The Turkish Audio-Visual Bipolar Disorder Corpus

Elvan Çiftçi
Department of Psychiatry
Cizre State Hospital
Cizre, Şırnak, Turkey
elvanlciftci@gmail.com

Heysem Kaya
Department of Computer Engineering
Namık Kemal University
Çorlu, Tekirdağ, Turkey
hkaya@nku.edu.tr

Hüseyin Güleç
Department of Psychiatry
Erenköy Psychiatric and Neurological Diseases Training and Research Hospital
Istanbul, Turkey
huseyingulec@yahoo.com

Albert Ali Salah
FVCRC, Nagoya University, Japan

Department of Computer Engineering, Boğaziçi University
Bebek, İstanbul, Turkey
salah@boun.edu.tr

***Abstract*— This paper introduces a new audio-visual Bipolar Disorder (BD) corpus for the affective computing and psychiatric communities. The corpus is annotated for BD state, as well as Young Mania Rating Scale (YMRS) by psychiatrists. The paper also presents an audio-visual pipeline for BD state classification. The investigated features include functionals of appearance descriptors extracted from fine-tuned Deep Convolutional Neural Networks (DCNN), geometric features obtained using tracked facial landmarks, as well as acoustic features extracted via openSMILE tool. Furthermore, acoustics based emotion models are trained on a Turkish emotional database and emotion predictions are cast on the utterances of the BD corpus. The affective scores/predictions are investigated with linear regression and correlation analyses against YMRS declines to give insights about BD, which is directly linked with emotional lability, i.e., quick changes in affect.**

***Index Terms*—bipolar disorder, audio-visual corpus, affective computing, multi-modal analysis**

I. Introduction

Bipolar Disorder (BD) lifelong prevalence is 2.1% worldwide, subthreshold forms affects 2.4% [1]. According to World Health Organization [2], it ranks among the top ten for young adults in the diseases of disability-adjusted life year (DALY) indicator.

Treatment resistance is one of the big challenges for bipolar disorder [3]. Despite advances in bipolar disorder treatment, remission rate and treatment compliance are low. Delay in the diagnosis of bipolar disorder, recurrent hospitalizations, low response to available treatment, and increase in inflammation due to insufficient recognition of depressive episodes cause treatment resistance.

Recently, machine learning studies are effectively applied to identify and classify mood disorders (MD) [4,5]. Since MD is directly linked to the affective states, affective computing and social signal processing methods can bring innovative approaches to the MD identification problem. Works on affective computing have been centered on audio-visual signal processing and machine learning, with support from clinicians and psychologists in annotating data and evaluating outcomes. A large set of affective issues are tackled, ranging from short-term states (e.g. laughter, emotion), to mid-term disorders (e.g. depression, bipolar disorder) and to long-term traits (e.g. personality traits) [6,7]. Social signal processing also pays attention to social and non-verbal signals, and investigates social interactions, and dialogues [8].

In this work, we apply signal processing and machine learning methods for recognition of BD from short video clips of subjects performing a small battery of affective tasks. A corpus is collected from 46 patients and 49 healthy controls. We implement several multimodal approaches for BD classification, and describe strong baselines. The contributions of this work are the new audio-visual BD corpus, as well as the comparative experimental results for several approaches to the problem. The core aim of the efforts on the corpus is to find biological markers/predictors of treatment response via signal processing and machine learning techniques to reduce treatment resistance. During treatment period, these biological markers can also help early detection of relapses. These discriminative markers are intended for early recognition of the bipolar disorder. Collectively, they are expected to provide an insight for personalized treatment of bipolar patients [9].

The rest of the paper is organized as follows. In the next section, we introduce the corpus, which remains unnamed to preserve double blind reviewing conditions. In section III, the methodology followed in multimodal analyses are explained, while Section IV presents the experimental results. Discussion and conclusions are given in Sections V and VI, respectively.

