Depression Severity Assessment for Adolescents at High Risk of Mental Disorders

Michal Muszynski
mmuszyns@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jamie Zelazny
jmz22@pitt.edu
University of Pittsburgh
Pittsburgh, PA, USA

Jeffrey M. Girard
jmgirard@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Louis-Philippe Morency
morency@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Recent progress in artificial intelligence has led to the development of automatic behavioral marker recognition, such as facial and vocal expressions. Those automatic tools have enormous potential to support mental health assessment, clinical decision making, and treatment planning.

In this paper, we investigate nonverbal behavioral markers of depression severity assessed during semi-structured medical interviews of adolescent patients. The main goal of our research is two-fold: studying a unique population of adolescents at high risk of mental disorders and differentiating mild depression from moderate or severe depression.

We aim to explore computationally inferred facial and vocal behavioral responses elicited by three segments of the semi-structured medical interviews: Distress Assessment Questions, Ubiquitous Questions, and Concept Questions. Our experimental methodology reflects best practise used for analyzing small sample size and unbalanced datasets of unique patients. Our results show a very interesting trend with strongly discriminative behavioral markers from both acoustic and visual modalities. These promising results are likely due to the unique classification task (mild depression vs. moderate and severe depression) and three types of probing questions.

CCS CONCEPTS

• **Applied computing** → **Health informatics**.

KEYWORDS

Health Informatics; Mental Disorders; Adolescent Depression; Depression Diagnosis; Non-verbal Behavioral Markers

ACM Reference Format:
Michal Muszynski, Jamie Zelazny, Jeffrey M. Girard, and Louis-Philippe Morency. 2020. Depression Severity Assessment for Adolescents at High Risk of Mental Disorders. In *Proceedings of the 2020 International Conference*

1 INTRODUCTION

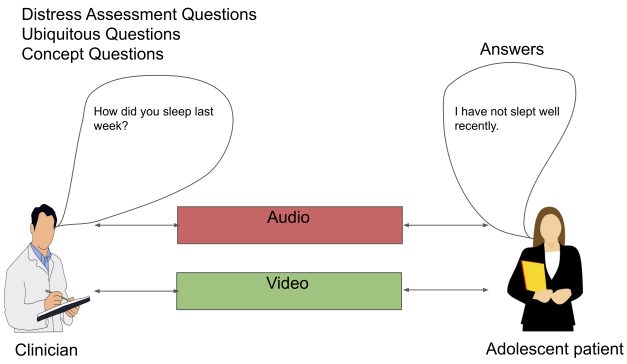

Figure 1: Overview of semi-structured clinician-patient medical interviews

Depression is a major global health concern and has been recognized as a complex cause of disability that reduces the quality of life and productivity of societies worldwide [25]. Moreover, depression can lead to high risk of suicidal behaviors [22]. Up to half of people committing suicide also meet clinical criteria of depression [27].

It is thus imperative that we develop high quality techniques for assessing and treating depression that can scale to address this concern [7]. However, depression assessment is a challenging task because depression has many subtle signs and symptoms that vary from patient to patient and even from time-to-time (depending on context).

In order to increase the reliability and validity of depression assessment, these signs and symptoms are typically scored by clinicians using standardized interviews [7, 18]. During such interviews, clinicians ask structured questions (to control the context) and use their best judgment to rate the presence of depression's signs and symptoms, including various behavioral markers (e.g., persistently monotone voice, slowed movements, and apparent sadness). This process can work well when clinicians are adequately trained and

enables them to use their expertise to determine whether an observed sign (e.g., slowed movements) is caused by depression or something else (e.g., fatigue), which is an important advantage of interviews over self-reported symptom measures.

However, it can also be time-consuming and subjective (e.g., two clinicians may disagree on whether a behavioral marker is present). Therefore, automatic measures of behavioral markers are desired to enhance the efficiency and objectivity of depression assessment [18].

In this paper, we develop a bi-modal approach to automatically measure acoustic and visual behavioral markers of depression severity during semi-structured clinical interviews with high-risk patients (Figure 1). We aim at assessing severity of adolescent depression that could be a next step when depression diagnosed. We are interested in discriminating between adolescent patients having mild depression and adolescent patients suffering from moderate or severe depression.

Our approach leverages the fact that semi-structured interviews create different contexts based on the questions being asked and answered, and we investigate whether contextualizing patients' behavior in this way will improve our ability to distinguish between patients with mild and more severe depression symptoms.
In this work, we address three main research questions:

- **RQ1**: Is it possible to identify computationally inferred acoustic and visual behavioral markers that distinguish between adolescents with mild depression from adolescents with moderate or severe depression?
- **RQ2**: Which parts of the semi-structured medical interview produce the most informative behavioral patterns for assessing depression symptom severity?
- **RQ3**: Is it possible to accurately classify the severity of depression based on acoustic and visual behavioral patterns of adolescents' responses?

The main contributions of this work are:

- We study a unique population of adolescents at high risk of mental disorders.
- We explore computationally inferred facial and vocal behavioral responses elicited by three segments of the semi-structured medical interviews: Distress Assessment Questions, Ubiquitous Questions, and Concept Questions.
- We provide the experimental methodology that reflects best practise to analyze small sample size and unbalanced datasets of unique patients.
- We reveal the strong relationship between some acoustic and visual behavioral markers and depression severity, with different question contexts.

