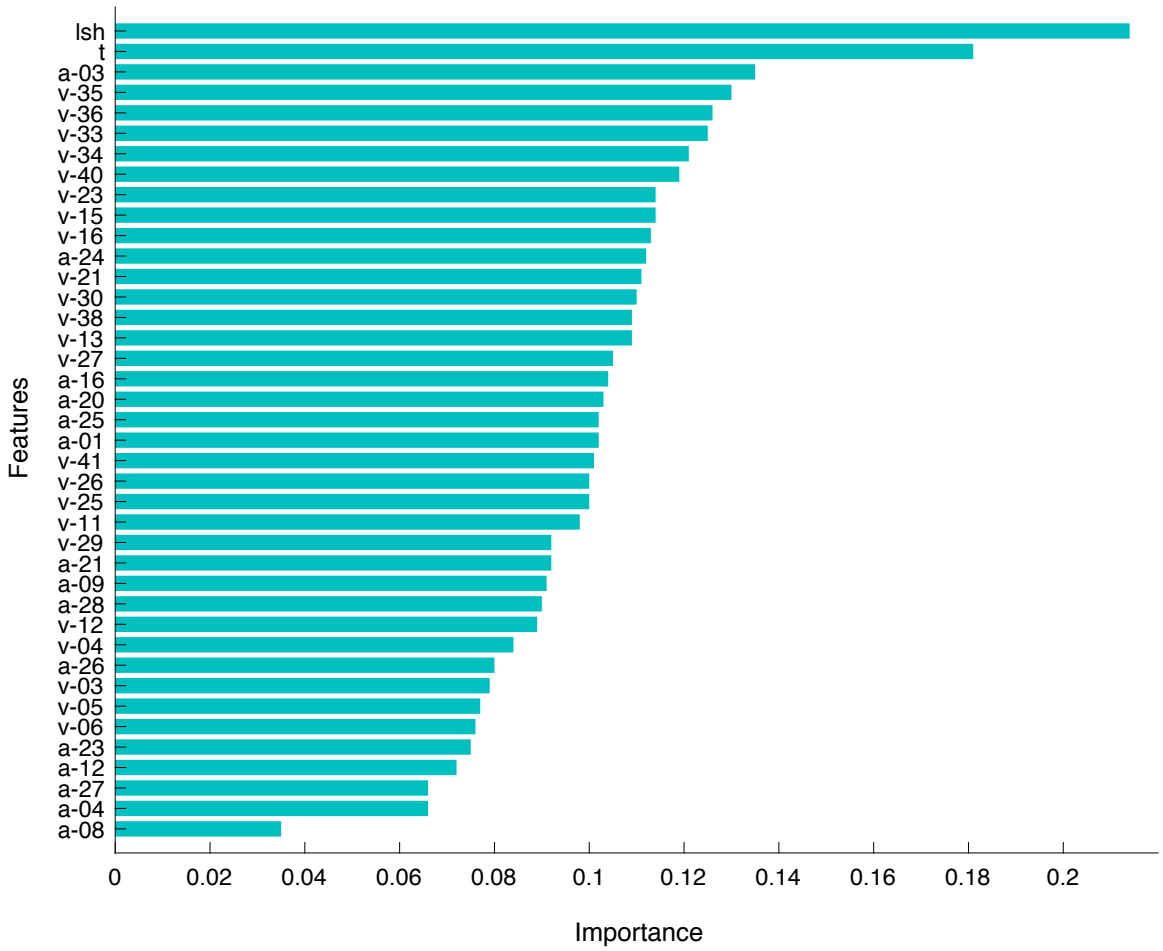


Figure 5.3: Feature importance: The importance is calculated as the decrease in AUC after iteratively removing one feature at a time. Last seen HDRS (lsh) from the previous week, time (t), features in the audio feature set (a-$i$: $i^{th}$), features in the video feature set (v-$i$: $i^{th}$).

Table 5.2 shows the proposed prediction method compared to the baselines in terms of pooled AUC. We used both temporal models and non-temporal models to show that not only the sequential nature of the data should be considered, but also among the temporal methods our proposed approach that combines state-estimation and value iteration outperforms the rest. The inferior performance of LSTM compared to other methods that include value iteration shows that state-estimation and modeling of long-term accumulated reward is essential to drawing a better representation of the recovery status of the patients. Finally, the better performance of (SGLM + value iteration) over (LSTM + value iteration) is likely due to the relative

simplicity of the SGLM model (i.e., smaller number of model parameters) compared to the more complex LSTM network, which tend to overfit on smaller datasets.

