A Multimodal Approach for Mania Level Prediction in Bipolar Disorder

Pınar Baki, Heysem Kaya *Member, IEEE,* Elvan Çiftçi, Hüseyin Güleç, Albert Ali Salah, *Senior Member, IEEE*

***Abstract*—Bipolar disorder is a mental health disorder that causes mood swings that range from depression to mania. Clinical diagnosis of bipolar disorder is based on patient interviews and reports obtained from the relatives of the patients. Subsequently, the diagnosis depends on the experience of the expert, and there is co-morbidity with other mental disorders. Automated processing in the diagnosis of bipolar disorder can help providing quantitative indicators, and allow easier observations of the patients for longer periods. In this paper, we create a multimodal decision system for three level mania classification based on recordings of the patients in acoustic, linguistic, and visual modalities. The system is evaluated on the Turkish Bipolar Disorder corpus we have recently introduced to the scientific community. Comprehensive analysis of unimodal and multimodal systems, as well as fusion techniques, are performed. Using acoustic, linguistic, and visual features in a multimodal fusion system, we achieved a 64.8% unweighted average recall score, which advances the state-of-the-art performance on this dataset.**

***Index Terms*—Affective disorders, bipolar disorder, multimodal fusion, mania level prediction**

I. Introduction

ASSESSMENT of mental health disorders from behavioral data using machine learning methods is a recently growing research area, with focused work including depression [1], anxiety disorders [2] and bipolar disorder [3]. Unobtrusive affective assessment makes it possible to observe multimodal responses during structured or semi-structured observation sessions, to derive indicators and deviations from behavior, or to observe subtle changes over time [3], [4]. While, fully automated diagnosis requires the integration of a comprehensive set of indicators and detailed patient history, automatic analysis of behavior can provide clinicians with useful quantitative measurement and monitoring tools [5].

Bipolar disorder (BD) is a mental health condition that causes extreme mood swings from elevated (mania, hypomania) to diminished state (depression), as well as mixed episodes, where depression and manic symptoms occur together. Its diagnosis is performed through a set of medical examinations administered by the psychiatrist, but may require lengthy observations of the patient as there is no comprehensive test [6]. There is a lot of co-morbidity with other mental disorders including, but not limited to, any anxiety disorder, conduct disorder, and substance use disorder [6]. The disease affects 2% of the population, sub-threshold forms (recurrent hypomania episodes without major depressive episodes) affect an additional 2%, and together, the lifetime prevalence estimates are 4.4% [7]. It is ranked as one of the top ten diseases of disability-adjusted life year indicator among young adults [8], and as the 17th leading source of disability among all diseases worldwide [9].

Diagnosis of mental health disorders rely on medical examinations administered by psychiatrists and reports from patients and their relatives or friends. But there is a need for more systematic and objective diagnosis methods, for remote treatment and diagnosis approaches assisted using automated methods. It is possible to collect behavioral data from people during their everyday lives [10], which creates an opportunity to create tools to monitor the symptoms of the patients for longer periods, screen patients before they see the psychiatrists, assist clinicians in the diagnosis, and capture patient behaviors in situations where they cannot act or hide the symptoms.

Different types of bipolar disorder are characterized by changes in the patient's mood, energy, and activity levels. The patient experiences periods of intense emotion and uncharacteristic behaviors, called *mood episodes*, which can be manic (high arousal and valence) or depressive (low arousal and valence). *Manic episodes*, the focus of this paper, include elated, erratic, charged behaviors. While a loss of appetite or decreased need of sleep is difficult to judge automatically from multimedia recordings, traces of elation and irritability, fast and incoherent thought, feelings of grandeur and recklessness can be gleaned from affective language and behavioral cues. *Hypomania* is a less severe form of mania, and *remission* is the period when the behavior is returning to normal. Patients admitted to the hospital with manic episodes are medicated, closely followed, and discharged only after entering the remission stage. In this paper, we work with data collected from such patients encompassing manic, hypomanic, and remission stages.