The rest of this paper is organized as follows: Section 2 reviews visual and acoustic behavioral markers of depression. Section 3 consists of description of semi-structured medical interviews and data collection protocols. Section 4 corresponds to descriptions of multimodal feature extraction. Section 5 consists of an analysis of the relationship between depression severity, and acoustic and visual features. Section 6 presents results of depression severity classification. Section 7 discusses performance of depression severity classification and limitations of our study. Section 8 includes conclusions and describes future directions of our research.

2 RELATED WORK

Visual and acoustic behavioral markers of depression. In the last decades, a lot of psychological research has focused on investigating the relationship between nonverbal behavior and depression [7]. One of the main findings with regard to facial behavior is that having depression is usually coupled with reduced amounts of positive facial expressions [4, 17, 36, 38, 41, 46]. Also, some studies on depression showed that general facial expressiveness was also reduced [16, 36].

The common symptoms of depression associated with facial expressions are sad facial expressions as well as an overall lack of facial expressions accompanied by the reduced affective responses [13]. Nevertheless, researchers have argued about the depression impact on negative facial expressions. Some studies found that depression increased amounts of negative facial expressions [3, 34, 40] while other studies reported that people suffering from depression displayed reduced amounts of negative facial expressions [16, 36]. As a result, several studies have attempted to detect depression from non-verbal behavior markers.

For example, Support Vector Machines (SVMs) individually fed with manually annotated Facial Action Units (AUs), Active Appearance Model, and vocal prosody features were used to detect depression [6]. The analysis of results showed that all of those features were discriminative for depression detection although the best detection accuracy of 88% was achieved by SVMs trained on AUs.

Moreover, Girard et al. [19] carried out an analysis of the relationship between non-verbal behavior markers, such as AUs and head pose manually annotated and automatically described. The study reported that subjects suffering from severe depression displayed less affiliative facial expressions (AU 12 and AU 15), and more non-affiliative facial expressions (AU 14) accompanied by diminished head motion. Also, the authors confirmed the outcomes of the analysis on manual annotations and automatically extracted descriptors of nonverbal behavior markers are consistent. Overall, those previous studies suggest that automatic nonverbal behavior marker analysis can lead to development of automatic depression analysis and detection.

Speech has also been considered as an objective marker of depression in the last decades. Prosodic abnormalities such as low voice, monotonous voice, and speaking slowly, stuttering and whispering have been associated with depression. Cummins et al. [7] hypothesized that depression causes cognitive and physiological changes that lead to noticeable acoustic changes in speech production. The process of speech production requires cognitive planning and complex motoric muscular actions, simultaneously. Many studies have investigated the relationship between patterns in speech of depressed individuals.

Darby et al. [8] found that listeners could perceive speech changes of depressed patients before and after treatment measured by pitch, loudness, speaking rate and articulation. Christopher et al. [5] showed that depression influences phonological loop resulting in phonation and articulation artefacts.

Williamson et al. [52] reported that depressed patients often hesitated slightly before answering and had problems with choosing words. Those previous findings motivate our study of acoustic

features as discrminative markers for depression severity classification.

Automatic depression analysis. In the last decade, automatic depression analysis has drawn a lot of attention. As a result, several challenges related to automatic depression analysis were organized [37, 47–49]. The primary research on automatic depression analysis used classical machine learning models, such as Support Vector Machine Regression [20, 49], Decision Tree [45, 53, 54], and Logistic Regression [11] to detect depression. Those studies often extracted hand-crafted features, such as Low Level Descriptors (LLDs) [54], Histogram Oriented Gradients (HOGs) [48], and Local Binary Patterns (LBPs) [10] with Edge Orientation Histograms (EOHs) [28]. An example approach to depression prediction is to extract LBPs and EOHs of images as visual features combined with LLDs, e.g., pitch, loudness, jitter, shimmer, and Harmonics to Noise Ratio, and Mel-Frequency-Cepstral-Coefficients (MFCCs) of acoustic signals [28]. In addition to various visual features, acoustic features appeared to be informative for automatic depression analysis [48, 49]. Williamson et al. [51, 52] extracted formant frequencies and delta-mel-cepstra from acoustic signals to describe changes in shape and dynamics of vocal expressions.

Recent developments in deep learning allow researchers to apply neural network architectures, such as Convolutional Neural Networks and Recurrent Neural Networks to perform automatic depression analysis. For example, Ma et al. [24] built the DeepAudioNet for depression classification based on audio signals. The DeepAudioNet is composed of Convolutional Neural Networks and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN). Rejaibi et al. [35] proposed a MFCC-LSTM-RNN based approach for automatic depression recognition. The approach relies on LSTM-RNNs fed with MFCC coefficients to detect depression and to assess its severity levels. Al Jazaery et al. [1] used a 3D convolutional neural network to extract short-term dynamic visual representation of depression from video segments. Then, a Recurrent Neural Network was applied to learn further from the sequence of the spatio-temporal information in order to predict depression. The video-level predictions of depressions were obtained by averaging segment-level predictions for each video. Also, the DepressNet was proposed to learn depression representations with visual explanations [55]. The facial regions that were the most discriminative were marked and used to predict depression in images.