Table 5.2: Comparison of AUC of the proposed method and the baselines

| **Non-Temporal Methods** | **Pooled AUC** |
|---|---|
| SVM | 0.70 |
| Ensembled Trees | 0.71 |
| Decission Tree | 0.74 |
| **Temporal Methods** | |
| LSTM | 0.80 |
| LSTM + value-iteration | 0.83 |
| SGLM + value-iteration | **0.88** |

To further investigate the results of the prediction model, we demonstrate the predicted state values for each subject in figure 5.4. It schematically shows how the measures are intuitively compared against each other. The blue curve is our derived measure, which represents the likelihood of the patient improving over the next 3 weeks. The purple horizontal line shows the 95% quantile of the expected estimated value in weeks 8-11. The red curve represents the HDRS measure in each week. When the blue line crosses the purple line it means that our model predicts a highly likely successful trial. It's worth noting that our measure is based on the value function (or return) and the higher value shows a better state of improvement. Firstly, our measure predicts the treatment result 3-4 weeks in advance. Moreover, it shows a more stable and robust estimate in contrast to the HDRS that varies a lot.

## 5.4 Conclusion

In this chapter, we proposed a value iteration-based prediction model for treatment outcomes, when the intermediate assessments of a patient's progress are likely noisy and imprecise. The framework combines the intermediate clinical feedbacks (i.e.,

HDRS) with information from success or failure of a trial to define an aggregated and accumulated learning signal for supervised learning. The resulting value network was able to learn the long-term value associated with a given clinical state. We showed that a feature derived from the learned state values over weeks 8-11 is able to predict the outcome of a DBS trial during the week 14 (i.e., three weeks in advance) with an AUC of 0.88. Such foresight can enable the clinical team to optimize the stimulation parameters, to devise an updated treatment plan, or to simply ignore outlier high HDRS values that are likely to be due to the patient just having a "bad day"; and thus likely not to be correlated with the long-term trajectory of the patient. Our future work includes using model-based RL (which is known to be more data efficient) and multi-task learning (which leverages a correlated set of prediction tasks) to achieve better performance. Other promising research directions include utilization of continuous measures of patient recovery based on wearable devices, and design of more comprehensive reward functions that take into account patient performance metrics measured at different time scales [88].


(a) Subject 2 (b) Subject 3 (c) Subject 4

(d) Subject 5 (e) Subject 6 (f) Subject 7

(g) Subject 8 (h) Subject 9 (i) Subject 10

(j) Subject 11 (k) Subject 12 (l) Subject 13

Figure 5.4: Trajectory of the estimated state value and HDRS for each subject. The weekly clinical scores (red circles; higher values indicate decline) are often noisy and may fluctuate from week to week. The proposed machine learning-based scores (blue diamonds; higher values indicate improvement) are less prone to weekly fluctuations and is able to predict the trajectory of a patient weeks in advance.

# Chapter 6

# Conclusion

The work presented here systematically studied the potential of using machine learning methods to extract biomarkers for clinical state changes during treatment of depression. The models developed here will be useful for both monitoring outcomes and justifying treatment interventions, which may include DBS parameters, making a medication adjustment, or initiating psychotherapy. While some progressive improvement in depression is observed with DBS initiation, after several weeks of DBS patients often enter a transitional phase marked by a return of subjective depressive symptoms, but with preserved emotional reactivity and somewhat heightened negative emotions and affect. Frequently this phase resolves and patients continue on a path of subjective improvement, culminating in stable treatment response or remission from depression. Hence, this recovery course is non-linear, with transient subjective worsening interrupting the improvement trajectory. It is also the case that DBS stimulation parameters are sometimes adjusted during the course of treatment, suggesting that clinicians suspect depressive relapse and make treatment adjustments accordingly. Thus, worsening depression rating scores may represent a transient subjective response that does not require treatment changes, or a disease relapse that does require treatment changes. A more objective biological marker that could dif-

ferentiate between these two states would guide treatment decisions. To this end, and toward the larger goal of establishing more refined clinical markers for depression severity, we focused on computational video and audio analysis for unsupervised (Chapter 2), supervised (Chapters 3 and 4) and reinforcement learning algorithms (Chapter 5). In what follows we review the chapters followed by discussing their findings, implications, limitations, and potential future works.