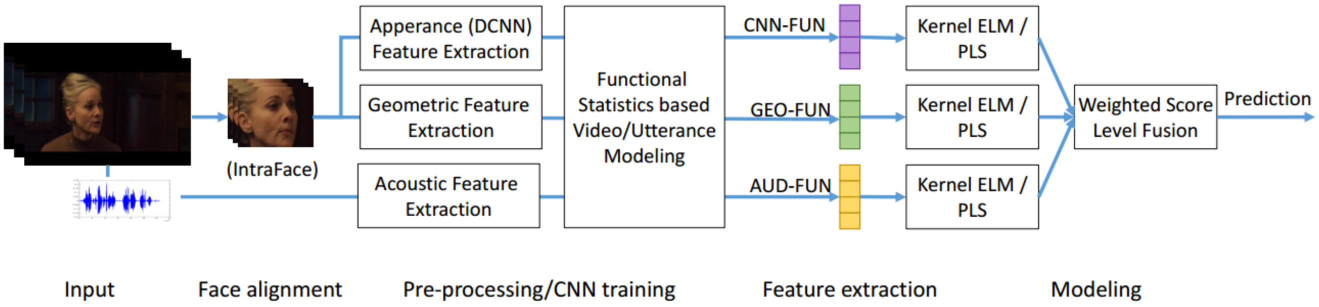
Apperance (DCNN) Feature Extraction
Geometric Feature Extraction
Acoustic Feature Extraction
(IntraFace)
Functional Statistics based Video/Utterance Modeling
CNN-FUN
GEO-FUN
AUD-FUN
Kernel ELM / PLS
Kernel ELM / PLS
Kernel ELM / PLS
Weighted Score Level Fusion
Prediction
Input
Face alignment
Pre-processing/CNN training
Feature extraction
Modeling

Fig. 1. Audio-visual processing pipeline for Bipolar Disorder classification from video

II. The Corpus

A. Participants

We have collected video data from patients with BD and healthy controls. 35 male and 16 female patients were recruited from the mental health service of a hospital. Inclusion criteria were as follows: (I) diagnosis of BD type I, manic episode according to DSM-5 [10] given by the following doctor, (II) being informed of the purpose of the study and having given signed consent before enrollment. Exclusion criteria were as follows:

I. being younger than 18 years or older than 60 years,
II. showing low mental capacity during interview,
III. expression of hallucinations and disruptive behaviors during interview,
IV. presence of severe organic disease,
V. presence of any organic disease that may affect cognition,
VI. having less than five years of public education,
VII. diagnosis of substance or alcohol abuse in the last three months (except nicotine and caffeine),
VIII. presence of cerebrovascular disorder, head trauma with longer duration of loss of consciousness, severe hemorrhage and dementia,
IX. having electroconvulsive therapy in the last one year.

For the healthy control group, the following additional criteria were considered for exclusion (I) presence of family history of mood or psychotic disorder, and (II) presence of psychiatric disorder during interview or in the past. Approval of Hospital Clinic Ethics Committee was obtained prior to data collection.

More than a hundred subjects participated the work, but recordings for some subjects were excluded due to the aforementioned criteria, and four subjects did not give approval for sharing the data. Consequently, data from 46 patients and 49 healthy controls (95 subjects) were retained for future experimentation and sharing. In this work, results on a subset of the original corpus (collected from 89 subjects) are reported.

B. Assessment and Data Collection

In order to gather sociodemographic and clinical information, all patients were assessed with semi-structured interviews based on the SKIP-TURK [11]. This form includes: identity, sociodemographic personal and family information, age at disease onset, severity, clinical presentation and used treatments.

During hospitalization, in every follow up day (0th- 3rd- 7th- 14th- 28th day) and after discharge on the 3rd month, presence of depressive and manic features were evaluated using Young Mania Rating Scale (YMRS) [12] and Montgomery-Asberg Depression Rating Scale (MADRS) [13]. In every follow up day, audiovisual recording is done by a video camera. Thus, each video session is separately annotated for bipolar mania/depression ratings.

