# Automated Mood-aware Engagement Prediction

Svati Dhamija, Terrance E. Boult
*University of Colorado Colorado Springs*
*Colorado Springs, CO 80918*
{*sdhamija,tboult*}*@vast.uccs.edu*

***Abstract*—Developing intelligent machines that recognize facial expressions, detect spontaneous emotions and infer affective states of an individual are all challenging problems. While significant amount of work in recent years has focussed on advancing machine learning techniques for affect recognition and affect classification, the prediction of mood from facial analysis and the usage of mood data have received less attention. Questionnaires for psychometric measurement of mood-states are common, but using them during interventions that target psychological well-being of people are arduous and may burden an already troubled population. In this work, we present mood prediction as a sequence learning problem that uses facial Action Units (AUs) as inputs to a Long Short-Term Memory (LSTM) machine. We create two separate automated LSTM models – a total mood disturbance predictor and a mood sub-scale predictor, and then use them to aid behavioral assessments of engagement. Our mood-aware engagement predictor uses total mood disturbance score, and our analysis compares both mood sub-scale predictors and an overall mood disturbance predictor for engagement prediction. We evaluate our mood models on a large scale dataset consisting of 8M+ frames from multiple videos collected from 110 subjects during a web-intervention for trauma recovery. Our experiments show that mood-aware engagement predictor using our novel visual analysis approach performs significantly better or on par with using self-reports.**

## 1. Introduction

Each year millions of people are affected by trauma. Emotional and psychological trauma is a result of extraordinarily stressful events such as military service, domestic violence, accidents that can shatter a person's sense of security, making them feel helpless [1]. Recovery from trauma is a complex process. In recent years, web-based treatment has shown promise to provide scalable, proactive, person-centric and evidence-based solution for trauma recovery [2], [3], [4]. In face-to-face psychotherapy sessions, psychologist monitor series of complex processes such as spontaneous facial expressions, emotions, mood swings, engagement and other behaviors when working with patients undergoing treatment for trauma recovery. Typically psychologists are trained to balance patients mental state with doctors broader goals. Due to dynamic and uncertain nature of trauma recovery, there is need to develop self-support tools and technologies that adapt to patient's needs while being sensitive to his/her behavior, moods and emotional state. Further, research in this domain suggests that adaptive self-help websites can aid people in coping with mental health issues and trauma recovery [5].

Facial analysis gives strong clues to the internal emotional state of a person [6], [7]. While emotions and expressions are important and frequently studied with respect to faces, basic emotion categories fail to capture the richness of facial behavior. In some scenarios, it is possible to judge emotion of a person through facial expressions, but internal mood states of person cannot be observed through such methods [8]. This effect is elevated in case of subjects suffering from trauma, where they are often reluctant to express themselves openly. Negative mood has been linked to poorer cognitive and independent functioning in trauma recovery subjects leading to lower quality of life, higher mortality and greater declines in physical and mental health status [1], [9]. Automated mood prediction methods from facial videos have huge the potential to have a significant impact not only for trauma recovery but across a wider-range of domains. Recently, Katsimerou *et al*. [10] conducted one of the first studies on automated mood prediction from visual input and noted that *"mood estimation from recognized emotions is still in its infancy and requires separate attention."*

In affective computing literature, often emotion and mood are used interchangeably. However, as noted by Ekkekais *et al*. [11], these are distinct characteristics of one's affective state. Emotion is characterized by a set of inter-related sub-events concerned with a particular object, person or thing. However, mood typically lasts longer than emotions and is generic rather than about a specific object or a thing. Mood is heavily influenced by external factors such as environment, physiology, and current emotions. Mood is dynamic and can last minutes, hours or even days. Moods are emotions over sustained time [8], [9], [12].

For this work, the ground truth for the subject's mood is a first 24-questions of the widely used standard *"Profile of Mood States - Short Form (POMS-SF)"* [13] questionnaire that seeks to measure six distinct mood subscales: tension, depression, anger, vigor, fatigue, and confusion. We perform training and evaluation of our mood prediction models on