In Chapter 2, we proposed three unsupervised feature sets to quantify clinical recovery phases of MDD in patients participating in a DBS for depression trial. Experimental validation on video recordings of 12 subjects confirms the power of these feature sets to discriminate between three pre-defined clinical phases. Multiscale Entropy is a very simple yet effective feature that can be computed without much computational burden. Dynamical analysis is effective when the subject facial dynamic is itself subject to change due to movement and variations in pose. Observability features are the most consistent of the three feature sets. It seems that by increasing complexity, the features can capture the discrimination more consistently, albeit with a computational cost The analysis presented in that chapter was limited in the sense that unsupervised features are used in an exploratory setting. A more systematic way that, for example, can effectively weight all the features according to their discriminatory power and classify the phases or subjects may be of more practical value. Applying these techniques to all of a subject's videos over time may show evidence of transition points in the recovery process that precede or predict subjective mood improvement. Further, while the features presented here can reliably discriminate between clinical phases in each subject, the pattern distinguishing the phases varies to some extent between subjects. This may be related to the reliance on each subject's dominant mode for analysis, which may be different across videos and across subjects. Development of a predictive marker based on video analysis would have to address this source of variability. In addition, detailed investigation of the SLDS modes and

entropy-based features and correlation with discrete features of facial expressivity remain as future directions of this work.

Chapter 3 leveraged the extracted features on the previous chapter to build predictive models that can generalize from limited observed labeled data, *i.e.* patient videos and associated depression severity score, and predict the depression severity level of unseen new patients. Our elastic net ordinal regression model outperforms the alternatives in prediction accuracy. This shows the significance of ordinal modeling of the severity classes and importance of elastic net terms on regularizing the space of feasible solutions. The former holds because the severity levels are actually ordered classes of depression rating scores, while the latter is effective because of the small number of data points compared to the dimensionality of the data points. Generalizing these findings to a wider clinical population is limited both by the relatively small number of subjects included here, as well as their uniqueness as a clinical population. For our specific target population (patients with treatment-resistant depression undergoing DBS treatment) and purpose (analyzing facial expressivity) the number of patients included here makes for a strong preliminary study that calls for including a greater number and greater clinical variability of patients for future work. In addition, eigenvalues and eigenvectors are not convenient for clinicians for translation into facial cues that convey information about the patients' facial dynamics unless they can be transformed into physical, observable phenomena. Extracting more interpretable features will be an interesting future work. However, predictive models can bridge the gap between unsupervised computer learning and clinical decision making by showing the utility of such features within a classification system meaningful to clinicians.

In Chapter 4, we tackled a depression severity classification problem from audio signals by accounting/adjusting for short-term emotions when assessing long-term changes in mood or depression state. This way we complemented the visual features

by adding properties extracted from patient's voice. Our proposed model benefits from both complexity of recurrent neural networks in extracting higher level representations and the robustness of SVM in handling small and sparse training data. Our preliminary results call for systematic study of using emotion features to build depression predictors of MDD patients. We acknowledge that the experimental result on only limited patients may not be extensible to the broader MDD patient population and leaves the large scale experiments for future works. Furthermore, the imbalanced emotion dataset needs to be addressed in the future analysis. Formulating the problem as transfer learning and building an end-to-end neural network are interesting directions for future work and explored some of them in the subsequent chapter.

In Chapter 5, we worked with both audio and visual features and proposed a value iteration-based prediction model for treatment outcomes, when the intermediate assessments of a patient's progress are likely noisy and imprecise. The framework combines the intermediate clinical feedbacks with information from success or failure of a trial to define an aggregated and accumulated learning signal for supervised learning. The resulting value network was able to learn the long-term value associated with a given clinical state. This foresight can enable the clinical team to optimize the stimulation parameters and to devise an updated treatment plan. The future work includes using model-based RL (which is known to be more data efficient) and multi-task learning (which leverages a correlated set of prediction tasks) to achieve better performance. Other promising research directions include utilization of continuous measures of patient recovery based on wearable devices.

Overall, we start with simple exploratory and unsupervised methods to find meaningful visual features explaining and distinguishing different phases. These features proved to be use full in building prediction models. Given these promising results on the visual features we improved our predictive models with audio features. With the

premise of audio-visual features on depression severity classification we expand our study to tackle the long-term prediction problem where reinforcement learning algorithms are leveraged to model the trade-off between short-term reward and long-term return of the treatment. We covered a spectrum of algorithms from unsupervised, to supervised to reinforcement learning methods. We used simple yet effective linear regression models who are trained using least squares methods, as well as complicated deep neural network models which are trained end-to-end using gradient based methods. We also showed that a variety of data sources, such as images or audios, may be used to build the machine learning models.

These results suggest that machine learning models may lead to quantitative biomarkers and predictive models (both short- and long-term) of depression states during the course of treatment. This may have a broader impact in terms of cost effectiveness, and resource allocation. The success of this work calls for more interdisciplinary research at the intersection of psychiatry, computer science, and machine learning.