Video recording is done by a presentation guide including seven tasks such as explaining the reason to come to hospital/participate in the activity, describing happy and sad memories, counting up to thirty, explaining two emotion eliciting pictures (see Fig.2). To increase the challenge of automatic discrimination, the Control Subjects were recorded with two additional conditions, where they are asked to portray mania and depression conditions. The collected corpus will be publicly available.

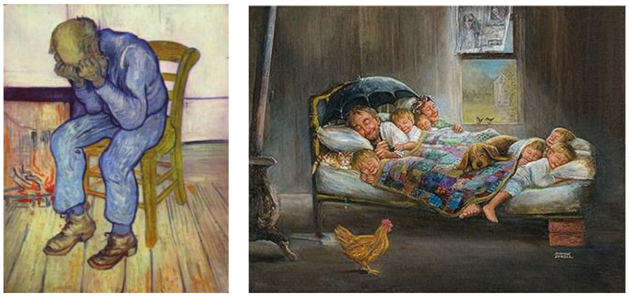

Figure 2. (left) van Gogh's *Depression* (right) Dengel's *Home Sweet Home*

III. Methodology for System Development

To provide a benchmark system and insight about the data, we investigated two different approaches. The first is a direct approach to classify video sessions into BD and normal classes, using audio-visual features. The second is an indirect approach, where we use emotion predictions for Pearson correlation analysis and regression of YMRS drop. The flowchart of the method proposed for direct classification of BP from video is given in Figure 1. The explanation of its components, and the details for the indirect approach are both given in the following subsections.

A. *Audio Processing: Acoustic Feature Extraction*

Speech utterances are first separated from the video signal and then a standard set of acoustic descriptors are extracted from the audio signal using the open-source openSMILE tool [14]. In order to obtain acoustic low level descriptors (LLDs) from audio, we use the 76-dimensional (38 raw, 38 temporal derivative) standard feature set used in the INTERSPEECH 2010 paralinguistic challenge as baseline [15]. These LLDs cover important speech signal characteristics, including prosody (energy, Fundamental Frequency – F0), voice quality features (jitter and shimmer), as well as Mel Frequency Cepstral Coefficients, which are commonly used in many speech technologies.

The LLDs are typically extracted from 25-40 seconds-length windows of the speech signal, and then summarized over the utterance. The most popular approach for summarization (also referred to as the utterance representation) is based on functionals, such as moments, extremes, coefficients of polynomials fit on the LLD contours, giving state-of-the-art results [15, 16].

In our preliminary work, we apply functionals over the whole utterance and compare performances of two sets of functionals. The first set consists of INTERSPEECH 2010 baseline features. These give a supra-segmental set of 1.582 acoustic features. Note that some functionals are not applied to all LLDs in this baseline set. The second is our proposed set of 10 functionals, which we apply on all LLDs to obtain a supra-segmental feature set of 76 x 10 = 760 dimensions. These 10 functionals are as follows: Mean, standard deviation, curvature coefficient (the leading coefficient of the quadratic polynomial [ax^2+bx+c] fit on the LLD contour), slope and offset (coefficients from the linear polynomial [ax+b] fit on the LLD contour), minimum value and its relative position, maximum value and its relative position, and the range (max. value – min. value).

We subsequently summarize the LLDs over analysis windows of length l = {10, 20, 30, 40, 50, 60} seconds with 50% overlap. The aim here is to find the optimal analysis window and to compare it against clip-level summarization.

For the second, indirect approach, we train an emotion classifier on a different corpus, and use it as a feature extractor. Recognition of emotions is related with all affective states/traits and the knowledge to recognize emotions from audio and video can be transferred to other related problems [7,17]. For example, the dynamics of short-term (3-5 sec.) emotional state predictions in audio/video recordings can be used to better explain and predict unipolar/bipolar disorder and to discriminate these from healthy controls. Emotional databases are collected in popular languages (e.g. English [18], German [19], Russian [20] and recently in Chinese [21]), however in languages such as Turkish the language resources are scarce. To alleviate this problem, recently there is an increasing research interest in cross-corpus/cross-language emotion recognition, which enables benefiting from available corpora effectively on original problems in a target corpus.