A gold standard tool used to rate the severity of the manic episodes of a patient is the Young Mania Rating Scale (YMRS) [11] (see Section. IV-D2). During the interviews, psychiatrists observe and rate the patient's symptoms via eleven indicators. Using a structured interview, it is possible to observe some of these from speech patterns, body or facial movements, and from the content of what was spoken during the interview.

In this work, we propose a multimodal machine learning

P. Baki and A.A. Salah are with the Department of Computer Engineering, Boğaziçi Univ., Turkey.

H. Kaya and A. A. Salah are with the Department of Information and Computer Sciences, Utrecht Univ., the Netherlands.

E. Çiftçi is with NP Brain Hospital and Üsküdar Univ., Turkey.

H. Güleç is with Psychiatric and Neurological Diseases Training and Research Hospital, Health Science Univ., Turkey.

Manuscript received Month Day, Year

system that uses information from acoustic, linguistic, and visual modalities to classify the bipolar patients into remission, mania and hypomania classes. Our aim is to investigate to what extent automatic analysis approaches can provide the psychiatrists with quantitative indicators to help in their diagnosis. Despite recently increasing interest [12], [13], there are very limited publicly available resources in this area. We evaluate our proposed multimodal approach using the Turkish Audio-Visual Bipolar Disorder corpus that we have recently collected and made available to the research community [14], [3], and push the state-of-the-art performance achieved on the corpus so far. We discuss our results extensively in the light of our quantitative findings, provide insights and point out to challenges in this problem.

The rest of the paper is organized as follows. Section II discusses the previous work on bipolar disorder and related mental conditions, including Section II-B on the Turkish Audio-Visual Bipolar Disorder corpus used in our study. Section III explains the features used for each modality, the preprocessing methods, classification algorithms, and the modality fusion approach used in our study. Section IV presents the results for uni- and multimodal experiments. We discuss our findings in Section V and provide some final remarks.

II. Related Work

In this section, we first briefly summarize the main findings in the related area of multimodal depression analysis. Then we describe our dataset, before moving to a more technical exposition of specific works on BD estimation.

A. Depression analysis

Research on depression analysis has shown that multimodal fusion of features in various levels increase the performance of single modalities [4], [15]. Fusion of textual, acoustic and visual features extracted from the clinical patient interviews outperforms unimodal models [15], [16], [17]. Recently, using a feature selection framework, F0, HNR, formants, and MFCC for the speech, and left-right eye movement, gaze direction and yaw head movement for visual modality are shown to be the most distinctive features for depression analysis [1]. These are in line with the former research showing that stillness of eyes [18] and low acoustic variability are important indicators of depression [4]. Additionally, lexical content of what people say during interviews is also useful in the detection of depression [19], [20]. Using only audio and textual information in a multimodal system is useful when there is no visual data, such as during phone call conversations [21], [15].

While low energy and acoustic / visual variability are indicators of a higher level of depression, BD patients show an inverse pattern during mania episodes. Higher bodily and acoustic energy, higher variability and lack of focus in the spoken content are correlated with mania levels [6].

B. The Turkish Audio-Visual Bipolar Disorder (BD) Corpus

In this paper, we use the Turkish Audio-Visual Bipolar Disorder Corpus [14] to report experimental results. Before discussing the related work performed on this corpus, we provide some details about the data. In our experiments, we have adhered to the 2018 AVEC Bipolar Disorder and Cross-cultural Affect Recognition Competition [3] protocol to ensure comparability of results with the literature. The aim of the AVEC competition series is developing and comparing machine learning models using audio and visual components on various affective computing problems. Participants were encouraged to achieve the highest performance, considering the baseline performance provided by the organizers. The BD corpus was used in the 2018 AVEC challenge for the first time, and only a part of it was opened for the challenge.