978-1-5386-0563-9/17/$31.00 

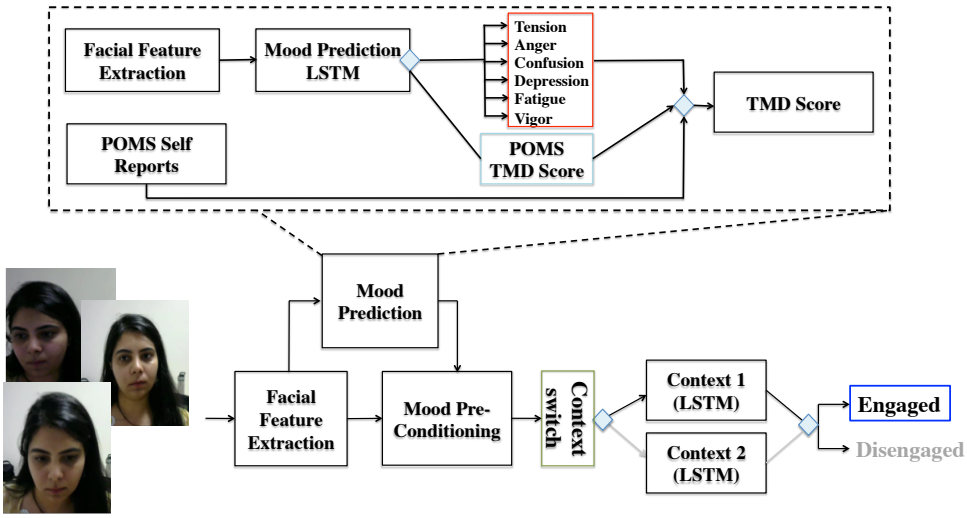


**Figure 1: Automated Mood Prediction for Mood-Aware Context Sensitive Engagement Prediction:** Trauma patients are often reluctant to express themselves openly, suffer from mood changes that last an extended period, which in turn effects their cognitive abilities. We propose Automated Mood-Aware Contextual Engagement prediction for trauma subjects. The top part of the figure shows mood prediction pipeline aimed at predicting the mood of trauma patients from facial videos. The mood estimates are then used to pre-condition learning for context sensitive engagement prediction models. Temporal deep learning methods are used to learn long-term dependencies to estimate mood and its interplay with contextual engagement.

a dataset that is collected from trauma subjects while they work on a recovery website *http://ease.vast.uccs.edu/*. The dataset consists of hundreds of videos collected from 110 subjects with self-reported POMS at three time-points per session, more details in Section 4. Subjects do up to six different modules; herein we consider only two: triggers and relaxation.

The assessment of mood is an important indicator for the evaluation of intervention effects. For example, the standard POMS-SF consists of a detailed questionnaire of 37 questions, which some subjects find intrusive and burdening. Subjects suffering from trauma are often distracted, lack focus, or at times are incapable of providing such detailed feedback [14]. Hence, there is a need to develop automated non-intrusive methods for mood prediction. As shown in Section 5 some aspect of mood is significantly changed by each module, so if mood is to be used during treatment it would require multiple measurements, which further increases the need for automated mood estimation.

In this work, we take advantage of recent advances in computer vision and deep learning and show that facial video data captured over sustained periods can reliably predict the mood of a person with sufficient accuracy to use it in engagement prediction. We build on top of our recent work [15] on contextual engagement prediction from facial videos and develop a framework for mood-aware contextual engagement prediction. An overview of the employed system is shown in Fig.1. The contribution of this work are as follows:

1) We explore automated prediction of mood disturbance based on automatically computed AUs and LSTMs. We develop mood prediction models in the domain of trauma recovery across two contextually different tasks: *Relaxation* and *Triggers*. To the best of our knowledge, this is the first work of its kind.
2) We show that contextual engagement models can be enhanced by incorporating automated mood predictions for trauma recovery subjects.
3) We build automated models for mood prediction at sub-scale and total mood disturbance levels, demonstrating the importance of sub-scale modeling for mood prediction.
4) We evaluate the proposed mood prediction methodology on large scale facial video dataset consisting of 8M+ frames and hundreds of videos. The associated dataset for AUs and Profile of Mood States (POMS) will be publicly released.

## 2. Related Work

Our work has relations to methods and techniques explored from multiple communities such as psychology, affective computing, deep learning, web-based intervention, trauma recovery. Further, our work draws inspiration and builds on top of existing research works from these areas which are reviewed in this section.

**Mood Assessment and Measures:** Affect, expressions, emotions, and mood are related, yet conceptually distinct in terms of the phenomenon they represent [12]. Klienke *et al.* [9] and Adelmann *et al.* [16] have demonstrated the effects of facial expression on mood states of subjects through extensive psychological studies. Studies in psychology (see Ekkekakis *et al.* [11] for a detailed survey) has also lead to the development of various mood assessment measures such as Profile of Mood States and Positive and Negative

Affect Schedule (PANAS). In the domain of sports physiology, Wang *et al*. [17] presented a measurement of mood states from physiological signals. More recently Sano *et al*. [18] presented a system for automatic stress and mood assessment from daily behaviors and sleeping patterns. More recently, Katsimerou *et al*. [10] proposed a novel framework for predicting mood, as perceived by other humans, from the emotional expressions of a person. The key differences between the approach of Katsimerou *et al*.and ours are that we predict mood from user self-reports rather than external annotators and further, we demonstrate the interplay of mood on engagement tasks for web-based trauma recovery.

**Engagement Prediction:** Engagement prediction from facial video data followed by user inputs (either self-reports or external annotation) has been looked into by researchers from the domains of student learning [19], [20], [21] and human-robot interaction [22], [23]. In the domain of student learning, methods typically involve extracting facial features followed by machine learning algorithms to predict student engagement. In the case of human-robot interaction, works of Castellano *et al*. [24] and Salam *et al*. [22] have shown that engagement prediction is contextual and task dependent. They have demonstrated that additional knowledge of user context (e.g. who the user is, where they are, with whom they are, the task at hand, etc.) can better predict an affective state of the user during Human-Robot Interactions (HRI). In the case of both student learning and HRI, it is assumed that subjects (students) are co-operative and in control of their emotions, which is not typically the case for trauma subjects.