For training the speech based emotion classifier, we use the BUEMODB corpus having 484 utterances: 11 affirmative sentences portrayed for four basic emotions (Anger, Happiness, Neutral, and Sadness) by 11 amateur theatre actors/actresses [22]. The portrayal scenario is based on Stanislavsky Effect, where the actor imagines himself in a situation that arouses the intended emotion [23]. Three affect classifiers are trained, one with the original four classes, one for valence (showing the pleasantness of emotion) and one for arousal (reflecting the activity/vitality of the speaker), respectively. The arousal and valence labels are obtained by clustering the original classes into corresponding binary labels (e.g. Happy to positive valence and high arousal, Anger to negative valence and high arousal). The emotion predictions (scores and labels) from these affect classifiers are then used as mid-level features. We apply the proposed functionals on 4-second analysis windows with 2 second shifts (50% overlap), and cast predictions for each short-term clip.

B. *Video Processing and Visual Features*

Approximately 2.2 million images are collected over all videos, each of which having a frame rate of 30 Hz. On each image, the faces are detected, cropped and registered using the method proposed by Xiong and de la Torre [24] with additional Procrustes analysis for frontalization, as in [25]. The registered faces are then saved as 128x128 pixel grayscale images. Along with the faces, we record both the original and the aligned landmark points for subsequent geometric feature extraction.

From each face, we extract 23 geometric features as suggested by [25] for video based emotion recognition in uncontrolled conditions. Also, we extract appearance descriptors from registered faces, using a pre-trained Deep Convolutional Neural Network (DCNN) fine-tuned on a face based emotion corpus. This DCNN is recently applied on emotion and apparent personality trait recognition in uncontrolled conditions, giving state-of-the art results in both of these challenging tasks [26]. Using DCNN, we extracted 4096 dimensional features from the last convolutional layer. These image level descriptors are then summarized over fix-length analysis windows (10, 20, and 40 seconds tested) with 50% overlap using functional statistics, as described in the previous subsection.

C. *Classification methods*

Feature vectors extracted from audio and video are modelled using Partial Least Squares (PLS) regression and Extreme Learning Machines classifiers. Both of these machine learning methods learn projection matrices that map input features to the target (dependent) variable via regularized least squares, and hence, they are both fast and accurate [25, 26], Each method has a parameter that regularizes the model complexity against the training set classification error. To optimize these parameters, each method is trained with 10 values and best validation set results are reported. For brevity, the technical descriptions of the classifiers are excluded in this work, interested readers are referred to [26].

D. *Regression on Predicted Emotion Scores*

To assess the extent that the cross-corpus acoustic emotion predictions on the zeroth day recording of the Bipolar patient can "predict" the clinical progress (i.e. response to treatment), we devised meta-variables from clinician-annotated scores. Let $\{T_i\}$, i=1..7 denote the ground truth YMRS such that T_1=0th day, T_2=3rd day, T_3=7th day, T_4=14th day, T_5= 28th day, T_6=90th day score; the meta target variables are set as $X_1=T_2/T_1$, $X_2=T_3/T_1$, $X_3=T_4/T_1$, $X_4=T_5/T_1$ and $X_5=T_6/T_1$. These ratios are stepwise regressed to functionals of valence, arousal and basic emotion scores, cast on the first audio recording.

IV. EXPERIMENTAL RESULTS

A. *Preliminary Analyses: Demographics and Session Lengths*

A total of 50 bipolar manic episode (34 male and 16 female) patients aged between 18 to 54 years and 39 healthy controls (23 male and 16 female) aged between 18 to 57 years are included in the analyses reported here. Sociodemographics and clinical characteristics of the groups are given in Table I.

TABLE I. DEMOGRAPHIC AND CLINICAL CHARACTERISTICS OF BIPOLAR DISORDER AND HEALTHY CONTROL GROUP. (ED.: EDUCATION IN YEARS, TE: TOTAL EPISODE, TDD: TOTAL ILLNESS DURATION.)