# Appendix A

# Switching Linear Dynamical Systems

Switching Linear Dynamical Systems (SLDSs) are powerful and expressive and are capable of modeling physical processes governed by state equations that may switch its behavior from time to time (ie, "mode switching"). They model event sequences using the evolution of 2 hidden layers of states. The top layer is governed by a Markov Chain with $J$ modes. It evolves according to $J \times J$ transition probability matrix $\mathbf{Z}$ in discrete time steps. A $J \times 1$ vector $\pi$ specifies the initial distribution of modes.

## A.1 Modeling

Let variables

$$s_t \in \{1, \ldots, J\}$$

show the mode at time step $t$. The bottom layer of states are $D$ dimensional state variables, $\mathbf{x}_t \in \mathbb{R}^D$, evolving according to a LDS whose characteristic is determined by the mode (from the top layer). Let $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(J)}$ be different state dynamics

associated to latent modes, then,

$$\mathbf{x}_t = \mathbf{A}^{(s_t)}\mathbf{x}_{t-1} + \mathbf{v}_t \tag{A.1}$$

where $v_t$ is a noise process (Gaussian with zero mean covariance $\mathbf{Q}^{(s_t)}$). Figure A.1 schematically demonstrates how the observations, states and modes evolve over time steps in a SLDS.

To complete the generative model, let the latent state $\mathbf{x}_t$, generate observation $\mathbf{y}_t$ at time $t$. Observations are assumed to come from a $M$-dim space, *i.e.*, $\mathbf{y}_t \in \mathbb{R}^M$. The dynamical system produces observed variables according to

$$\mathbf{y}_t = \mathbf{C}^{(s_t)}\mathbf{x}_t + \mathbf{w}_t, \tag{A.2}$$

where the noise $w_t$ on observation is a Gaussian distribution with zero mean and covariance $\mathbf{R}^{(s_t)}$.

## A.2 System Identification

The purpose of system identification is to learn the unknown parameters and latent variables of the generative model using the observed variables, noisy images (videos) in our case. We have two types of model parameters:

1) parameters associated to dynamical systems $\mathbf{A}^{(j)}$, $\mathbf{Q}^{(j)}$,$\mathbf{C}^{(j)}$, and $\mathbf{R}^{(j)}$ for different modes $1 \leq j \leq J$;

2) parameters associated to mode switches $\mathbf{Z}$ and $\pi$.

We utilized memory-efficient and fast implementation of the expectation maximization algorithm [89] that iterates between inferring the switching parameters using an approximate inference algorithm and learning the parameters of each dynamical system (or modes) using another approximate algorithm. These approximation al-

gorithms let us learn parameters and latent variables of the model on a standard machine in a few minutes.

## A.3 Experimental Setup

Here we discuss the settings of our experiments. The number of modes is set to 3. The latent state and observation dimensions are set to 20 and 900 respectively. Note that 900 is the number of pixels in resized images ($900 = 30 \times 30$). After performing down-sampling we used 1000 frames of the video recordings . It roughly corresponds to the first 5 minutes of the video recordings. The inference usually converges after 3-5 iterations. The hyperparameters (e.g. number of modes) are selected based on the best performance. SLDS with 3 modes worked better than a system with 2 modes. For more than 3 modes, we haven't observed a significantly different result. Regarding to the number of significant eigenvalues, the magnitude of the eigenvalues usually drops significantly after 10. A smaller number of eigenvalues would have insufficient information while a larger number could be misleading because the magnitude of the eigenvalues drop significantly towards the size of the matrix (*i.e.* 20) and make the result unstable. Finally, we have used Bayesian optimization to find the optimum state space dimension among the values in the set $\{10, 20, 30, 40, 50\}$.

## A.4 Latent Dynamical Analysis

We performed an evaluation and analysis on the most dominant mode, ie, the inferred mode that includes the most video frames or mathematically,