In this paper, we aim at studying a unique population of adolescents at high risk of mental disorders and distinguishing between mild depression, and moderate or severe depression. We are interested in investigating acoustic and visual behavioral markers of depression severity assessed during semi-structured medical interviews. Those acoustic and visual behavioral makers could help clinicians to distinguish patients having mild depression from patients suffering from moderate or severe depression by means of automatic depression analysis.

3 DATASET

We collected a new dataset of clinical interviews with adolescent patients at high risk of mental disorders, including depression and suicidal ideation. Participants had to be between 13 and 25 years old, fluent in English, and undergoing treatment at an Intensive Outpatient Program at Western Psychiatric Hospital in the United States for severe depression and/or suicidality at the time of enrollment. Each patient participated in up to 4 interview sessions (baseline and 3 follow-up sessions) conducted by the same clinician. The interviews were designed with the intention of eliciting conversation to carefully produce various verbal and non-verbal behaviors. The goal was to study behavioral markers of depression and suicidal ideation during patients' responses. We selected three different question sets, such as Distress Assessment Questions (DQs) [21], Ubiquitous Questions (UQs) [32, 50], and Concept Questions (CQs) [23] to address the design goal. The dataset contains recordings of semi-structured medical interviews with 18 patients between the ages of 13 and 23 (average of 16.78 years with standard deviation of 2.90).

In this work, we focus on depression severity assessment based on the baseline session of 18 patients conducted at the University of Pittsburgh Medical Center in the United States. Eight patients identified as male and ten patients identified as female (two patients identified as a gender different from their sex at birth). Depression severity was assessed by a clinician at the end of each session using the Montgomery and Asberg Depression Rating Scale (MADRS) [29], unlike the DAIC-WOZ database in which depression severity was estimated based on self-reports [21]. Each patient received 25 USD compensation for participation in each interview session. The semi-structured medical interviews were audio and video recorded in a private examination room with controlled lighting and minimal distractions. Video recordings were collected using a webcam directed at the patients only, while audio recordings were collected using head-mounted microphones on both the clinician and patient. Speech segments of the clinician and patients were manually annotated at the utterance level using the ELAN annotation software [42].

Depression severity is reflected in the patients' MADRS scores, which range from 0 to 60 points and can be discretized into several categories: normal (0-6), mild symptoms (7-19), moderate symptoms (20-34), or severe symptoms (34-60). All patients in our dataset were undergoing treatment in an Intensive Outpatient Program at the time of medical interviews, and they had MADRAS scores above 6. Therefore, all adolescent patients had at least mild symptoms of depression. In our analysis, we formulate the depression severity assessment as a binary classification problem based on MADRAS scores. Our motivation to combine the moderate and severe categories [9] comes from a clinical perspective where a MADRS score in the mild category typically does not require treatment in a clinical setting, whereas a MADRS score in the moderate or severe category indicates a need for treatment referral [33, 39]. Therefore, being able to aid in the classification of mild vs. moderate or severe depression would have clinical utility. We split the 18 patients into two classes: patients with mild symptoms (4) and patients with moderate or severe symptoms (14).

3.1 Question Sets

There were three main parts of each interview: Distress Assessment Questions (DQs) [21], Ubiquitous Questions (UQs) [32, 50], and Concept Questions (CQs) [23]. The main goal was to study which

facial and vocal behavioral markers of depression were probed by those three different sets of assessment questions. We were interested in studying whether those behavioral makers could be discriminative to determine depression severity.

Distress Assessment Questions (DQs). This question set includes the six positively valenced questions (1-6) and six negatively valenced questions (7-12) shown in Table 1. It was originally developed to support the assessment of psychological distress conditions, such as anxiety, depression, and post traumatic stress disorder [21]. Previous research showed that these event-specific questions revealed nonverbal behavioral markers of psychological distress [43].

Table 1: List of 12 Distress Assessment Questions consisting of 6 positively valenced and 6 negatively valenced questions.

1. Are you from Pittsburgh originally?
2. Where do you go to school? What grade are you in?
3. Are you more of a people person or shy (or extrovert/introvert)?
4. What is your dream job? What would you like to do if you could do anything?
5. How close are you to your family?
6. Who has been a great/positive influence on your life?
7. What are the things that make you mad/pissed?
8. What's something you feel guilty about?
9. When was the last time you were annoyed/angry with someone?
10. Tell me about a situation that you wish you had handled differently?
11. What advice would you give yourself 5 years ago?
12. When was the last time you were really happy?

Ubiquitous Questions (UQs). This question set includes five open-ended questions selected to elicit conversational responses (Table 2). It was constructed to study vocal and facial behavioral markers in response to both positive and negative valence, and has been used in previous studies identifying adolescents with suicidal thoughts and behaviors [32, 50].

Table 2: List of 5 Ubiquitous Questions.

1. Do you have hope?
2. Do you have any fear?
3. Do you have any secrets?
4. Are you angry?
5. Does it hurt emotionally?

Concept Questions (CQs). This set includes three negatively valenced concepts and three positively valenced concepts (Table 3) that have previously been shown to generate discriminative brain imaging scans in subjects with and without suicidal thoughts [23].

Table 3: List of 6 Concept Questions consisting of 3 negatively valenced and 3 positively valenced questions.