	People with BP-1			**Healthy Controls**	**t/x^2**	**p**
	Female	**Male**	**All**			
AGE	40.2±8.8	35.02±10.6	36.7±10.3	37.3±10.9	0.36	0.72
ED	12.6±2.9	9.5±3.3	10.5±3.5	11.2±3.7	0.89	0.11
TE	7.13±7.7	7.67±5.7	6.26±6.4	-	0.71	0.48
TID	15.9±9.9	12.02±9.7	13.07±9.8	-	1.41	0.16

TABLE II. VIDEO LENGTH STATISTICS FOR HEALTHY CONTROLS AND MANIA PATIENTS WITH VARYING LEVELS

Diagnosis	**Number of videos**	**Average Time (s.)**	**Standard Deviation**
Healthy	120	138.9	68.4
Remission	62	151.9	65.4
Hypomania	82	221.1	171.4
Mania	88	276.4	246.3

The most remarkable result at a first glance is that the average response time for manic patients is longer than healthy controls (see Table II). When the data are subdivided into four groups as healthy, remission, hypomania, and mania according to the YMRS total score, it can be observed that average time increases gradually. However, due to increase in the standard deviation of hypomania and mania, this feature alone is not sufficient for discrimination of the disorder.

B. *Direct Approach: Audio-Visual Mania Classification*

For classification experiments, healthy and bipolar subjects are evenly distributed considering gender: a total of 182 videos of 44 subjects are set as the training group; the rest, 170 videos of remaining 45 subjects are arranged as the test group. Apart from the binary (healthy/bipolar) classification task, we also grouped the bipolar subjects's YMRS scores annotated at session level into three disjoint groups, thus obtaining a ternary classification task. Let Y_t denote YMRS score of session t, bipolar sub-groups are arranged as follows:

1. Remission: $Y_t <= 7$
2. Hypomania: $7 < Y_t < 20$
3. Mania: $Y_t >= 20$.

Both classification tasks are tackled using the pipeline shown in Fig. 1. In the following, we provide component-level results for this pipeline. Unweighted Average Recall (UAR), which is mean of class-wise recall scores, is commonly used as performance measure, instead of accuracy, which can be misleading in the case of class-imbalance. In the subsequent machine learning experiments, we report UAR as performance measure. In K-class classification, UAR has a constant chance-level baseline of 1/K.

Firstly, the LLDs are summarized over the whole session. Comparing the baseline set of functionals from IS2010 configuration and the proposed set of 10 functionals on the 76 acoustic LLDs, we observe that using a smaller set of functionals yields significantly better performance with ELM, whereas PLS gives similar performances (see Table III). All results reported in Table III are found significantly higher than the chance-level baseline (McNemar's Test [27], $p < 0.01$). This result means that the proposed system can differentiate bipolar mania/ hypomania/ remission from healthy control, simulation of mania and depression. Summarizing LLDs over analysis windows length of {10, 20, 30, 40, 50, 60} seconds with 50% overlap was not found to increase the success rate.

TABLE III. VALIDATION SET PERFORMANCES FOR BASELINE AND PROPOSED FUNCTIONALS APPLIED ON THE WHOLE UTTERANCE (IS10: BASELINE INTERSPEECH 2010 FUNCTIONAL SET, P10: PROPOSED 10 FUNCTIONALS)

Functional Set	**Dimensions**	**PLS**	**ELM**
IS10	1582	65.7	64.8
P10	760	65.3	**69.4**

Next, geometric and appearance descriptors are investigated using the same set of functionals over the whole session. Table IV reports comparative performances of individual audio and video features summarized over the whole session (i.e. one supra-segmental feature vector for each recording). In the Table, GEO23 denotes the frame level 23 geometric features, CNN4096 represents 4096 dimensional features from the last convolutional layer of the FER fine-tuned CNN. We observe that the best audio and visual performances are similar for the binary task; however, they differ in the ternary mania level classification task. Session level summarization of geometric features extracted from each face gives an UAR of 67.3%, while 4096 dimensional features from convolutional neural network trained with just average functionals reach an UAR score of 69.9%. Interestingly, acoustic and appearance descriptors perform better than geometric features in discriminating between healthy and bipolar subjects; however, their best result in the ternary task (48.6%) is lower compared to the one obtained with geometric features (51.5%). We also observe that removing the "range" functional effectively improves performance of the geometric features in the binary task.

