Addressing the Credit Assignment Problem in Treatment Outcome Prediction using Temporal Difference Learning

Sahar Harati†

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
†*E-mail: harati@stanford.edu*

Andrea Crowell

Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA, USA
E-mail: andrea.crowell@emory.edu

Helen Mayberg

Center for Advanced Circuit Therapeutics, Mount Sinai, New York City, NY, USA
E-mail: helen.mayberg@mssm.edu

Shamim Nemati

Department of Biomedical Informatics,UC San Diego Health, San Diego, CA, USA
E-mail: shamim.nemati@alum.mit.edu

Mental health patients often undergo a variety of treatments before finding an effective one. Improved prediction of treatment response can shorten the duration of trials. A key challenge of applying predictive modeling to this problem is that often the effectiveness of a treatment regimen remains unknown for several weeks, and therefore immediate feedback signals may not be available for supervised learning. Here we propose a Machine Learning approach to extracting audio-visual features from weekly video interview recordings 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 patient's response and the delayed nature of feedbacks. Our results based on longitudinal recordings from 12 patients with depression show that the learned state values are predictive of the long-term success of DBS treatments. We achieve an area under the receiver operating characteristic curve of 0.88, beating all baseline methods.

Keywords: Machine Learning; Temporal Difference Learning; Depression.

1. Introduction

Major Depressive Disorder (MDD) is a common psychiatric illness, but unfortunately lacks an objective, non-verbal, automated biomarker that can reliably predict treatment outcome. These patients often suffer from psychomotor slowing and limited emotional reactivity and affective range. Patients with MDD are diagnosed with treatment-resistant depression (TRD) according to non-response or poor response to standard antidepressant treatments. Deep brain stimulation (DBS) of the subcallosal cingulate cortex is a promising treatment for TRD patients.[1] It has been reported that increased facial expressivity and other psychomotor factors are correlated with improvement in patients who have received DBS.[2]

Response to treatment is typically measured with the Hamilton Depression Rating Scale (HDRS),[3] a standardized clinician-administered measure based on the patient's self-report and the current gold standard for measuring treatment response in depression studies. Recovery from DBS treatment is usually non-linear, with transient subjective worsening sometimes interrupting the improvement trajectory.[2] Automatically predicting the outcome of DBS treatment has several useful implications for clinical management of TRD patients, such as optimization, stopping, or continuation of the ongoing treatment. Furthermore, being aware of the stability and time course of patient's state during the recovery process has the potential to help clinical trial design. In this paper, we propose an automated biomarker of patient progress based on vocal and facial dynamics that can serve as an early predictor of DBS treatment outcomes.

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.[4–6] For example, Darby *et al.*[7] reported a quantifiable change in the pitch, speaking rate, loudness, and articulation of depressed patients before and after treatment. Harati *et al.*[8] 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.*[9] 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[10] or using accepted clinical classification of severity: Pampouchidou, *et al*[11] achieved 55% accuracy , Ramasubbu, *et al*[12] reported 52-66% accuracy , Anis, *et al*[13] achieved an accuracy of 66%, and Dibeklioglu, *et al*[14] reached an accuracy of 66-84%. Recently, Harati *et al.*[15] explored the use of video analysis of facial expressivity in a cohort of MDD patients before and after DBS, to propose a feature set and leveraged them to build predictive models for depression.

Our proposed model based on Machine Learning is an extension to our previous work [15] and it differs from the other works that are mentioned above in four ways. First, both audio and visual features are considered and combined to achieve improved prediction accuracy. Second, we utilize the framework of temporal difference (TD) learning to model the sequential nature of treatment and assessment/feedback delay over course of time. Third, we use state-estimation to infer the hidden state of the patient over time, thus exploiting the temporal information embedded in longitudinal patient recordings. Fourth, the proposed deep neural

network structure for state-estimation and outcome prediction (via value iteration) is trained end-to-end using gradient descent optimization. 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.

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 for future prediction by showing that credit assignment via back-propagation allows us to train the model without immediate feedback. We learn from the accumulated rewards rather than 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/video 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. Figure 1 shows the overall architecture of the proposed method to be elaborated later in this paper.

The rest of this manuscript is organized as follows. Section 2 introduces the DBS dataset that was used to develop the proposed method. Section 3 describes the utilized features, proposed prediction and state-estimation modules, and the baseline techniques. We present the experimental results in Section 4, and finally the paper is concluded in Section 5 with discussions and several directions for future works and extensions.