The original BD corpus contains video clips of 46 bipolar disorder patients and 49 healthy controls collected at the Istanbul Health Sciences University, Erenkoy Mental Health Research and Training Hospital[1]. Mania level of the patients is evaluated on 0th, 3rd, 7th and 28th days of the hospitalization and after discharge on the 3rd month. On those days, psychiatrists performed an interview with the patients, asking the same questions each time, and taking audiovisual recordings of the sessions. Annotation was done based on the Young Mania Rating Scale (YMRS) score [11], which is a continuous clinical interview assessment scale used for rating the severity of manic episodes of a patient. Scores range from 0 to 60, where higher scores represent severe mania. In the BD corpus, bipolar patients are grouped into three ordinal classes (remission, hypomania, and mania, respectively) based on their session-wise YMRS score, as described in [14].

During recordings, patients were asked to perform seven tasks, designed to arouse different emotions in the patients. The first three tasks can be considered as negative emotion eliciting tasks, the subsequent two tasks are neutral, and the two final tasks are positive emotion eliciting tasks. The performed tasks are 1) explaining the reason for coming to the hospital, 2) describing van Gogh's Depression painting, 3) describing a sad memory, 4) counting from one to thirty, 5) counting from one to thirty faster, 6) describing Dengel's Home Sweet Home painting and 7) describing a happy memory. The paintings used in the study are shown in Figure 1.

Fig. 1. Van Gogh's Depression (left), Dengel's Home Sweet Home (right)

Clips were recorded in a room where only the participant and the psychiatrist were present. The participants were recorded with a camera while performing tasks. They read the descriptions of the tasks they were asked to perform from

[1]Please see [14] for patient sociodemographics, clinical characteristics, and exclusion criteria.

the computer screen. After completing a task, the participants pushed a button and a description of the next task appeared on the screen, while a 'knock' sound was played to mark the beginning of a new task. This sound helps to split tasks if one wants to use the tasks separately for classification. Our preliminary experiments have shown that task-based analysis results in too small data partitions for training, and does not result in higher overall accuracy [22], since some negative-emotion eliciting tasks are skipped by a number of patients.

In the AVEC 2018 Challenge, only data from the bipolar patients are used for a three-class (R: remission, H: hypomania, M: mania) classification. The healthy controls have visual properties that may help in their identification (e.g. clothing colors for doctors), and subsequently, they are not used in the AVEC Challenge or in this paper. In the competition, there were 104 (R: 25, H: 38, M:41), 60 (R: 18, H: 21, M:21), and 54 (18 each) clips in the training, development, and test sets, respectively. As it is the case with other mental-healthcare datasets, the number of session-wise annotated samples is small, which may lead to overfitting, and here is a mild data imbalance that can cause bias in favor of the majority class.

C. Multimodal supervised learning for mania estimation

The first comprehensive set of investigations into the extension of multimodal methods to the analysis of BD started with the 2018 Audio/Visual Emotion Challenge (AVEC) [3], which introduced the Turkish BD corpus described in the previous section to the larger affective computing community in form of a challenge. Several groups have worked on this corpus within the AVEC Challenge [23], [24], [25], [26], [27], [28], [31]. Table I summarizes the major works reporting results on the BD corpus to date. In this section, we summarize the feature extraction and machine learning approaches that were used for the mania level estimation problem. We caution the reader that the reported accuracies in these works (including the present paper) are not clinical results, but a good indication of the possibilities of automatic analysis approaches.

As the classification of manic episodes is correlated with increased arousal levels, audio-visual detection of arousal is a good place to start. In [23] arousal-related features extracted from speech and from visual upper body motion of patients were fused. Another important source of information is the dynamics of affective cues. Syed et al. [25] proposed to use turbulence features that represent the sudden changes in feature contours of both audio and visual modalities. In the extraction of audio features, they used a Fisher vector encoding with a feature set extracted via the openSMILE tool [32]. They have used a standard feature set introduced for the Interspeech Computational Paralinguistics Challenge (ComParE). Other groups (e.g. [24]) have used the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for acoustic feature extraction [33].