What comes to mind when you think about the following concepts?
1. Trouble
2. Death
3. Cruelty
4, Carefree
5. Good
6. Praise

4 MULTIMODAL FEATURES

4.1 Acoustic Features

We extracted acoustic features from the patient audio recordings to use as input for depression severity analysis. Audio source separation between patient and clinician was very high quality because the audio signals were collected with separate microphones and simultaneous speech was infrequent. For 10 ms windows, we extracted selected acoustic features from the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [14], such as frequency-related features (i.e., Pitch, Jitter), Loudness, Shimmer, Harmonics to Noise Ratio (HNR) describing energy and amplitude of acoustic signals, Alpha Ratio (AR), Spectral Slope 0-500 (SS1), Spectral Slope 500-1500 (SS2) representing spectral balance features, and Mel-Frequency-Cepstral-Coefficients (MFCC) 1-4 that are spectral shape-related features. The GeMAPS is a standard acoustic feature set for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis.

4.2 Visual Features

We follow previous work on depression severity estimation [31, 44] found that these 17 AUs extracted by means of the OpenFace 2.0 [2] were more useful for predicting depression severity than other behavior markers i.e., descriptors of eye gaze and head pose [44]. The OpenFace 2.0 toolkit demonstrates state-of-the-art results in facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. We applied the OpenFace 2.0 toolkit [2] to the patient video recordings to estimate the frame-level occurrence of 17 action units (i.e., AUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45) from the Facial Action Coding System [12] that describes facial muscle movements.

4.3 Feature processing

Acoustic features were computed for the time intervals corresponding to patients' responses to each interview question and then were averaged per question. For visual features, we calculated activation frequency of each AUs for time intervals corresponding patients' responses to each of interview questions. To obtain robust behavioral markers of depression severity at the question set level and perform patient-independently depression severity assessment, we averaged the mean values of acoustic features and activation frequencies of AUs over each segment of three question sets, i.e. DQs, UQs, and CQs.

5 STATISTICAL ANALYSIS

Table 4: Median values of acoustic features for patients with mild depression (class mild) vs. patients suffering from moderate or severe depression (class severe). ∗ stands for statistically different medians with p-value < 0.05.

	DQs		UQs		CQs	
Feature	mild	severe	mild	severe	mild	severe
Pitch	14.590	21.902	14.678	20.861	15.415	21.350
Jitter	0.027*	0.040*	0.038	0.035	0.025*	0.038*
Loudn.	0.378	0.414	0.351	0.412	0.314	0.391
Shim.	0.899	0.966	1.089	0.943	0.878	0.982
HNR	1.768	4.176	1.380	4.050	2.008	3.583
AR	-12.649	-15.737	-13.543	-14.501	-15.953	-14.697
SS1	0.015	0.010	0.016	0.019	0.016	0.017
SS2	-0.015	-0.014	-0.015	-0.014	-0.017	-0.014
MFCC1	31.392	30.203	32.448	31.064	33.448	30.539
MFCC2	2.757	5.514	2.955	2.997	5.550	3.762
MFCC3	10.608	10.558	9.743	9.157	11.291	10.523
MFCC4	-5.263	-1.332	-3.499	-2.717	-5.421	-1.555

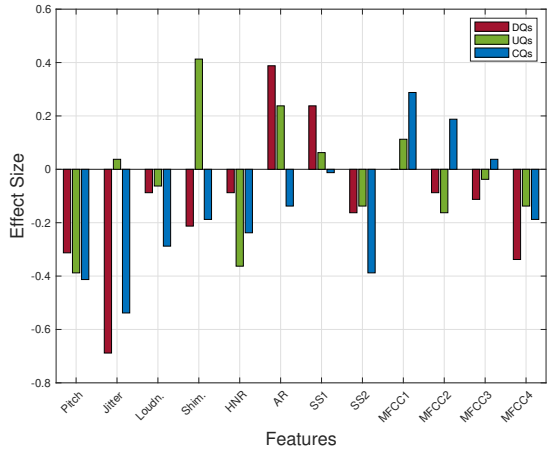

Figure 2: Effect size of Mann-Whitney U-test between class mild and severe of depression calculated for acoustic features.

One goal of our study is to investigate patients' responses to three different sets of questions (DQs, UQs, and CQs). We aim to explore the nonverbal content of these answers. Therefore, we carried out statistical analyses to determine which acoustic and visual features significantly differed between adolescents with mild depression (i.e., mild) and adolescents with moderate and severe depression (i.e., severe). We first ran Mann-Whitney U-tests to compare the medians of each acoustic and visual feature between the two classes, and then calculated the effect size (r) for the Mann-Whitney U-test [15]. We interpret a U-test as statistically significant if $p < 0.05$. Table 4 presents median values of acoustic features for both classes while Figure 2 shows effect sizes of depression for acoustic feature

Table 5: Median values of visual features for patients with mild depression (class mild) vs. patients suffering from moderate or severe depression (class severe).∗ stands for statistically different medians with p-value < 0.05.