**Facial Features:** Work in the domain of engagement prediction, emotion prediction, and facial expression analysis has benefitted tremendously from the recent advances in the domain of facial feature extraction, face tracking, and facial action unit coding. Automated detection of facial action units (AUs) [25], [26], [27] have proved to advance multiple face based affective computing systems [6]. In our work, we rely on AUs extracted from video frames as an intermediate representation provided as input to our sequence learning models. Recent facial expression recognition systems can recognize several AUs with reasonable accuracies [25], [28], [29]. Finally, there have been multiple notable works in the domain of facial expression and affect analysis that has pushed state of the art in affect recognition beyond six basic emotion categories [30], [31].

**Deep Learning for Affect Detection:** Significant advances in deep learning have lead to the development of various affect detection methods based on deep learning. More specifically, researchers have applied deep learning techniques to problem such as continuous emotion detection [32], facial expression analysis [33], facial action unit detection [34] and others. Deep learning methods have used video data, sensor data or multi-modal data [35]. As noted earlier, moods are diffused over longer durations than emotions, and hence there is need to employ methods that can integrate long-term temporal information in learning framework. Such problems are often modeled as sequence learning problems [36], [37]. To address sequence learning, various methods such as Recurrent Neural Networks, Gated Recurrent Units, and Long Short-Term Memory were proposed and employed in wide range of problems [38]. In this work, we explore LSTMs, a specialized form of recurrent neural networks, to model long-term mood prediction and mood-aware engagement prediction.

## 3. Long Short-Term Memory for Mood Prediction

In this section, we present the details about LSTMs used to model long-term dependencies of AUs. LSTMs have the ability to handle longer sequences. We model the problem of mood prediction as a sequence learning problem, where input consists of sequences $x_i$ of AUs computed from facial video data of a particular length. Each sequence is associated with a label $y_i$ which relates to POMS self-report provided by trauma subjects. Our implementation is based on TensorFlow which in turn is based on [39], [40], and we follow their notation.

We let subscripts denote timesteps and superscripts denote layers. All our states are $n$-dimensional equal to the number of AUs tracked, currently 20. Let $h_t^l \in \mathbb{R}^n$ be a hidden state in layer $l$ at time-step $t$. Let $T_{n,m} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be an affine transform from $m$ to $n$ dimensions, i.e. $T_{n,m}x = Wx + b$ for some $W$ and $b$). Let $\odot$ be element-wise multiplication and let $h_t^0$ be an input data vector at time-step $t$. We use the activations $h_t^L$ to predict $y_t$, since $L$ is the number of layers in our deep LSTM.

The LSTM has complicated dynamics that allow it to easily "memorize" information for an extended number of time-steps using *memory cells* $c_t^l \in \mathbb{R}^n$. Although many LSTM architectures that differ in their connectivity structure and activation functions, all LSTM architectures have explicit memory cells for storing information for long periods of time, along with weights for updating the memory cell, retrieving it, or keeping it for the next time step. The LSTM architecture used in our experiments is given by the following equations [36], as implemented in TensorFlow basic LSTM cell:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$
$$h_t^l = o \odot \tanh(c_t^l)$$

where $\text{sigm}$ is the sigmoid function, $\text{sigm}$ and $\tanh$ are applied element-wise, $i, f, o, c, h$ are the input gate, forget gate, output gate, cell activation vector and hidden vectors, respectively. In this work, we assume the length of sequence is known a-priori and hence use single layer LSTM with static RNN cells.

## 4. EASE dataset

In this section, we present details of the data used for our analysis of mood-aware engagement prediction. The dataset we use is called EASE (Engagement Arousal Self-Efficacy) [15] [1].

| | SESSION 1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MODULE 1 | | | | MODULE 2 | | | |
| | Task | Number of Videos | Number of Frames | Number of Self Reports | Task | Number of Videos | Number of Frames | Number of Self-Reports |
| Trigger Task followed by Relaxation Task | Trigger | 52 | 806855 | 166 | Relaxation | 52 | 1579927 | 122 |
| Relaxation Task followed by Trigger Task | Relaxation | 43 | 1391803 | 91 | Trigger | 43 | 590953 | 98 |

| | SESSION 2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MODULE 1 | | | | MODULE 2 | | | |
| | Task | Number of Videos | Number of Frames | Number of Self Reports | Task | Number of Videos | Number of Frames | Number of Self Reports |
| Trigger Task followed by Relaxation Task | Trigger | 33 | 454510 | 94 | Relaxation | 33 | 1553409 | 63 |
| Relaxation Task followed by Trigger Task | Relaxation | 47 | 1139996 | 105 | Trigger | 47 | 544298 | 149 |

**Figure 2:** Information about participants and the distribution of modules taken by them in each session considered for mood analysis in this work. Participants consisted of total 110 subjects with 88 Female, 17 Male, 5 unspecified in the age group of 18-79 years, with 80% being under the age of 46.