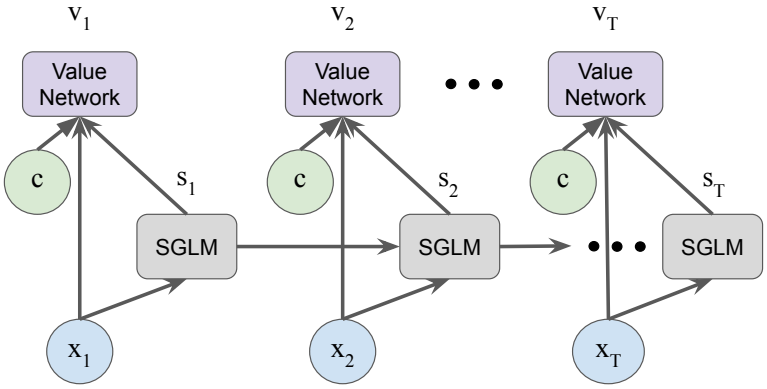

Fig. 1. The proposed model. SGLM modules estimate the latent state of the patients at different time steps while value networks predict the treatment outcome given the patient state.

2. Dataset

We have collected our audio-visual data in an ongoing TRD DBS study performed at Emory University[a]. A cohort of 12 TRD patients were evaluated weekly by study psychiatrists starting 1 month before DBS surgery and throughout the first 6 months of chronic stimulation. 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

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

the full 7 months for each patient. The videotaped interviews document the evolution of DBS treatment and clinical response. The subjects included in the analysis were primarily Caucasian (one African American patient) and female (two male participants) and were aged between 35 and 68. Treatment response was measured with the Hamilton Rating Scale for Depression (HDRS),[3] which is based on the patient's self-report of their depression symptom relative to the previous week and is the gold standard assessment tool for depression. According to Crowell *et al.*,[2] 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 after 6 month, resulting in nine improved and three depressed subjects.

All of the data gathering and analytic procedures were approved by the Emory University Institutional Review Board (IRB).

3. Methods

First, we introduce the features that are being extracted from videos of patients, including visual and audio features. Second, the temporal difference learning method in association with a parametric value-network is presented to estimate the long-term value of each patient state. Finally, we present the baseline methods and measures with which we compare our method.

3.1. *Feature Extraction*

In this paper, we used both audio and visual features to test the hypothesis that fusion of multimodal data can improve prediction accuracy.

For the audio features we used the same technique described in a previous work by our group,[8] 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. Due to the lack of training data on depression, the LSTM network is pre-trained on the Interactive emotional dyadic motion capture (IEMOCAP)[16] dataset with replicated architecture used in other studies. For pre-training, we used a stacked LSTM network[17] that has two hidden LSTM layers. 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 IEMOCAP 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. 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. Finally, per emotion, seven statistics over all utterances in each interview session is 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.[8]

For the visual features, we leveraged another previous work to extract facial features for MDD subjects[15] that are shown to be effective in distinguishing the recovery phases of DBS patients during treatment. Briefly, first the images are put through face detection, contrast normalization, and image registration and alignment. Then, three types of dynamical features are extracted using a Multiscale Entropy (MSE) and Switching Linear Dynamical Systems (SLDS) approach. MSE measures the randomness or unpredictability exists in a sequence of patient's facial expression. We used scales from 1 to 12 to get 12 features, and calculated the average entropy across all the video pixels. Second, a 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 significant eigenvalues (or spectral properties) of the state transition matrix of the most dominant dynamical mode were used as another variability feature set. Third, the top 15 largest singular values of the observability matrix were taken for a comprehensive coverage of dynamical behavior of facial expression. These led to an overall feature set size of 42. For more details please refer to our previously published work .[15] In summary, we extracted 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.

3.2. *Temporal Difference Learning*

We developed a TD-learning approach that leverages the value iteration algorithm to predict treatment outcome.[18] Given the multivariate time series of features described in the previous section, a Switching Generalized Linear Model (SGLM) [19] was utilized to identify patient-specific clinical states, which was then fed into the value iteration network to assess the long-term value of a given clinical state. The overall model was optimized end-to-end as described next.