TABLE IV. COMPARISON OF AUDIO-VISUAL FEATURES FOR BINARY (MANIA VS. HEALTHY CONTROL) AND TERNARY (REMISSION, HEALTHY AND MANIA) CLASSIFICATION. P9: PROPOSED FUNCTIONALS EXCEPT RANGE. RESULTS IN **BOLD** DENOTE SIGNIFICANTLY BETTER SCORES COMPARED TO CHANCE-LEVEL BASELINE (P < 0.01, MCNEMAR'S TEST).

Feature Attributes			**Mania/Control**		**Remission/Hipomania/ Mania**	
Descriptor	***Func.***	***Dim.***	***PLS***	***ELM***	***PLS***	***ELM***
IS10	P10	760	**65.3**	**69.4**	**48.5**	**48.6**
CNN4096	Average	4096	**69.9**	**65.2**	42.3	43.7
GEO23	P10	230	**64.7**	58.3	**48.6**	**51.3**
GEO23	P9	207	**67.3**	60.0	**47.1**	**51.5**

When the DCNN based appearance descriptor is summarized via mean and range functionals over sub-clips and the decisions are voted at video level, an UAR performance of 72.2% is obtained (Table V). Finally, the best results from each feature-level model (namely geometric, appearance, and acoustic models) are fused with equal weight. Fusion improves the binary classification UAR slightly to 73%, while the performance in the ternary (mania/hypomania/remission) classification task is observed to improve markedly to 55.6%.

TABLE V. UAR PERFORMANCE COMPARISON FOR SUMMARIZING 4096-DIMENSIONAL APPEARANCE FEATURES EXTRACTED FROM DCNN OVER VIDEO SUB-CLIPS OF VARYING WINDOW LENGTHS AND VIDEO LEVEL DECISION FUSION. STATISTICALLY SIGNIFICANT RESULTS (P < 0.01, MCNEMAR'S TEST) ARE SHOWN IN BOLD.

Analysis Window (sec.)		**Mania/Control**		**Remission/Hipomania/Mania**	
Window Length	***Shifts***	***PLS***	***ELM***	***PLS***	***ELM***
40	20	**71.3**	61.8	44.4	43.6
20	10	**72.2**	61.8	43.1	**47.3**
10	5	**72.2**	60.9	38.3	45.5

C. Indirect Approach: Analysis of Predicted Affect Primitives

After cross-corpus emotion recognition, clip-level functional statistics of predictions of affect classifiers used as midlevel features. These scores are regressed against the YMRS decline ratios as explained in Section III-D. This regression aims to find which mid-level features extracted from the first audio recordings can predict the YMRS decline (i.e. treatment response) observed in future sessions. When arousal and valence scores were regressed, the slope and relative position of the minimum value of arousal is observed to predict YMRS decline on the third day. When four basic affect scores are analyzed, standard deviation of neutral affect and mean value of sadness are observed to be predictors of YMRS decline on the third day. This shows that these audio parameters can be used as treatment predictors. Other potential treatment response predictors are shown in the Table VI.

TABLE VI. REGRESSING YMRS DECLINES ON PREDICTED AFFECTIVE SCORES. ONLY PREDICTORS HAVING STATISTICALLY SIGNIFICANT RELATIONSHIP WITH THE TARGET VARIABLE ARE LISTED.