**Dataset details:** The web-intervention used to collect the data was based on the findings of Social Cognitive Theory [41] and consisted of subjects undergoing six tasks (modules) namely: social-support, self-talk, relaxation, unhelpful coping, professional help and triggers. The broader study was divided into three sessions/visits in the form of a Randomized Control Trial (RCT). Each participant was assigned two out of the six modules in each visit. The first two visits were restricted to "Relaxation" (RX) and "Triggers" (TR) modules only and in the third visit the participants were free to choose from the remaining four modules. Each visit lasted for approx. 30 minutes - 1.5 hours. In the first visit, subjects were randomly allocated Relaxation or Triggers as the first module and a reverse order during the second visit and second module. At the beginning of each visit, the subjects listened to a neutral introductory video. During these sessions, a LogiTech webcam with a resolution of 640x480 at 30 fps was placed on top of the monitor recording the participants face video along with audio. Physiological data was also recorded for the entire session. The participants could freely interact with the trauma recovery website, and their interactions were recorded in the form of Picture in Picture video using a Camtasia recorder (with screen and webcam recording simultaneously). During the module, participants provided self-reports about their engagement level, Profile of Mood States before the module (pre-POMS) and after the module (post-POMS). Although the EASE data is significantly rich in terms of its multi-modal nature, for this work we focus our attention primarily on facial video data captured by webcam placed on monitor, POMS responses provided by the participants and the engagement self-reports.

As mentioned earlier, each participant came in for three sessions/visits. The first two (controlled) visits are used for experiments in this paper. In each visit, the subjects undergo relaxation and trigger tasks. The relaxation module presents the user with video demonstrations of various exercises like breathing, muscle relaxation, etc. The triggers module presents educational material to the user about trauma symptoms and prevention. Since few subjects dropped out during the study, for the first session, we have data from 95 subjects and from the second session we use data from 80 subjects. Some of the collected data was unusable due to either system issues, data corruption or lack of engagement self-reports.

**Profile of Mood States (POMS):** The Profiles of Mood States-Short Form (POMS-SF) [13], [14] are used to measure reactive changes in the mood of a person. It is a list of 37 items related to depression, vigor, tension, anger, fatigue, and confusion. In EASE data the TMD score was calculated from a 24-point questionnaire instead of 37 to reduce the cognitive load of trauma subjects. Participants rate items for how they feel at the moment on a 5-point scale, ranging from 1 (not at all) to 5 (extremely). Each question corresponds to a specific sub-scale of mood, e.g., tension: negative sentiment (5 questions), depression: negative sentiment (6 questions), anger: negative sentiment (5 questions), fatigue: negative sentiment (2 questions), confusion: negative sentiment (2 questions) and vigor: positive sentiment (4 questions). The final TMD (Total Mood Disturbance) level is computed as difference of sum of negative n(x) and positive p(x) sentiments:

$$\text{TMD} = \sum_{x \in \text{negative sentiments}} n(x) - \sum_{x \in \text{positive sentiment}} p(x) \quad (1)$$

This self-report is given to the subjects at baseline and immediately before and after each module. Each video consisted of mood states report before and after the module (pre-POMS report and post-POMS report). Participants also provided self-reports about their engagement level approx. three times (start, middle, and end of the segment) during each session. Figure 3 shows the changes in mood at the TMD and sub-scale level in the EASE dataset. Greater TMD scores are indicate of subjects with more unstable mode profiles and lower scores indicate stable mood profiles.

## 5. Experiments

We now describe the AU computation procedure to extract intermediate feature representation, followed by the methodology for sequence learning using LSTMs and the various LSTM mood/engagement models. Lastly, we elaborate the data used for training and testing.

1. We have been granted IRB approval to release unidentifiable data for research purposes, including AUs extracted from facial videos and the associated POMS data. Since identifiable data cannot be released, we cannot display raw video frames from the dataset in this paper.

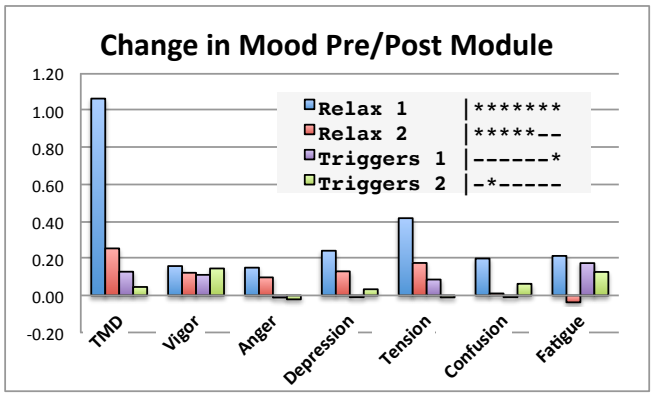


**Figure 3:** Changes in subject moods pre/post each module in each session. We tested each change for significant with a two-sided paired t-test. The legend, a * means that scale/subscale showed a statistically significant decrease change at the .01 level. Because of the significant changes during each module one cannot compute TMD once and reuse it across the treatment – we need automated mood estimation.