$$argmax_j \sum_t Prob(S_t = j).$$

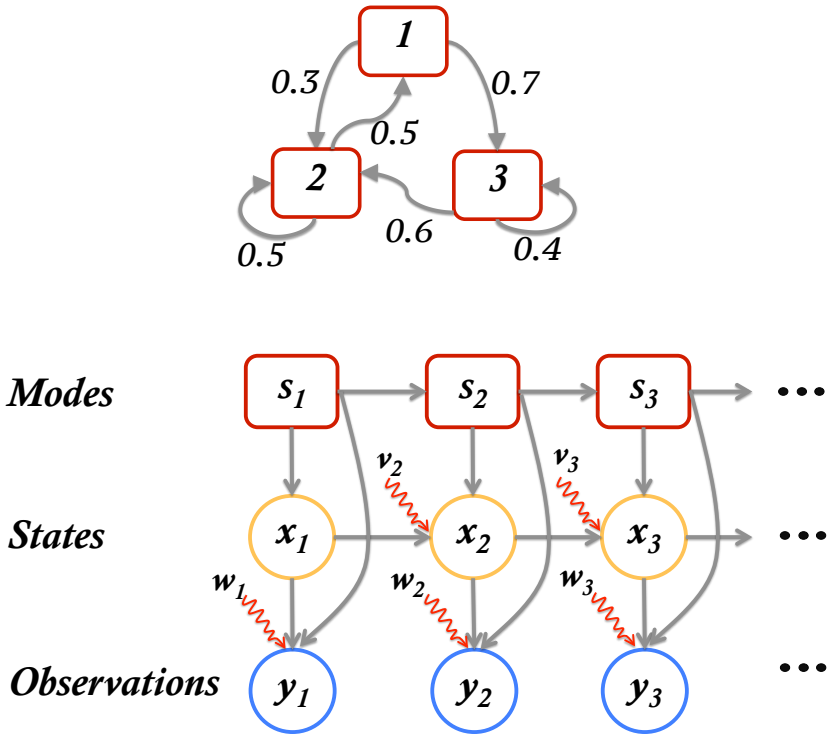


Figure A.1: Graphical representation of the SLDS; Top) probability transitions between the three modes; Bottom) evolving SLDS over time

The dynamical behavior of the video recordings is determined by eigenvalues (or spectral properties) of the state transition matrix **A**. The 10 largest eigenvalues (in magnitude) are used as a representation the dynamical properties of the latent state. Eigenvalues with higher absolute real values indicate a smoother state transitions, while smaller values show significant variations, that's why they are useful for discriminating phases and distinguishing between them.

### Observability Analysis

Linear dynamical systems are characterized by the system matrices $C$ and $A$. The linear dynamical system is said to be *observable* if and only if the observability matrix, defined as

$$[C^\top (CA)^\top (CA^2)^\top \dots (CA^{D-1})^\top],$$

is full rank (*i.e.*, of rank $M$) [54]. This matrix is the key criterion for success in recovering the state sequence from the common measurement values and its spectral properties can be used to distinguish different phases for a patient. Thus the Singular Values (SVs) of observability matrix are more comprehensive than the eigenvalue features (which only describe the properties of the state transition dynamics matrix $A$). Similar to the analysis of dynamical properties, we used the most dominant mode of the inferred model for each video to construct the observability matrix and extracts its singular values.

For SLDS-based analysis, both dynamics within individual facial expressions and the dynamics between facial expressions are utilized in our approach. First, the dynamics between expressions is used to unroll the modes and switches between them. Then, we use the spectral properties of the dynamics within the expression to extract features for prediction.

Let $T$ be the length of the time series (in our case $T = 1000$), $J$ be the number of modes (in our case $J = 3$), and $D$ be the dimension of the latent space (in our case $D = 20$), then the SLDS inference take $O(TJ^2D^3)$ time complexity [89]. Computing eigenvalues is of $O(D^3)$. For the prediction task, if $d$ is the dimension of extracted features (in our case $d = 44$) and $n$ is the number of data points ($n = 11$) then every iteration of gradient descent is of $O(nd)$.

# Appendix B

# Elastic Net Ordinal Logistic Regression

Assume the feature space $\mathcal{X} \subset \mathbb{R}^d$ and output space contains $K$ classes: $\mathcal{Y} = \{1, 2, \ldots, K\}$. In ordinal logistic regression the cumulative probability is modeled as the logistic function:

$$P(y \leq k|x) = \phi(\theta_k - w^\top x) = \frac{1}{1 + \exp(w^\top x - \theta_k)}, \tag{B.1}$$

where,

$$w = (w_1, \ldots, w_d) \in \mathbb{R}^d$$

is the coefficient vector, and

$$\theta = (\theta_1, \ldots, \theta_{K-1}) \in \mathbb{R}^{K-1}$$

is the threshold vector, and

$$\phi(t) = 1/(1 + \exp(-t))$$

is the logistic function. $\theta$ is a non-decreasing vector *i.e.*,

$$\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{K-1}.$$

By defining $\theta_0 = -\infty$ and $\theta_K = \infty$ we will then assign the class $k$ if the linear prediction model $x^\top w$ lies in the interval $[\theta_{k-1}, \theta_k)$. We are interested in a vector $w$ such that $x^\top w$ generates a set of values that are well distributed in the different classes using the different thresholds $\theta$.