3.2.1. *State-Estimation*

In order to track the treatment process, we first identified 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 chose a supervised approach to hidden state-estimation (known as the SGLM model) under the assumption of Markovianity and a linear state transition model.[19] In the top layer, there were J possible hidden states (or *modes*), and the likelihood function of states takes the form of a softmax classifier with parameter α; mapping the observations to the likelihood of the J latent states. The network used 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

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

where, $P\left(s_t = j | \{\mathbf{x}_{1:t}\}\right)$ 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 α, 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 was a softmax classifier that was trained along with the rest of the value iteraton network.

3.2.2. *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 β. The value iteration algorithm tries to find the best value network satisfying

$$V_\beta(y_t) = r_t + \gamma V_\beta(y_{t+1}). \tag{2}$$

γ is the discount factor to control the importance of future rewards and is set to 1. 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.

3.2.3. *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 α. 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 β. 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:

$$\mathcal{L}(\Theta) = \sum_{t=1}^{T-1} (V_{\beta^{new}}(y_t) - (r_t + \gamma V_{\beta^{old}}(y_{t+1})))^2 \ , \tag{3}$$

where the dependence y_t on Z and α are omitted for brevity, and the β^{new} and β^{old} correspond to the updated and the previous values of the network parameters. The overall network parameters can be jointly optimized via gradient descent:

$$\Theta^{new} = \Theta^{old} + \eta \nabla_{\Theta} \mathcal{L}(\Theta), \tag{4}$$

where, η 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 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 $1e-4$).

3.3. *Baselines and Performance Measure*

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

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

- **LSTM**:[20] 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.

- **SVM** :[21] Support Vector classifier with linear kernel and LASSO regularization trained via stochastic gradient descent.
- **Decision Tree**:[22] This is a decision tree with Gini's diversity index as split criterion.
- **Ensemble Learner**:[23] 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.

- **Neural Network**: A 2-layer neural network is also implemented as a deep learning baseline while not benefiting from temporal information in the data.

For hyperparameter optimization and evaluation purposes cross-validation is typically used, however, Parker et al.[24] 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.[25] we used 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 Airola et al.,[25] this approach leads to a robust measure when the sample size is small. The confidence interval for reported AUC's is calculated using Hanley-McNeil method.[26] The Matlab implementation codes can be found online[b].

4. Results

First we report the effect of different features on the performance of the proposed method. Our hypothesis was that combining vocal, facial, HDRS, and time features provide the best performance. Figure 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 model.

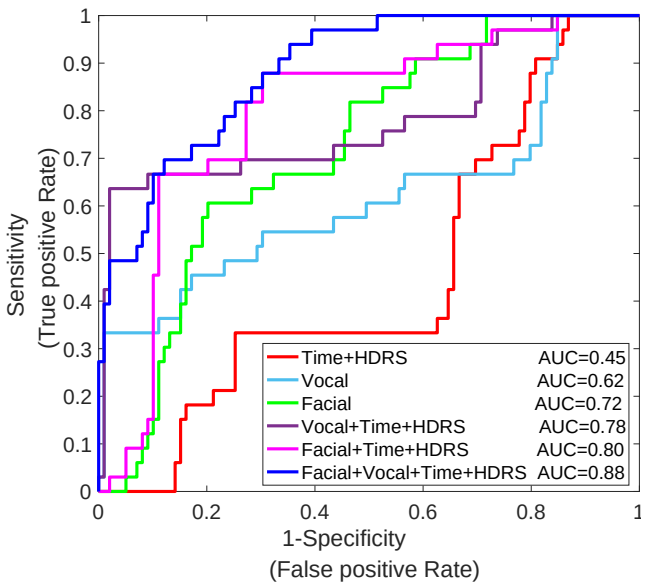

Fig. 2. Effect of feature set on performance

In order to demonstrate the significance of each feature we iteratively remove a single

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

feature from our feature set and measure the accuracy of the model. As it's shown in figure 3, including each of the features that are selected by the feature selection method 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.