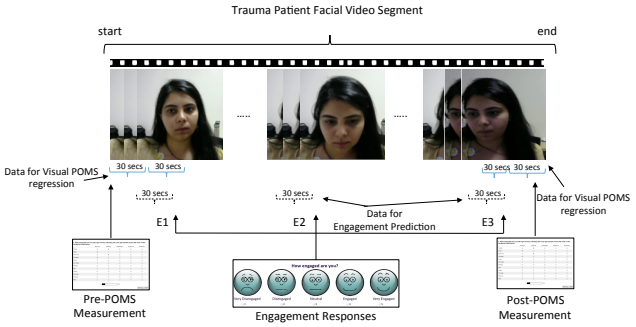


**Figure 4: Video segment selection:** The process of selecting video segments from a particular video session for training POMS models and engagement models is shown in above figure. Each video segment consists of POMS measurement at the start and end of the session. For training POMS regression models POMS-TMD and POMS-TMDS, we consider two 30 seconds segments from pre-POMS and post-POMS measurement. Engagement self-reports are collected at approx. 3 time-points during the module. Training data for engagement prediction consists of 30 second segments prior to each engagement self-report. From each of these frames (either for POMS or engagement), AUs are extracted from video frames and used as intermediate feature representation for training LSTMs

**Facial action units extraction:** As noted in earlier section 4, the collected dataset consisted of a large number of facial video and engagement self-reports. While there are number of software available for extracting facial landmark points and facial action units (e.g. [25], [26]) we use the recent work on OpenFace [28] proposed by Baltrušaitis *et al*.. It is an open-source tool which has shown state-of-the-art performance on multiple tasks such as head-pose, eye-gaze, and AU detection. For our work, we primarily focus on facial action units. The AUs extracted consisted of intensity based and presence based AUs. The list of AUs used in this paper are as follows: Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer (intensity), Upper Lid Raiser, Cheek Raiser, Nose Wrinkler, Upper Lip Raiser, Lip Corner Puller (intensity), Dimpler, Lip Corner Depressor (intensity), Chin Raiser, Lip Stretched, Lips Part, Jaw Drop, Brow Lowerer (presence), Lip Corner Puller (presence), Lip Corner Depressor (presence), Lip Tightner, Lip Suck, Blink.

**LSTM Engagement and Mood Models:** In this work, we model four variants of engagement prediction using the same basic LSTM cell which are as follows (each model is a single layered LSTM):

1) *Engagement multi-class classifier*
We optimize the LSTM to predict a discrete set of engagement levels 1 (very disengaged) through 5 (very engaged), by minimizing the cross-entropy loss of the predicted and actual class after applying softmax function. By doing this, each engagement level is treated as a separate and mutually exclusive output. The baseline from [15] was recomputed to include the subjects that had POMS data available.
2) *Mood-aware engagement prediction : POMS-SR*
We precondition the basic engagement multi-class LSTM with TMD scores obtained using self-reports. We max normalize the TMD scores during training and propagate the normalized score along with the intermediate AU representations by adding the normalized score to the AU representation. The model is a reproduction of our prior work [15].
3) *Mood Disturbance predictor : POMS-TMD*
POMS-TMD is a single-output regressor. This LSTM model is optimized to predict continuous POMS-TMD scores in range 0-24, by back-propagating the L2-regularization loss. Using L2-regularization, we penalize the outliers heavily by adding a penalty on the norm of the weights to the loss.
4) *Mood Sub-scale predictor : POMS-TMDS*
The POMS sub-scale LSTM (POMS-TMDS) is a multi-output regressor. It is optimized to predict six sub scales of tension, vigor, depression, anger, confusion and fatigue each in range 1-5, by computing the L2-regularization loss and an Adam optimizer with a learning rate of 0.1 trained over 15 epochs, see Figure 5 for a choice of epoch-size. The estimated sub-scale values are then used to compute a TMD score estimate.

**Cross-Validation and Train-test pipeline:** Our multi-class engagement model consists of a 20 fold validation with 420 segments of 30 seconds each in training set and 46 segments of 30 seconds each in testing set for Trigger (TR) module. Similarly, in Relaxation (RX) module, our training set consists of 313 segments of 30 seconds each and 35 segments of 30 seconds each in testing module. Each segment had a self-reported engagement score as ground-truth on a scale of 1-5. This led to the availability of AUs from 378K frames for training LSTM for engagement prediction for TR module and AUs from 282K frames for training in RX module. Mood-aware contextual engagement LSTMs were