In [25], the classification is performed using the Greedy Ensemble of Weighted ELMs model [34]. Because of the small number of samples, deep learning is not suitable for end-to-end classification, but transfer learning can be adopted for feature extraction. Using highly complex classifiers results in poor generalization due to the limits of the training set. In [24], Xing et al. used linguistic features in addition to visual and audio based features, and created 5,395 dimensional features by the early fusion of these three modalities. Using eGeMAPS features, Mel frequency cepstrum coefficients (MFCC), facial action units, and gaze features, they achieved the highest Unweighted Average Recall (UAR) on the validation set among the AVEC Challenge participants. However, the great difference between the UAR scores on the development and test sets (i.e. 86.7% vs 57.4%) shows that the proposed model cannot generalize well to the sequestered test set data. In [31], an Inception module was combined with an LSTM network, and L_1 regularization to deal with overlearning. 16-dimensional MFCC features are extracted from the speech files. Using only audio features, 65.1% UAR is achieved on the validation set. However, no score was reported for the test set. In [26], LSTM and Bi-LSTM models were trained on the challenge baseline features including MFCCs, eGeMAPS, Bag-of-Acoustic-Words [35], DeepSpectrum[2], Facial Action Units (FAU) and Bag-of-Visual-Words (BoVW). Their best result on the test set was achieved with the Bi-LSTM network trained on the concatenation of all the features.

AbaeiKoupei and Al Osman reach the test set baseline by only using visual features extracted from a pre-trained and fine-tuned deep neural network model [29], achieving 60.6% and 57.4% UAR on development and test sets, respectively, which shows that the model does not overfit. Capsule Neural Networks (CapsNet) [36] were used in [27], on Mel-frequency spectrograms extracted from small segments of raw audio files. In [28], audio clips were segmented into chunks to increase the dataset size. However, each clip has only one label and after segmenting the clip, each chunk becomes weakly labeled. This problem was solved using multi-instance learning, where training was performed with a bag of instances, instead of one single feature vector. Using ensembles of DNNs, 61.6% UAR on the development, and 57.4% UAR on the test set was achieved using the audio modality.

The works mentioned so far used audio and video features, but the text transcriptions of the speech of the patients during the tasks are also informative, particularly in a multimodal context. For instance, when a patient describes a sad memory, e.g. the death of a loved one, in a cheerful voice, this presents a strong case for elevated mania levels. Zhang et al. proposed fixed length, session-level paragraph-vector representations for the text modality [37]. They showed that early fusion on audio-visual and textual representation vectors was beneficial.

The highest test set score achieved on the BD corpus so far was 59.3% UAR [30], using eGeMAPS and MFCC acoustic features, as well as linguistic features, such as the number of words, number of types, letters per word, number of paragraphs, number of sentences, and number of words per sentence. Additionally, sentiment information was extracted using the SEANCE tool [38]. In the next section, we present a tri-modal system that advances the state of the art in this problem.

[2]https://github.com/DeepSpectrum/DeepSpectrum

TABLE I
SUMMARY OF THE WORKS THAT USE BD DATASET. (A: AUDIO, V:VISUAL, T:TEXTUAL). PLEASE SEE THE TEXT FOR THE USED ABBREVIATIONS.

Paper	Modalities	Features	Classifier
Ringeval *et al.* [3] (baseline)	A,V	eGeMAPS+FAUs	SVMs
Yang *et al.* [23]	A,V	Arousal and upper body posture features	Multistream
Xing *et al.* [24]	A,V,T	eGeMAPS+MFCC+Timing+FAUs+Emotion+Eyesight+Body movement +features from various NLP tools including SiNLP	Hierarchical recall model
Syed *et al.* [25]	A,V	FAUs+gaze+pose	GEWELMs
Ebrahim *et al.* [26]	A,V	MFCC+eGeMAPS+BoAW+DeepSpectrum+FAUs+BoVW	Bi-LSTM
Amiriparian *et al.* [27]	A	Mel-Spectogram	CapsNet
Ren *et al.* [28]	A	MFCC	Multi-instance learning
AbaeiKoupaei, Al Osman [29]	V	Facial Features	LSTM
AbaeiKoupaei, Al Osman [30]	A,T	MFCC+eGeMAPS+SiNLP+SEANCE	Stacked Ensemble Model

III. METHODOLOGY

Figure 2 illustrates the proposed multimodal framework for the classification of the patients into one of remission, hypomania, and mania classes, based on a short video interview where the patient performs several tasks, as described in Section II-B. The components of the pipeline are summarized in the following subsections. Besides leveraging multimodality, our goal was to reach a minimal set of modalities and features that provide the highest predictive performance. Thus, we conducted comparative unimodal experiments to choose the optimal sets of features in each modality.