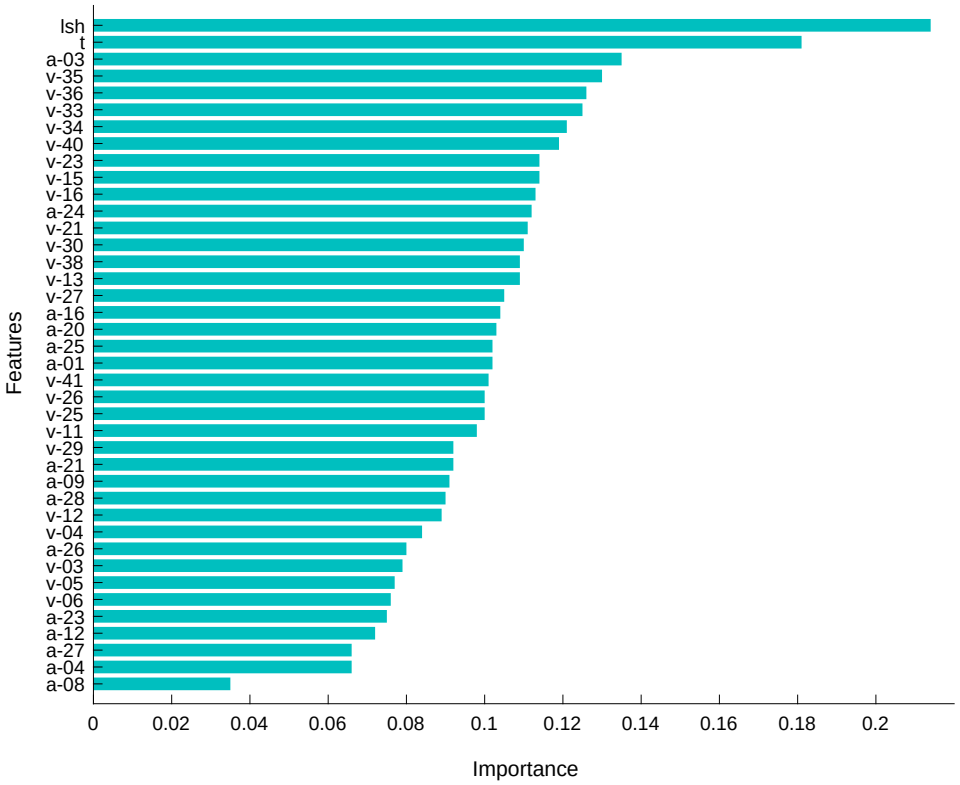

Fig. 3. Feature importance: The importance is calculated as the decrease in AUC after iteratively removing one feature at a time. Last seen HDRS (lst) from the previous week, time (t), features in the audio feature set (a-i: i^{th}), features in the video feature set (v-i: i^{th}).

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 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 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

Table 2. Comparison of AUC of the proposed method and the baselines

Non-Temporal Methods	Pooled AUC	CI
SVM	0.70	[0.63-0.79]
Ensembled Trees	0.71	[0.60-0.79]
Decission Tree	0.74	[0.65-0.81]
Neural Network	0.71	[0.61-0.80]
Temporal Methods		
LSTM	0.80	[0.78-0.91]
LSTM + value-iteration	0.83	[0.79-0.89]
SGLM + value-iteration	**0.88**	[0.83-0.94]

To further investigate the results of the prediction model, we demonstrate the predicted state values for three randomly sampled subjects in figure 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 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 better predicts the treatment result weeks in advance. Moreover, it produces a more stable and robust estimation in contrast to the fluctuations of the HDRS.

5. Discussion and Conclusion

In this paper, 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

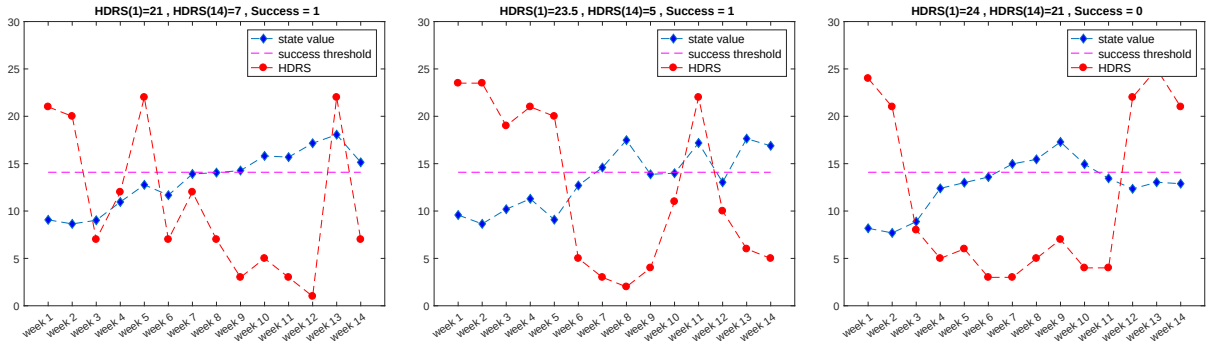

Fig. 4. Trajectory of the estimated state value and HDRS for three subjects. The weekly clinical scores (red circles; higher values indicate decline) are often noisy and may fluctuate from week to week. The proposed machine learning-based scores (blue diamonds; higher values indicate improvement) are less prone to weekly fluctuations and are able to predict the trajectory of a patient and success of the trial weeks in advance.