trained based on context. Our automated mood LSTM's i.e. POMS-TMD and POMS-TMDS had 600 segments in training set and 67 segments in test set. As shown in Figure 4 two 30 second segments were selected with same POMS label to increase the effective size of training samples. Each segment was accompanied by POMS self report collected before/after the module as shown in Figure 4. TMD scores from self reports were used to condition the AUs and form the POMS-SR baseline. We formulate the problem of automated mood prediction as a sequence based single-output (POMS-TMD) or multi-output (POMS-TMDS) regression and train LSTM to predict a TMD score or individual mood clusters. Finally, we used automated mood prediction pipeline to compute TMD scores from both POMS-TMD and POMS-TMDS models. These automated TMD scores were used to pre-condition AUs and LSTMs were trained for engagement prediction. Convergence plot for one fold of data on the POMS-TMD LSTM is shown in Figure 5. It took approximately 5-7 epochs for both TR and RX contextual models to stabilize test loss. After 60 epochs the model starts overfitting, therefore, we selected 15 epochs as a constant for all cross-validation model evaluations irrespective of the context.

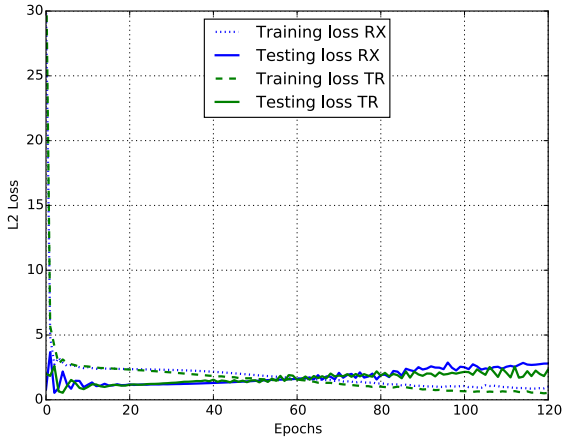


**Figure 5: Convergence Plot:** The above figure shows L2 loss for training LSTM model for Mood Prediction as a function of Epochs. It can be observed that it takes approximately 50-60 epochs for train and test losses to converge for both RX and TR modules. After the convergence point, the model is prone to over-fitting.

## 6. Results and Evaluation

In this section, we discuss in detail the results obtained for engagement prediction across variety of tasks and its task specificity. The results are summarized in Table 1.

We train separate contextual LSTMs for each visual model (RX and TR) i.e. engagement predictions, POMS-TMD and POMS-TMDS. LSTMs trained with only AUs from the engagement segments serve as baseline for all predictions. We obtain 48.55 $\pm$ 17.7 % average prediction accuracy across 20-folds on TR module and 42.04 $\pm$ 11.7 % average prediction accuracy for RX module (margins of error correspond to standard deviation computed over 20-folds). We noticed significant improvements in performance for TR module in a 2-sided tailed paired t-test (p=.07) with average prediction accuracy of 52.57 $\pm$ 17.21% for POMS-SR, whereas RX accuracy increased to 43.88 $\pm$ 11.9% but was not statistically different (p=.26), consistent with the findings in [15]. Using the visual estimations from POMS-TMD regressor increased engagement accuracy significantly for TR at p=.05 level when compared to POMS-SR, highlighting the effect of using visual POMS rather than self-reports. On the other hand, for RX although the visual predictors of POMS-TMD and POMS-TMDS show improvements in accuracy from baseline 46.14 $\pm$ 14.3 % and 44.42 $\pm$ 12.9 %, there is statistically weak evidence at p=.19 when evaluated with POMS-SR.

**TABLE 1:** The first row shows engagement prediction results without mood conditioning. The second row represents mood pre-conditioning performed with POMS self reports for engagement prediction. The third and fourth row show engagement prediction accuracy with mood conditioning performed POMS-TMD estimates from temporal deep learning based mood prediction. We note that the performance of engagement predictions can be improved by pre-conditioning with POMS self-report data. Accuracy is further enhanced by pre-conditioning with visual POMS-TMD estimates, suggesting effectiveness of automated mood prediction. The performance numbers are obtained using 20-fold cross validation, with accuracy along with standard deviation reported in the above table.

| | Trigger | Relaxation |
|---|---|---|
| Engagement-baseline | 48.55 $\pm$ 17.7 % | 42.04 $\pm$ 11.7 % |
| Engagement-Mood Aware **POMS-SR** | 52.57 $\pm$ 17.2% | 43.88 $\pm$ 11.9 % |
| Engagement-Automated MA **POMS-TMD** | 54.04 $\pm$ 18.14 % | 46.14 $\pm$ 14.38 % |
| Engagement-Automated MA **POMS-TMDS** | 55.54 $\pm$ 18.43 % | 45.42 $\pm$ 12.99 % |

While evaluating, POMS-TMD regressor obtained 0.03673 average Mean Squared Error (MSE) with 0.0153 standard deviation in mood prediction for TR module and 0.03633 average MSE with 0.01227 standard deviation in mood prediction for RX module using the POMS-TMDS predictor. Average MSE and standard deviation for mood prediction were computed over 10-fold cross validation. However, evaluating the performance of visual mood-aware regressors on a standalone basis is not required, since the visual regressors surpass the POMS-SR. However, for completeness, the POMS-TMD and POMS-TMDS results are shown in Figure 6 alongwith the actual scores from self-reported POMS data. Moreover, the statistically insignificant results in RX could be attributed to the significant changes in mood states during the task of RX vs TR, analyzed in Figure3. Such significant changes in mood-states during a task, strongly suggest the need to build continuous predictors of mood states from video in order to monitor the users during an intervention.