Having observed a dataset of $n$ samples

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

the log-likelihood is

$$\mathcal{L}(w, \theta) = \sum_{i=1}^{n} \log(\phi(\theta_{y_i} - w^\top x_i) - \phi(\theta_{y_i - 1} - w^\top x_i)) \tag{B.2}$$

To enhance generalizability and prediction capability of the model, we extend the ordinal logistic regression by incorporating elastic net regularization [55]. Elastic net is a combination of LASSO (least absolute shrinkage and selection operator) and ridge regression. Similar to LASSO regularization, elastic net results in sparse solutions, however it also has the advantage of performing well with highly correlated variables. Therefore we define our new objective function as

$$\mathcal{F}(w, \theta) = \min_{w,\theta} -\mathcal{L}(w, \theta) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 \tag{B.3}$$

where $\lambda_1$ and $\lambda_2$ are regularization parameters chosen via cross validation. The gra-

dient of the objective function is computed as:

$$\begin{aligned}\nabla_\theta \mathcal{F}(w,\theta) = \sum_{i=1}^{n} e_{y_i}(1-\phi(\theta_{y_i}-x_i^\top w) - \frac{1}{1-\exp(\theta_{y_i-1}-\theta_{y_i})} \\ + e_{y_i-1}(1-\phi(\theta_{y_i-1}-x_i^\top w) - \frac{1}{1-\exp(\theta_{y_i-1}-\theta_{y_i})}\end{aligned} \tag{B.4}$$

where $e_i = (0,\ldots,1,\ldots,0)$ is the canonical vector where only the $i$-th location is 1. Also,

$$\begin{aligned}\nabla_w \mathcal{F}(w,\theta) = \sum_{i=1}^{n} x_i(1-\phi(\theta_{y_i}-x_i^\top w) - \phi(\theta_{y_i-1}-x_i^\top w)) \\ + 2\lambda_2 w + \lambda_1\, sign(w),\end{aligned} \tag{B.5}$$

where in $sign(w)$ the sign operation is performed element wise. Given the analytical form of the function and its gradient the well-known Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [90] is used to find the optimum parameter minimizing the regularized loss.

# Appendix C

# Audio Features

Table C.1: Audio features. *ANG: angry, EXC: excited, NEUT: neutral, var: variance, vrb: variability

| Feature ID | Feature Name |
|---|---|
| a-01 | min(ANG*) |
| a-02 | min(EXC*) |
| a-03 | min(NEUT*) |
| a-04 | min(SAD) |
| a-05 | max(ANG) |
| a-06 | max(EXC) |
| a-07 | max(NEUT) |
| a-08 | max(SAD) |
| a-09 | mean(ANG) |
| a-10 | mean(EXC) |
| a-11 | mean(NEUT) |
| a-12 | mean(SAD) |
| a-13 | var*(ANG) |
| a-14 | var(EXC) |
| a-15 | var(NEUT) |
| a-16 | var(SAD) |
| a-17 | skew(ANG) |
| a-18 | skew(EXC) |
| a-19 | skew(NEUT) |
| a-20 | skew(SAD) |
| a-21 | Kurt(ANG) |
| a-22 | Kurt(EXC) |
| a-23 | Kurt(NEUT) |
| a-24 | Kurt(SAD) |
| a-25 | vrb*(ANG) |
| a-26 | vrb(EXC) |
| a-27 | vrb(NEUT) |
| a-28 | vrb(SAD) |

# Bibliography

[1] Danilo Bzdok and Andreas Meyer-Lindenberg. Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230, 2018.

[2] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997. ISBN 9780071154673. URL `https://books.google.com/books?id=EoYBngEACAAJ`.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[4] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[5] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9):1426–1448, 2019.

[6] Wajid Mumtaz, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, and Aamir Saeed Malik. A machine learning framework involving eeg-based functional connectivity to diagnose major depressive disorder (mdd). *Medical & biological engineering & computing*, 56(2):233–246, 2018.

[7] João R Sato, Jorge Moll, Sophie Green, John FW Deakin, Carlos E Thomaz, and Roland Zahn. Machine learning algorithm accurately detects fmri signature

of vulnerability to major depression. *Psychiatry Research: Neuroimaging*, 233 (2):289–291, 2015.

[8] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag, and Rosalind Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 325–332. IEEE, 2017.

[9] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[10] James R Williamson, Diana Young, Andrew A Nierenberg, James Niemi, Brian S Helfer, and Thomas F Quatieri. Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language*, 55:40–56, 2019.

[11] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 43–50. ACM, 2016.