A. Feature Extraction

1) Acoustic Feature Extraction: Acoustic cues are used for the diagnosis of bipolar disorder by psychiatrists. Rapid, pressured speech and speaking too much (amount of speech) are common indicators of manic episodes. Subsequently, we perform acoustic analysis of the patient interviews.

For acoustic feature extraction, we use the openSMILE feature extraction toolkit [32], which provides many built-in configuration files that extract the baseline audio features from INTERSPEECH and AVEC challenges, and some parameter sets proposed for voice research and affective computing studies on audio. In our experiments, we use eGeMAPS [33] (Extended Geneva Minimalistic Acoustic Parameters Set), which is a parsimonious set of audio features, chosen for their ability to represent affective physiological changes in voice production. 23 eGeMAPS low level descriptors (LLD) are summarized using the functionals from the original eGeMAPS configuration [33], and this set is enriched by 10 functionals we have added [14] (see Table II for the entire set). Additionally, the baseline acoustic feature set from INTERSPEECH 2010 Paralinguistic Challenge (IS10) [39] is used, with 38 low-level descriptors and their temporal derivatives, indicating paralinguistic activity.

2) Linguistic Feature Extraction: Clinicians assess the presence of risk of suicide, risk of violence to persons or property, risk-taking behavior, sexually inappropriate behavior, substance abuse, patient's ability to care for himself/herself, etc., using the patient interview contents [40].

TABLE II
LIST OF STATISTICAL FUNCTIONALS APPLIED TO LLDS.

Functional	Description
Mean	Arithmetic mean
Std	Standard deviation
Curvature	Leading coefficient of the 2^{nd} order polynomial fit to LLD contour
Slope + offset	Coeff. of the 1^{st} order polynomial fit to LLD contour
Min	Minimum value
Relative Min Location	Location index of min value divided by the length of LLD contour
Max	Maximum value
Relative Max Location	Location index of max value divided by the length of LLD contour
ZCR	Zero crossing rate of LLD contour normalized to [-1,1]

We use the Google Automatic Speech Recognition (ASR) tool [3] to convert the interviews to text, and obtain one text segment per task for each interview.

Transformer language embeddings (GPT-2 [41], BERT [42], GPT-3 [43]) are the state-of-the-art natural language processing (NLP) models in representing language features. However, these complex models show unreliable results on small datasets. Consequently, we use three alternative feature sets for the linguistic experiments, which are linguistic inquiry and word count (LIWC) [44], term frequency-inverse document frequency (TF-IDF), and polarity features.

LIWC is a text analysis tool that calculates the linguistic or psychological categories of words where the categories indicate social, cognitive, and affective processes. It extracts 93 features for an input text file. Before using LIWC, the Turkish transcripts extracted from the patient clips are translated into English via Google Translation engine[4].

TF-IDF is a statistical measure that shows how much a word is important in a document. They are used commonly in NLP, information retrieval, and text mining tasks. As a preprocessing step, stop words are removed using the NLTK library [45], and stemming is applied using the Porter algorithm [46]. After these steps, TF-IDF features are computed over the set of uni-grams and bi-grams.