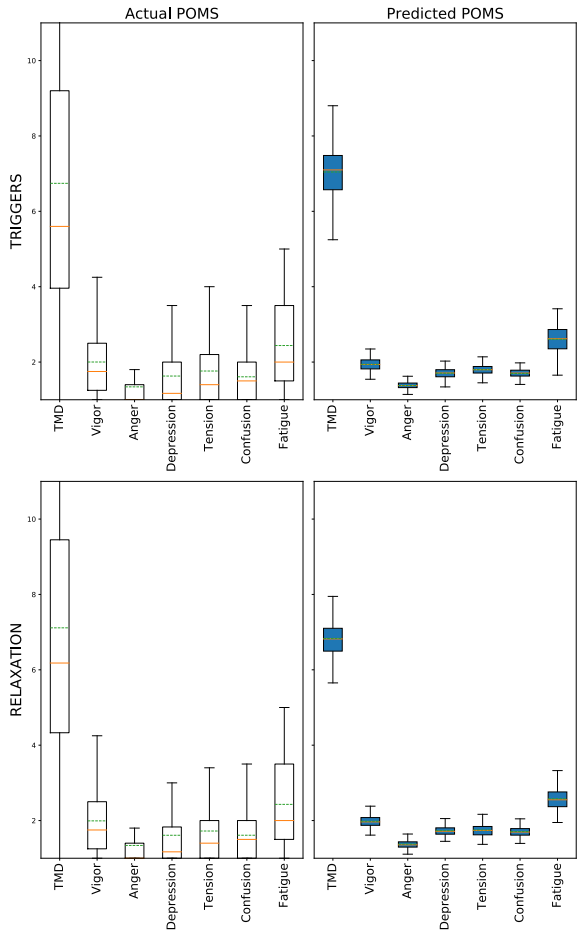


**Figure 6: Automated Mood Predictors:** This figure shows the POMS-TMD and POMS-TMDS values for Relaxation as well as Triggers modules. The subplots on the left show the Actual TMD and sub-scale values from subject self-reports and the subplots on the right show the predicted TMD and sub-scale values. Each subplot shows the mood-scores on the y-axis and the mood-type on the x-axis. Each boxplot shows the mean (green line), median (orange line), and the inter-quartile range i.e. upper (75th percentile) and lower (25th percentile) lines of the box. Notice, that even though the predicted values do not predict over the full dynamic range of mood-scores, the mean of the actual and predicted POMS is similar at TMD and sub-scale levels.

## 7. Conclusion

In this work we presented a method for automated detection of total mood disturbance and mood-subscale prediction from facial action units and LSTMs. Our experiments show that mood prediction using visual estimates of POMS-TMD and POMS-TMDS performs better or at-par with self-reported TMD. We used the detected mood to aid engagement prediction models. The presented mood-aware engagement prediction models outperformed baseline engagement prediction model that relies only on video data. We evaluated the proposed method on large scale dataset collected from subjects during web-intervention for trauma recovery. However, in this work, we considered a singular automated mood-model unlike the contextual engagement models. Future work may explore the possibilities of contextual mood. We also used LSTM's in this work which are highly parametric models and require extensive optimization. In this work we optimized basic parameters like number of epochs. Next obvious step is to fine-tune other hyper-parameters for enhanced accuracy.

Beyond psychological studies, automated mood prediction provides multiple applications in other areas as well. In human robot interaction [22], mood assessment can lead to adapting robot behavior based on subjects affective state. In the domain of multi-player games [42], one can start addressing questions like do a player's actions reveal a friendly/aggressive mood towards the other players? Can we use the player's actions to predict subsequent actions? In domain of computer-aided instruction [20], understanding mood and engagement in automated and scalable way can help devise better learning tools or personalized instruction mechanisms. In workplaces, understanding performance during job-interview, analyzing stress of employees, assessing performance of call center employees during long conversations can be significantly improved by automated mood and engagement prediction methods. In the domain of computational Advertising [43], automated mood-aware engagement prediction methods can help create more sophisticated tools to better align advertising needs of users and content creators.

There are multiple implications of the proposed work. Mood and its relationship with other cognitive abilities has been widely studied in psychology literature. Collecting questionnaires from subjects is cumbersome process and in many cases like trauma recovery, poses additional burden on the subjects. Automated mood prediction using facial video data provides a scalable platform to study wide range of behavioral tasks and opens up multiple opportunities in the domains of trauma recovery, elderly care, treating mood disorders and other rehabilitation. In the domain of web-intervention, automated mood and engagement prediction provides a mechanism to build adaptive evidence-based treatment [2], [44], [45].