Target	**Affective Predictors**	**B**	**t**	**p**
X_1	Slope of arousal score	21.653	2.761	0.008
	Relative position of minimum value of arousal score	0.264	2.103	0.041
	Standard deviation of neutral score	2.067	3.667	0.001
	Mean of sadness score	0.553	2.251	0.029
X_2	Standard deviation of valence	1.720	3.209	0.003
	Standard deviation of neutral score	2.606	3.315	0.002
X_3	Curvature of valence score	-19.903	-2.391	0.023
	Curvature of happiness score	347.699	3.018	0.005
	Relative position of maximum value of neutral score	0.320	2.549	0.016
	Mean of happiness score	-0.602	-2.745	0.010
X_4	Range of arousal score	0.279	2.160	0.047

V. DISCUSSION

On the basis of DSM-5 criteria related with talkativeness and pressured speech, we can discriminate healthy and bipolar disorder speech. Video length is a candidate predictive feature in this sense, as subjects are asked to perform a standard set of tasks. Due to the high standard deviation of (especially) mania/hypomania video length, other speech predictors are needed for discrimination.

UAR performance for acoustic analysis shows that the audio-only system can differentiate bipolar subgroups (mania/ hypomania/remission) from healthy controls (including depression and mania simulation) with 69.4% success rate, which is statistically significantly higher compared to chance level UAR score (50% for two classes). Supporting these results with other studies have potential for effective characterization of bipolar speech from healthy control and from their uni/bipolar depression simulations. Proposed 10 functionals indicate potential for this characterization. Subdivision of audio-clips was not found to increase the success rate, and this may be related to the fact that lability (i.e. quick changes) of speech for bipolar subgroups can be better estimated in clips longer than 60 seconds.

UAR performance for visual systems were lower, or on par with acoustic systems. The appearance of the subjects may be affected from the high dose of antipsychotics used for treatment of bipolar mania, which can result in the slowing of facial mimics. The best results are attained using audio-visual fusion: in the binary task (healthy/bipolar) a UAR score of 73% and in ternary classification task (mania/ hypomania/ remission) a UAR of 55.6% is reached. We observe that multimodal fusion has higher contribution to the latter, more difficult task.

Effective application of machine learning techniques to identify and to classify mood disorders has potential in both identifying BD and in quantizing treatment response earlier [28]. In the proposed approach, short-term affective primitives predicted from acoustic features are summarized over each utterance using 10 proposed functionals. Subsequently, these are regressed against YMRS declines. Results reveal that affect primitives estimated from zeroth day recording in this way have a potential to predict treatment response in the third day of treatment. As treatment resistance is a big challenge for BD, using these predictors will make it possible to select the best treatment approach for each individual at the beginning of treatment.

When classification performance improves up to 90% UAR for bipolar/ healthy discrimination and 80% UAR for detection for treatment response, these parameters can serve as treatment response predictors; then the whole framework can be deployed as a decision support system for psychiatrists and neurologists.

VI. Conclusion

In this work, we presented a new audio-visual BD corpus, as well as experiments using both modalities for binary (healthy/bipolar) and ternary (mania level) classification. We also analyzed the bipolar corpus using cross-corpus acoustic emotion recognition, to reveal statistically significant mid-level predictors of YMRS declines from the zeroth day recording. The promising results obtained in both classification (direct approach) and regression (indirect approach) motivate future studies for automatic monitoring of people with BD. In follow up works, a compact set of predictive high level features (as in the case of predicted affect primitives), their usage in the linguistic and multimodal system development will be investigated.

References

[1] K. R. Merikangas, M Ames, L Cui, et al., “The impact of comorbidity of mental and physical conditions on role disability in the US adult household population,” Archives of General Psychiatry, vol. 64(10), pp. 1180-1188, 2007. doi:10.1001/archpsyc.64.10.1180.,

[2] World Health Organization, “The global burden of disease: 2004 update, Table A2: Burden of disease in DALYs by cause, sex and income group in WHO regions, estimates for 2004”, WHO Press, Geneva, 2008.

[3] I. E. Bauer, J.C. Soares, S. Selek, T.D. Meyer, “The link between refractoriness and neuroprogression in treatment-resistant bipolar disorder,” Mod Trends Pharmacopsychiatry, vol.31, pp 10-26, 2017. doi: 10.1159/000470803.

[4] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D.Mehta, “Vocal biomarkers of depression based on motor incoordination,” In *Proc. AVEC*, pp. 41-48, October 2013.