	DQs		UQs		CQs	
Feature	mild	severe	mild	severe	mild	severe
AU 1	0.259*	0.179*	0.362	0.131	0.220	0.135
AU 2	0.182	0.151	0.271	0.178	0.281	0.162
AU 4	0.180	0.070	0.158	0.041	0.353	0.062
AU 5	0.697	0.232	0.755	0.177	0.776	0.122
AU 6	0.000	0.346	0.000	0.268	0.000	0.398
AU 7	0.000	0.051	0.000	0.071	0.000	0.053
AU 9	0.018	0.057	0.000	0.000	0.000*	0.029*
AU 10	0.009	0.205	0.093	0.157	0.000*	0.370*
AU 12	0.178	0.295	0.058	0.227	0.015*	0.361*
AU 14	0.030	0.436	0.163	0.362	0.018	0.569
AU 15	0.403	0.203	0.293	0.160	0.137	0.097
AU 17	0.162	0.119	0.300	0.155	0.262	0.116
AU 20	0.046	0.105	0.022	0.017	0.000	0.074
AU 23	0.245	0.399	0.270	0.431	0.165	0.392
AU 25	0.614	0.632	0.556	0.611	0.512	0.533
AU 26	0.499	0.408	0.646	0.400	0.390	0.364
AU 45	0.438	0.343	0.345	0.292	0.531	0.405

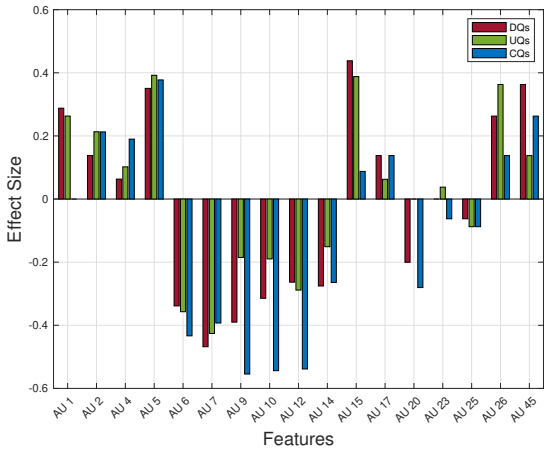

Figure 3: Effect size of Mann-Whitney U-test between classes mild and severe of depression calculated for visual features.

values enhanced during medical interviews. We observe that speech jitter (i.e., frequency instability) was significantly higher for the more severe class of patients during two of the three question sets. Table 5 reports median values of visual features for both classes of depression supported by effect sizes shown in Figure 3. We observe the trend of decreased activation frequency of facial AU 5 for severely depressed adolescents responding to three different sets of questions compared to mild depressed adolescents. Also, we find the trend of increased activation frequency of AU 6 and 7 for

severely depressed adolescents. The magnitude and direction of the effects supported by our significant statistical comparison reveal the high discriminability of the CQs. In particular, the segment of CQs shows significant differences in the activation frequency of AUs 9, 10, and 12 for two classes of depression severity. That might suggest that adolescent patients with moderate or severe depression displayed significantly more facial expressions related to these AUs when they were asked to discuss the concepts e.g., death. In addition to CQ set, it is worth mentioning DQ set emphasizes that adolescent patients with mild depression had significantly more often activated facial muscles described by AU 1 than the other patients.

6 DEPRESSION CLASSIFICATION

6.1 Depression classifiers

We selected three machine learning classifiers that are particularly well suited for small sample size datasets including Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Decision Tree (DT). Our main goal is to accurately predict unbalanced depression severity classes using acoustic and visual behavioral markers intended to study predictive power of the three different sets of questions. We selected acoustic and visual features that could be interpretable by clinicians and can lead to finding a nonverbal characterization of depression severity. To study the impact of each modality, we explored classifiers with only acoustic features, only visual features, or with both sets of features in both early and late fusion manners.

6.2 Experimental Setup

We performed 18-fold testing with nested 17-fold validation. All folds were patient-independent. The optimal hyperparameters were found during each validation step, we retrained a model using data from all 17 training patients and then used it for testing. Balanced accuracy was used as a validation metric for tuning hyperparameters. Classification performance during testing was calculated using balanced accuracy (B-ACC) and F1 score (F1) because the number of patients per class was strongly unbalanced.

We briefly describe hyperparameters of those three depression classifier below. We tuned only the hyperparameter $\gamma \in [0, 1]$ of LDA classifiers with a step of 0.1 that corresponds to an amount of regularization applied when estimating a covariance matrix of predictors. For KNN classifiers, we searched for an optimal value of the number K of nearest neighbors ranging from 1 to 10. We controlled the depth of the decision trees by using two hyperparameters: the minimum number of leaf node observations from the set $\{1, 4, 7, 10\}$ and the maximal number of decision splits from the set $\{1, 6, 11\}$.

All the hyperparameters of our three classifiers were tuned by performing a grid search over a set of possible hyperparameter values. An optimal value of a hyper-parameter was selected based on classification performance (i.e., balanced classification accuracy) on the validation set.

6.3 Results

In this section, we present details of our classification results for depression severity. Table 6 shows a summary of these results. We performed experiments for all three machine learning classifiers (i.e., LDA, KNN, and DT) with acoustic or visual features, and for the three question sets (i.e., DQs, UQs, and CQs). We also included early and late fusion of the multimodal features in our experiments. We used DT classifier for the late fusion since its predictions were most accurately for unimodal classification. For statistical analysis, we used McNemar test to compare the predictions of the different classifiers. The statistical tests were applied to study whether or not we could accurately classify the severity of depression based on acoustic and visual behavioral patterns of adolescents' responses. In Table 6, the symbol † was used when the best performing classifier was significantly better than the second best classifier while the symbol ‡ showed statistical difference between the best classifier and the third classifier.