## References

[1] K. E. Cherry, L. D. Marks, R. Adamek, and B. A. Lyon, "Traumatic stress and long-term recovery," 2015.

[2] V. S. Mehta, M. Parakh, and D. Ghosh, "Web based interventions in psychiatry: An overview," *International Journal of Mental Health & Psychiatry*, vol. 2015, 2016.

[3] D. Macea, K. Gajos, Y. D. Calil, and F. Fregni, "The efficacy of web-based cognitive behavioral interventions for chronic pain: a systematic review and meta-analysis," *J. of Pain*, vol. 11, no. 10, pp. 917–929, 2010.

[4] C. Benight, J. Ruzek, and E. Waldrep, "Internet interventions for traumatic stress: A review and theoretically based example," *J. of Trauma Stress*, vol. 21, no. 6, pp. 513–520, 2008.

[5] R. A. Calvo, K. Dinakar, R. Picard, and P. Maes, "Computing in mental health," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016, pp. 3438–3445.

[6] F. D. la Torre and J. Cohn, "Facial expression analysis," *Visual Analysis of Humans - Springer*, pp. 377–409, 2011.

[7] J. Cohn, "Foundations of human computing: facial expression and emotion," *ICMI*, pp. 233–238, 2006.

[8] P. Ekman and A. Fridlund, "Assessment of facial behavior in affective disorders," *Depression and expressive behavior*, pp. 37–56, 1987.

[9] C. L. Kleinke, T. R. Peterson, and T. R. Rutledge, "Effects of self-generated facial expressions on mood." *Journal of Personality and Social Psychology*, vol. 74, no. 1, p. 272, 1998.

[10] C. Katsimerou, I. Heynderickx, and J. A. Redi, "Predicting mood from punctual emotion annotations on videos," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 179–192, 2015.

[11] P. Ekkekakis, *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press, 2013.

[12] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[13] S. Shacham, "A shortened version of the profile of mood states," *Journal of Personality Assesment*, vol. 47, no. 3, pp. 305–306, 1983.

[14] R. R. Edwards and J. Haythornthwaite, "Mood swings: variability in the use of the profile of mood states," *Journal of pain and symptom management*, vol. 28, no. 6, p. 534, 2004.

[15] S. Dhamija and T. Boult, "Exploring contextual engagement for trauma recovery," *CVPR Workshop on Deep Affective Learning and Context Modelling*, 2017.

[16] P. K. Adelmann and R. B. Zajonc, "Facial efference and the experience of emotion," *Annual review of psychology*, vol. 40, no. 1, pp. 249–280, 1989.

[17] J. Wang, P. Lei, K. Wang, L. Mao, and X. Chai, "Mood states recognition of rowing athletes based on multi-physiological signals using pso-svm," *E-Health Telecommunication Systems and Networks*, vol. 2014, 2014.

[18] A. Sano, Z. Y. Amy, A. W. McHill, A. J. Phillips, S. Taylor, N. Jaques, E. B. Klerman, and R. W. Picard, "Prediction of happy-sad mood from daily behaviors and previous sleep history," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6796–6799.

[19] H. Monkaresi, P. Bosch, R. Calvo, and S. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. on Affective Computing*, 2017.

[20] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "Faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. on Affective Computing*, vol. 5, no. 3, pp. 86–98, 2014.

[21] J. Grafsgaard, J. Wiggins, K. Boyer, E. Wiebe, and J. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," *Educational Data Mining*, 2013.

[22] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. 5, pp. 705–721, 2017.

[23] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *ICMI*. ACM, 2009, pp. 119–126.

[24] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in hri: An exploration of social and task-based context," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 421–428.

[25] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[26] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[27] M. Cox, J. Nuevo-Chiquero, J. Saragih, and S. Lucey, "Csiro face analysis sdk," *Brisbane, Australia*, 2013.

[28] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[29] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 944–958, 2015.

[30] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[31] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[32] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.

[33] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial expression intensity estimation," *CVPR*, 2017.

[34] W. Chu, F. D. la Torre, and J. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," *Automatic Face and Gesture Conference*, 2017.

[35] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[36] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

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

[38] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, 2016.

[39] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[40] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[41] C. Benight and A. Bandura, "Social cognitive theory of posttraumatic recovery: the role of perceived self-efficacy," *Behaviour Research and Therapy,Elsevier*, 2004.

[42] G. N. Yannakakis and A. Paiva, "Emotion in games," *Handbook on affective computing*, pp. 459–471, 2014.

[43] D. McDuff, "New methods for measuring advertising efficacy," *Digital Advertising: Theory and Research*, 2017.

[44] C. Yeager, "Understanding engagement with a trauma recovery web intervention using the health action process approach framework," *PhD Thesis*, 2016.

[45] K. Shoji, C. Benight, A. Mullings, CarolynYeager, S. Dhamija, and T. Boult, "Measuring engagement into the web-intervention by the quality of voice," in *International Society for Research on Internet Interventions*. ISRII, 2016.