[12] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *Journal of the American Medical Association*, 289(23): 3095–3105, 2003.

[13] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[14] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[15] Christina Sobin and Harold A Sackeim. Psychomotor symptoms of depression. *The American Journal of Psychiatry*, 154(1):4, 1997.

[16] John F Greden, Nancy Genero, H Laurence Price, Michael Feinberg, and Simon Levine. Facial electromyography in depression: Subgroup differences. *Archives of General Psychiatry*, 43(3):269–274, 1986.

[17] Aaron T Beck, C Ward, M Mendelson, J Mock, and J Erbaugh. Beck depression inventory (bdi). *Arch Gen Psychiatry*, 4(6):561–571, 1961.

[18] Kurt Kroenke and Robert L Spitzer. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515, 2002.

[19] Yossef S Ben-Porath. Assessing personality and psychopathology with self-report inventories. *Handbook of psychology*, pages 553–577, 2003.

[20] A John Rush, Madhukar H Trivedi, Stephen R Wisniewski, Andrew A Nierenberg, Jonathan W Stewart, Diane Warden, George Niederehe, Michael E Thase, Philip W Lavori, Barry D Lebowitz, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a star* d report. *American Journal of Psychiatry*, 163(11):1905–1917, 2006.

[21] Helen S Mayberg, Andres M Lozano, Valerie Voon, Heather E McNeely, David Seminowicz, Clement Hamani, Jason M Schwalb, and Sidney H Kennedy. Deep brain stimulation for treatment-resistant depression. *Neuron*, 45(5):651–660, 2005.

[22] Andrea L Crowell, Steven J Garlow, Patricio Riva-Posse, and Helen S Mayberg. Characterizing the therapeutic response to deep brain stimulation for treatment-resistant depression: a single center long-term perspective. *Frontiers in Integrative Neuroscience*, 9, 2015.

[23] Alexander R Daros, Anthony C Ruocco, and Nicholas O Rule. Identifying mental disorder from the faces of women with borderline personality disorder. *Journal of Nonverbal Behavior*, 40(4):255–281, 2016.

[24] Naomi Jane Scott, Robin Stewart Samuel Kramer, Alex Lee Jones, and Robert Ward. Facial cues to depressive symptoms and their associated personality attributions. *Psychiatry Research*, 208(1):47–53, 2013.

[25] Anastasia Pampouchidou, Panagiotis Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 2017.

[26] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

[27] Anastasia Pampouchidou, Kostas Marias, Manolis Tsiknakis, P Simos, Fan Yang, and Fabrice Meriaudeau. Designing a framework for assisting depression severity assessment from facial image analysis. In *Signal and Image Processing Applications (ICSIPA), 2015 IEEE International Conference on*, pages 578–583. IEEE, 2015.

[28] Rajamannar Ramasubbu, Matthew RG Brown, Filmeno Cortese, Ismael Gaxiola,

Bradley Goodyear, Andrew J Greenshaw, Serdar M Dursun, and Russell Greiner. Accuracy of automated classification of major depressive disorder as a function of symptom severity. *NeuroImage: Clinical*, 12:320–331, 2016.

[29] Kacem Anis, Hammal Zakia, Daoudi Mohamed, and Cohn Jeffrey. Detecting depression severity by interpretable representations of motion dynamics. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 739–745. IEEE, 2018.

[30] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 22(2):525–536, 2018.

[31] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[32] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[33] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 147–152. IEEE, 2013.

[34] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender. *Journal on Multimodal User Interfaces*, 9(1):17–29, 2015.

[35] Sayan Ghosh, Moitreya Chatterjee, and Louis-Philippe Morency. A multimodal context-based approach for distress assessment. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 240–246. ACM, 2014.

[36] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing*, 32(10):641–647, 2014.

[37] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, pages 1–13, 2015.

[38] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[39] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinicianpatient interaction. *Applied and Preventive Psychology*, 1995.

[40] J. K. Darby and H. Hollien. Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica*, 29(4):279–291, 1977.

[41] L. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *ICASSP*, pages 5154–5157. IEaEE, 2010.

[42] N. Cummins, J. Epps, and E. Ambikairajah. Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013.

[43] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.

[44] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10. ACM, 2016.

[45] E. Moore, M. Clements, J. Peifer, and L. Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 2008.

[46] B. Stasak, N Cummins J Epps, and R Goecke. An investigation of emotional speech in depression classification. In *Interspeech*, 2016.

[47] Sahar Harati, Andrea Crowell, Helen Mayberg, Jun Kong, and Shamim Nemati. Discriminating clinical phases of recovery from major depressive disorder using the dynamics of facial expression. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2254–2257. IEEE, 2016.