Table 6: Depression classification performance: leave one patient out cross validation.

	LDA		KNN		DT	
Questions	B-ACC	F1	B-ACC	F1	B-ACC	F1
Acoustic						
DQs	0.80	0.78	**0.84**	0.84	**0.96**$^{\dagger\ \ddagger}$	0.93
UQs	0.68	0.68	0.71	0.73	0.75	0.60
CQs	**0.88**	0.91	0.71	0.73	0.82	0.70
Visual						
DQs	0.52	0.52	**0.80**	0.78	**0.96**$^{\dagger\ \ddagger}$	0.93
UQs	**0.71**	0.73	0.59	0.60	0.75	0.60
CQs	0.68	0.68	0.64	0.63	0.82	0.70
Early Fusion of Acoustic and Visual						
DQs	0.71	0.73	0.68	0.68	**0.96**†	0.93
UQs	0.63	0.65	0.63	0.65	**0.96**†	0.93
CQs	**0.75**	0.80	**0.75**	0.80	0.75	0.60
Late Fusion of Acoustic and Visual using DT						
DQs	0.93	0.86	0.93	0.86	**0.93**‡	0.86
UQs	0.89	0.80	**0.96**	0.93	0.75	0.60
CQs	**1.00**	1.00	**0.96**	0.93	0.71	0.55

To summarize, acoustic and visual behavioral markers are significantly predictive when we take into account the context of Distress Assessment Questions (DQs). Another observation from Table 6 is that early and late fusion of acoustic and verbal features strongly increase depression classification accuracy in the context of Concept Questions (CQs). That suggests that acoustic and verbal behavioral markers elicited by CQs are complementary to each other. This result is not observed when behavioral markers of responses to DQs are fused. Classification experiments were performed in a patient-independent manner, allowing us to study generalizability of the nonverbal behavioral markers. The decision tree classifier most accurately performed depression severity classification based on all combinations of acoustic and visual features for all three sets of questions. In general, late fusion outperformed early fusion. We

Table 7: Acoustic and visual feature importance estimates averaged over leave one patient out cross validation: 3 the most important features for each question set.

	DQs		UQs		CQs	
	feat.	est.	feat.	est.	feat.	est.
			Acoustic			
1.	Jitter	0.389	Shim.	0.026	SS2	0.146
2.	MFCC4	0.000	HNR	0.023	Jitter	0.069
3.	MFCC3	0.000	MFCC2	0.017	MFCC4	0.000
			Visual			
1.	AU 6	0.117	AU 5	0.102	AU 9	0.217
2.	AU 5	0.078	AU 7	0.062	AU 12	0.041
3.	AU 45	0.000	AU 45	0.017	AU 45	0.000

also analyzed which features the classifiers selected to discriminate between classes. We focused on the DT classifier since it best performed among the three classifiers.

Table 7 presents average importance estimates for acoustic and visual features, respectively. We observe high values for speech jitter. That is in line with the results of our statistical analysis for acoustic features described in Section 5. For visual behavioral markers, we observe that facial action unit 5 and 6 are informative, especially for DQ context. In the UQ context, AU 5 and 7 are also helpful. In addition, AU 9 and 12 are informative in the CQ context.

7 DISCUSSION

In this section, we discuss the three research questions introduced in Section 1. we also discuss potential limitations of our work and present open issues with regard to the selected modalities and nonverbal behavioral markers, the number of patients in the study, and classifier selection.

RQ1: Our experiments and analyses identified a set of interpretable acoustic and visual features that are predictive for depression severity of adolescents at high risk of mental disorders. The knowledge of nonverbal behavioral markers could eventually help clinicians to distinguish patients having mild depression from patients suffering from moderate or severe depression. Our statistical analyses of acoustic features show that speech jitter is strongly informative to discriminate between depression severity [30]. Speech jitter measures frequency instability in speech [14]. We observe that speech jitter is higher for adolescents suffering from moderate or severe depression than adolescents with mild depression. This effect is strongest in the DQ and CQ context. Both sets of questions contain questions for both positive and negative valence that might accentuate that relationship. We found acoustic frequency instability as a behavioral marker could help to assess the depression severity for adolescent patients.

Our depression severity classification experiments using facial expressions were also aligned with our statistical analysis of action units. While responding to all three question sets, activation frequency of AU 5 was lower for adolescents suffering from moderate or severe depression when compared adolescents having mild depression. AU 5 corresponds to the upper lid raiser and is involved in the prototypical expressions of anger, fear, and surprise [26]. Also, the activation frequency of AU 6 and 7 increases for adolescents suffering from moderate or severe depression. This observation was also found for all three different sets of questions. AU 6 describes movements of cheek raiser connected to the prototypical expression of happiness and joy while AU 7 corresponds to movements of lid tightener associated with the prototypical expressions of anger and fear [26]. Those trends of AU 5, 6 and 7 might suggest that adolescents with moderate or severe depression express less positive emotional facial expressions and more negative emotional facial expressions than adolescents having mild depression.