[5] H. Kaya, and A. A. Salah, “Eyes whisper depression: A CCA based multimodal approach,” In *Proc. ICMI*, pp. 961-964, 2014.

[6] R. W. Picard, an R. Picard, Affective Computing, vol. 252, Cambridge: MIT press, 1997.

[7] Schuller, B. “Voice and speech analysis in search of states and traits,” In Computer Analysis of Human Behavior, Springer London, pp. 227-253, 2011

[8] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” Image and Vision Computing, vol. *27*(12), pp. 1743-1759, 2009.

[9] S. Prendes-Alvarez, and N. Charles, “Personalized medicine: prediction of disease vulnerability in mood disorders,” Neuroscience Letters, 2016.

[10] The Diagnostic and Statistical Manual of Mental Disorders, 5th ed.; DSM–5; American Psychiatric Association, 2013.

[11] A. Özerdem, O. Yazıcı, Z. Tunca, K. Tırpan and Mood Disorders Study Group, “Establishment of computerized registry program for bipolar illnes in Turkey: SKİP-TÜRK,” *J. Affect. Disord*, vol. *84*, 82-86, 2004.

[12] R.C. Young, J.T. Biggs, V.E. Ziegler, and D.A. Meyer, “A rating scale for mania: reliability, validity and sensitivity,” The British Journal of Psychiatry, vol. 133(5), pp. 429-435, 1978.

[13] S. A. Montgomery, M. Asberg, “A new depression scale designed to be sensitive to change,” The Bristish Journal of Psychiatry, vol. 134, pp. 382–389, 1979.

[14] F. Eyben, F. Weninger, F. Gross, & B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” In Proc. ACM Multimedia, pp. 835-838, 2013.

[15] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” In Proc. INTERSPEECH, pp. 2794-797, 2010.

[16] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, and others, “The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring,” In Proc. INTERSPEECH 2017, pp. 3442-3446, 2017.

[17] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, et al, “AVEC 2014: 3d dimensional affect and depression recognition challenge,” In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3-10, November 2014.

[18] O. Martin, I. Kotsia, B. Macq, and I. Pitas. “The eNTERFACE’05 audio-visual emotion database,” In Proceedings of 22nd International Conference on Data Engineering Workshops, pp. 8, IEEE, 2006.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” In INTERSPEECH, vol. 5, pp. 1517-1520, 2005.

[20] V. Makarova, and V. A. Petrushin, “RUSLANA: A database of Russian emotional utterances,” In Seventh International Conference on Spoken Language Processing, 2002.

[21] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, “CHEAVD: a Chinese natural emotional audio–visual database,” Journal of Ambient Intelligence and Humanized Computing, *vol.* 8(6), pp. 913-924, 2017.

[22] H. Kaya, A. A. Salah, S. F. Gurgen, and H. Ekenel, “Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus,” In *22nd IEEE Signal Processing and Communications Applications Conference (SIU)*, pp. 1698-1701, 2014.

[23] K. Stanislavsky, An Actor Prepares. Taylor & Francis, 1989.

[24] X. Xiong, and F. De la Torre, “Supervised descent method and its applications to face alignment,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 532-539, 2013.

[25] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, “Contrasting and combining least squares based learners for emotion recognition in the wild,” In Proc. ACM ICMI, pp. 459-466, 2015

[26] H. Kaya, F. Gürpınar, and A. A. Salah, “Video-based emotion recognition in the wild using deep transfer learning and score fusion,” Image and Vision Computing, vol. 65, pp. 66-75, 2017.

[27] Q. McNemar, Psychological Statistics. New York: Wiley, 1969.

[28] D. Librenza-Garcia, B. Kotzian, J. Yang, B. Mwangi, B. Cao, L. Nunes Pereira Lima, M. Bagatin Bermudez, M. Vianna Boeira, F. Kapczinski, I. Passos, “The impact of machine learning techniques in the study of bipolar disorder: A systematic review,” Neuroscience & Biobehavioral Reviews, vol. 80, 2017. doi:10.1016/j.neubiorev.2017.07.004.