[48] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[49] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[50] Carlo Tomasi and T Kanade Detection. Tracking of point features. Technical report, Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, 1991.

[51] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[52] Paul Ekman. Expression and the nature of emotion. *Approaches to emotion*, 3: 19–344, 1984.

[53] Madalena Costa, Ary L Goldberger, and C-K Peng. Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters*, 89(6):068102, 2002.

[54] Aswin C Sankaranarayanan, Pavan K Turaga, Rama Chellappa, and Richard G Baraniuk. Compressive acquisition of linear dynamical systems. *SIAM Journal on Imaging Sciences*, 6(4):2109–2133, 2013.

[55] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[56] Shigeru Morishita and Seizaburo Arita. Possible predictors of response to fluvoxamine for depression. *Human Psychopharmacology: Clinical and Experimental*, 18(3):197–200, 2003.

[57] Richard C Shelton, Anne C Andorn, Craig H Mallinckrodt, Madelaine M Wohlreich, Joel Raskin, John G Watkin, and Michael J Detke. Evidence for the efficacy of duloxetine in treating mild, moderate, and severe depression. *International Clinical Psychopharmacology*, 22(6):348–355, 2007.

[58] Max A Little, Gael Varoquaux, Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, 6(5):gix020, 2017.

[59] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[60] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[61] Kellie J Archer, Jiayi Hou, Qing Zhou, Kyle Ferber, John G Layne, and Amanda E Gentry. ordinalgmifs: An r package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, 13:187, 2014.

[62] Michael J Wurm, Paul J Rathouz, and Bret M Hanlon. Regularized ordinal regression and the ordinalnet r package. *arXiv preprint arXiv:1706.05003*, 2017.

[63] Bo Chen, Rui Castro, and Andreas Krause. Joint optimization and variable selection of high-dimensional gaussian processes. *arXiv preprint arXiv:1206.6396*, 2012.

[64] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[65] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*. IEEE, 2013.

[66] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[67] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[68] T. Giannakopoulos and S. Petridis. Fisher linear semi-discriminant analysis for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):1913–1922, 2012.

[69] T. Giannakopoulos and A. Pikrakis. *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014.

[70] T. Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610, 2015.

[71] C Busso, M Bulut, C Lee, A Kazemzadeh, E Mower, S Kim, J Chang, and S Narayanan S Lee. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335, 2008.

[72] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[73] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers. Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning. *arXiv preprint arXiv:1708.03920*, 2017.

[74] Fei Tao and Gang Liu. Advanced lstm: A study about better time dependency modeling in emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910. IEEE, 2018.

[75] Vladimir Chernykh and Pavel Prikhodko. Emotion recognition from speech with recurrent neural networks. *arXiv preprint arXiv:1701.08071*, 2017.

[76] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In

*2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[77] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2011.

[78] Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.

[79] Sahar Harati, Andrea Crowell, Helen Mayberg, and Shamim Nemati. Depression severity classification from speech emotion. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5763–5766. IEEE, 2018.

[80] Amy H Farabaugh, Stella Bitran, Janet Witte, Jonathan Alpert, Sarah Chuzi, Alisabet J Clain, Lee Baer, Maurizio Fava, Patrick J McGrath, Christina Dording, et al. Anxious depression and early changes in the hamd-17 anxiety-somatization factor items and antidepressant treatment outcome. *International Clinical Psychopharmacology*, 25(4):214, 2010.

[81] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[82] Shamim Nemati, H Lehman Li-wei, and Ryan P Adams. Learning outcome-discriminative dynamics in multivariate physiological cohort time series. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7104–7107. IEEE, 2013.

[83] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[84] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[85] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[86] Brian J Parker, Simon Günter, and Justin Bedo. Stratification bias in low signal microarray studies. *BMC bioinformatics*, 8(1):326, 2007.

[87] Antti Airola, Tapio Pahikkala, Willem Waegeman, Bernard De Baets, and Tapio Salakoski. A comparison of auc estimators in small-sample studies. In *Machine learning in systems biology*, pages 3–13, 2009.

[88] Erik Reinertsen, Supreeth P Shashikumar, Amit J Shah, Shamim Nemati, and Gari D Clifford. Multiscale network dynamics between heart rate and locomotor activity are altered in schizophrenia. *Physiological measurement*, 39(11):115001, 2018.

[89] Shamim Nemati and Mohammad M Ghassemi. A fast and memory-efficient algorithm for learning and retrieval of phenotypic dynamics in multivariate cohort time series. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 41–44. IEEE, 2014.

[90] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.