RQ2: Our goal was to study three different question sets, namely the Distress Assessment Questions (DQs), the Ubiquitous Questions (UQs), and the Concept Questions (CQs). When analyzing differences across these three question sets, we found speech jitter in the DQ and CQ context, facial AU 1 in the DQ context, and facial AU 9, 10, and 12 in the CQ context as significantly informative features. In particular, speech jitter best discriminates when the question set contains both positively and negatively valenced questions (i.e., DQs and CQs). We observe that responses to these question sets include the speech with high frequency instability. AU 1 is associated with inner brow raiser to and it is linked to feelings of fear and sadness. The frequency activation of AU 9, 10 and 12 was higher for moderately or severely depressed adolescents than for mildly depressed adolescents. AU 9 describes movements of nose wrinkler and AU 10 corresponds to upper lip raiser. Both action units are associated with feelings of disgust. AU 12 is connected with movements of lip corner puller when contempt is felt.

One might hypothesize that these adolescents might experience feelings of contempt, disgust and nervousness because AU 9, 10 and 12 were frequently activated and speech jitter was high. When we assessed the results of our statistical analyses and the importance of visual features selected by the decision tree classifier, facial action unit 9 and 12 in the CQ context most accurately distinguished between adolescents having mild depression and adolescents suffering from moderate or severe depression.

RQ3: Our classification experiments show non-verbal behavioral markers that can help to distinguish between adolescents with mild depression and adolescents with moderate or severe depression. All three classifiers achieved promising performance for depression severity prediction. Unimodal classifiers with acoustic or visual behavioral markers lead to satisfactory classification performance, retaining good interpretablity of results and learned features. Late fusion generally improved depression severity classification. This result hints at the fact that acoustic and visual modalities may have complementary information.

7.1 Discussion on Potential Limitations

In our work, we were able to analyze non-verbal responses of 18 participants who had to be 13-25 years old, fluent in English, and currently undergoing treatment in an intensive outpatient program for treatment of severe depression and/or suicidality at the time of enrollment. We limited our analysis to non-verbal interpretable features that can be automatically extracted from video recordings.

Symptoms of depression can be reflected in different multimodal channels. The importance of those multimodal channels is not

the same for depression severity diagnosis. Each patient can have slightly different behavioral responses to the same set of questions. Facial expressions and speech can be affected by age, gender, and cultural differences. Furthermore, nonverbal responses of patients can vary from one person to another due to mental comorbidities, such as suicidal ideation and post-traumatic stress disorder. A lot of information on depression severity might be encoded in the content of adolescents' responses. Language could be a complementary modality to acoustic and visual modality.

The possibly largest caveat of our study is that we analyzed a small population of adolescents at high risk of mental disorders. Although our conclusions are supported by the magnitudes of effect sizes and performance of depression classification, we cannot generalize about all patients suffering from depression and/or with various mental comorbidities based on such a small number of adolescent patients.

We developed our experimental methodology to analyze our small sample size dataset with unbalanced classes of depression severity. For example, we selected classifiers that do not require tuning multiple hyper-parameters and have good interpretability.

8 CONCLUSION

In this work, we investigated depression severity assessment for adolescents at high risk of mental disorders. We explored both facial and voice behavioral markers. We also studied three different interview contexts: Distress Assessment Questions, Ubiquitous Questions, and Concept Questions. The goal of our study was twofold: studying a unique population of adolescents at high risk of mental disorders and differentiating mild depression from moderate or severe depression. Our experimental methodology reflects best practise used for analyzing small sample size and unbalanced datasets of unique patients.

Our results showed some interesting differences happening with different question contexts. We revealed the strong relationship between some acoustic and visual behavioral markers and depression severity. These results are a good step in building healthcare decision support tools for adolescent populations at high risk of mental disorders. This can also help to support mental health diagnosis, clinical decision making, and therapy planning.

In the future, we plan to expand our study to investigate mental comorbidities like suicidal ideation and post-traumatic stress disorder that have also been reported. Moreover, we would like to extend our evaluation of depression severity to longitudinal assessment of vowel space in adolescents' responses in order to identify a wide range behavioral markers of psychological conditions.

ACKNOWLEDGMENTS

MM was supported by the Swiss National Science Foundation (#P2GEP2_184518). JZ was supported by the National Institute of Mental Health (#T32MH018951-27). JG and LP were partially supported by the National Science Foundation (#1722822, #1734868) and the National Institute of Mental Health (#5R01MH096951-07, #U01MH116925 and #U01MH116923).

REFERENCES

[1] Mohamad Al Jazaery and Guodong Guo. 2018. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing* (2018).

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*. IEEE, 59–66.

[3] Alizah Z Brozgold, Joan C Borod, Candace C Martin, Lawrence H Pick, Murray Alpert, and Joan Welkowitz. 1998. Social functioning and facial emotional expression in neurological and psychiatric disorders. *Applied Neuropsychology* 5, 1 (1998), 15–23.

[4] Yulia E Chentsova-Dutton, Jeanne L Tsai, and Ian H Gotlib. 2010. Further evidence for the cultural norm hypothesis: Positive emotion in depressed and control European American and Asian American women. *Cultural Diversity and Ethnic Minority Psychology* 16, 2 (2010), 284.

[5] Gary Christopher and John MacDonald. 2005. The impact of clinical depression on working memory. *Cognitive neuropsychiatry* 10, 5 (2005), 379–399.