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 may reflect temporary mood fluctuations rather than a change in illness state. 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.[27] Also, extracting and analyzing visual and audio features altogether to feed to a deep learning model directly is another direction for future research. 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. Moreover, interpretability of the measures derived from machine learning methods remains for future work and research in the field of computational psychiatry.

References

1. H. S. Mayberg, A. M. Lozano, V. Voon, H. McNeely, D. Seminowicz, C. Hamani, J. M. Schwalb and S. H. Kennedy, Deep brain stimulation for treatment-resistant depression, *Neuron* **45** (2005).
2. A. L. Crowell, S. J. Garlow, P. Riva-Posse and H. 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).
3. M. Hamilton, A rating scale for depression, *Journal of neurology, neurosurgery, and psychiatry* **23**, p. 56 (1960).
4. 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*, 2010.
5. N. Cummins, J. Epps and E. Ambikairajah, Spectro-temporal analysis of speech affected by depression and psychomotor retardation, in *Acoustics, Speech and Signal Processing*, 2013.

6. 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*, 2013.
7. J. K. Darby and H. Hollien, Vocal and speech patterns of depressive patients, *Folia Phoniatrica et Logopaedica* **29**, 279 (1977).
8. S. Harati, A. Crowell, H. Mayberg and S. Nemati, Depression severity classification from speech emotion, in *EMBC*, 2018.
9. J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou and F. 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*, 2009.
10. A. H. Farabaugh, S. Bitran, J. Witte, J. Alpert, S. Chuzi, A. J. Clain, L. Baer, M. Fava, P. J. McGrath, C. Dording *et al.*, Anxious depression and early changes in the hamd-17 anxiety-somatization factor items and antidepressant treatment outcome, *International Clinical Psychopharmacology* **25**, p. 214 (2010).
11. A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang and F. 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*, 2015.
12. R. Ramasubbu, M. R. Brown, F. Cortese, I. Gaxiola, B. Goodyear, A. J. Greenshaw, S. M. Dursun and R. Greiner, Accuracy of automated classification of major depressive disorder as a function of symptom severity, *NeuroImage: Clinical* **12**, 320 (2016).
13. K. Anis, H. Zakia, D. Mohamed and C. Jeffrey, Detecting depression severity by interpretable representations of motion dynamics, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.
14. H. Dibeklioğlu, Z. Hammal and J. F. Cohn, Dynamic multimodal measurement of depression severity using deep autoencoding, *IEEE Journal of Biomedical and Health Informatics* **22** (2018).
15. S. Harati, A. Crowell, H. Mayberg, J. Kong and S. 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*, 2016.
16. C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower and S. Kim, Iemocap: Interactive emotional dyadic motion capture database, *Language resources and evaluation* **42**, p. 335 (2008).
17. A. Graves, A. Mohamed and G. Hinton, Speech recognition with deep recurrent neural networks, in *ICASSP*, 2013.
18. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
19. S. Nemati, H. L. Li-wei and R. P. Adams, Learning outcome-discriminative dynamics in multivariate physiological cohort time series, in *EMBC*, 2013.
20. S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9**, 1735 (1997).
21. C. Cortes and V. Vapnik, Support-vector networks, *Machine learning* **20**, 273 (1995).
22. L. Breiman, *Classification and regression trees* (Routledge, 2017).
23. J. Friedman, T. Hastie, R. Tibshirani *et al.*, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics* , 337 (2000).
24. B. J. Parker, S. Günter and J. Bedo, Stratification bias in low signal microarray studies, *BMC bioinformatics* **8**, p. 326 (2007).
25. A. Airola, T. Pahikkala, W. Waegeman, B. De Baets and T. Salakoski, A comparison of auc estimators in small-sample studies, in *Machine learning in systems biology*, 2009.
26. J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., *Radiology* **143**, 29 (1982).
27. E. Reinertsen, S. P. Shashikumar, A. J. Shah, S. Nemati and G. D. Clifford, Multiscale network dynamics between heart rate and locomotor activity are altered in schizophrenia, *Physiological measurement* **39**, p. 115001 (2018).