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

[7] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71 (2015), 10–49.

[8] John K Darby and Harry Hollien. 1977. Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica* 29, 4 (1977), 279–291.

[9] David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert A Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*. 193–202.

[10] Abhinav Dhall and Roland Goecke. 2015. A temporally piece-wise fisher vector approach for depression analysis. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 255–259.

[11] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. 2017. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics* 22, 2 (2017), 525–536.

[12] W. V. Ekman, P. Friesen and J. Hager. 2002. Facial action coding system: A technique for the measurement of facial movement. *Research Nexus* (2002).

[13] Heiner Ellgring. 2007. *Non-verbal communication in depression*. Cambridge University Press.

[14] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[15] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* 141, 1 (2012), 2.

[16] Wolfgang Gaebel and Wolfgang Wölwer. 2004. Facial expressivity in the course of schizophrenia and depression. *European archives of psychiatry and clinical neuroscience* 254, 5 (2004), 335–342.

[17] Jean-Guido Gehricke and David Shapiro. 2000. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research* 95, 2 (2000), 157–167.

[18] Jeffrey M Girard and Jeffrey F Cohn. 2015. Automated audiovisual depression analysis. *Current opinion in psychology* 4 (2015), 75–79.

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

[20] Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 69–76.

[21] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews.. In *LREC*. Citeseer, 3123–3128.

[22] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. 2013. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders* 147, 1-3 (2013), 17–28.

[23] Marcel Adam Just, Lisa Pan, Vladimir L Cherkassky, Dana L McMakin, Christine Cha, Matthew K Nock, and David Brent. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature human behaviour* 1, 12 (2017), 911–919.

[24] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 35–42.

[25] Colin D Mathers and Dejan Loncar. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *Plos med* 3, 11 (2006), e442.

[26] David Matsumoto and Paul Ekman. 2008. Facial expression analysis. *Scholarpedia* 3, 5 (2008), 4237.

[27] Alexander McGirr, Johanne Renaud, Monique Seguin, Martin Alda, Chawki Benkelfat, Alain Lesage, and Gustavo Turecki. 2007. An examination of DSM-IV depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study. *Journal of Affective Disorders* 97, 1-3 (2007), 203–209.

[28] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 21–30.

[29] Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry* 134, 4 (1979), 382–389.

[30] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering* 51, 9 (2004), 1530–1540.

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

[32] John P Pestian, Michael Sorter, Brian Connolly, Kevin Bretonnel Cohen, Cheryl McCullumsmith, Jeffry T Gee, Louis-Philippe Morency, Stefan Scherer, Lesley Rohlfs, and STM Research Group. 2017. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide and Life-Threatening Behavior* 47, 1 (2017), 112–121.

[33] Laura A Pratt. 2014. *Depression in the US household population, 2009-2012*. Number 172. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

[34] Lawrence Ian Reed, Michael A Sayette, and Jeffrey F Cohn. 2007. Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of abnormal psychology* 116, 4 (2007), 804.

[35] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2019. MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech. *arXiv preprint arXiv:1909.07208* (2019).

[36] Babette Renneberg, Katrin Heyn, Rita Gebhard, and Silke Bachmann. 2005. Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry* 36, 3 (2005), 183–196.

[37] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 3–12.

[38] Jonathan Rottenberg, Karen L Kasch, James J Gross, and Ian H Gotlib. 2002. Sadness and amusement reactivity differentially predict concurrent and prospective functioning in major depressive disorder. *Emotion* 2, 2 (2002), 135.

[39] Albert L Siu. 2016. Screening for depression in children and adolescents: US Preventive Services Task Force recommendation statement. *Annals of internal medicine* 164, 5 (2016), 360–366.

[40] Denise M Sloan, Milton E Strauss, Stuart W Quirk, and Martha Sajatovic. 1997. Subjective and expressive emotional responses in depression. *Journal of affective disorders* 46, 2 (1997), 135–141.

[41] Denise M Sloan, Milton E Strauss, and Katherine L Wisner. 2001. Diminished response to pleasant stimuli by depressed women. *Journal of abnormal psychology* 110, 3 (2001), 488.

[42] Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

[43] Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. Multimodal Analysis and Estimation of Intimate Self-Disclosure. In *2019 International Conference on Multimodal Interaction*. 59–68.

[44] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. 2020. Spectral Representation of Behaviour Primitives for Depression Analysis. *IEEE Transactions on Affective Computing* (2020).

[45] Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 61–68.

[46] Jeanne L Tsai, Nnamdi Pole, Robert W Levenson, and Ricardo F Muñoz. 2003. The effects of depression on the emotional responses of Spanish-speaking Latinas. *Cultural Diversity and Ethnic Minority Psychology* 9, 1 (2003), 49.

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

[48] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.

[49] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.

[50] Verena Venek, Stefan Scherer, Louis-Philippe Morency, John Pestian, et al. 2017. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing* 8, 2 (2017), 204–215.

[51] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. 65–72.

[52] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. 2013. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 41–48.

[53] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 89–96.

[54] Le Yang, Dongmei Jiang, and Hichem Sahli. 2018. Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures. *IEEE Transactions on Affective Computing* (2018).

[55] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2018. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing* (2018).