[3]https://cloud.google.com/speech-to-text
[4]https://cloud.google.com/translate

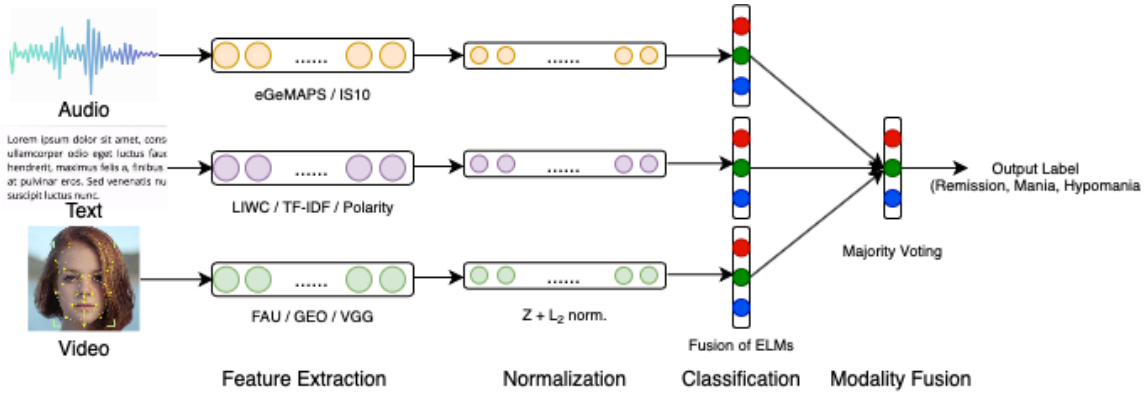


Fig. 2. Pipeline of the multimodal system. For each unimodal system, different feature sets are used. The best performing feature sets are combined.

As **polarity** features, we use the outputs of three sentiment analysis tools together, which are Natural Language Toolkit Valence Aware Dictionary for sEntiment Reasoning (NLTK Vader) [47], TextBlob [48] and Flair [49] due to their complementary information and generalization performance on a recent mood recognition challenge [50]. NLTK Vader uses a sentiment lexicon together with grammatical rules for expressing polarity, but performs weakly on unseen words. Flair uses a character-level LSTM network for sentiment analysis, with good generalization to unseen words. The TextBlob library returns a sentiment with polarity and subjectivity scores, where subjectivity represents the amount of personal and factual information in the sentence, which is a good feature for the valence dimension. However, it does not consider negation in sentences for the polarity score, which can be misleading.

Sentiment and subjectivity features obtained from these three libraries are combined into a feature vector. Then each feature is summarized with mean, standard deviation, max, min, and sum functions.

3) Visual Feature Extraction: Clinicians gain significant insight from visual cues and some items of YMRS can be obtained from visual cues like increased motor activity-energy, irritability, elevated mood, appearance, and disruptive-aggressive behavior. Besides, the speech rate and the amount can also be observed in the facial actions.

For the visual experiments, we use facial action units (FAUs), as well as geometric features extracted from each face, provided as baseline features in the AVEC challenge [3]. The FAUs are based on the Facial Action Coding System (FACS), which describes the movements of specific facial muscles [51]. Emotional expressions typically correspond to combinations of various action units. In [3], intensities of 16 FAUs along with a confidence score are extracted using the OpenFace toolkit [52].

The set of 23 geometric features we use are based on our early work for video-based emotion recognition in uncontrolled conditions. They are extracted from detected and aligned faces of the BD corpus and represent different geometric aspects like distance, angle, and aspect ratio based on facial landmarks (see [53] for a full list).

B. Preprocessing

The feature vectors extracted for each clip contain representations of auditory, visual, and textual signals with different ranges and scales. Subsequently, feature standardization or normalization needs to be performed before model training.

The features we used for the classification of the clips are represented as two-dimensional matrices, where columns are the functionals of the low level descriptors and each row contains the feature vector of a clip. We experiment with both row-level and column-level normalization. For the column-level, standardization (z-normalization) brings each feature to the same scale. For the row-level, we apply L_2 normalization, which effectively transforms a linear kernel into a cosine similarity kernel.

C. Feature Selection

We use high-dimensional feature sets in our experiments. Considering the sample size of the BD dataset, we considered reducing the feature dimensionality with feature selection. We tested the *tree feature selection* method [54], which is expected to be robust against overfitting. In this approach, a random forest [55] is trained, and features are ranked based on the information gain for each feature. We report experimental results with this approach, even though ultimately, it did not yield improved test set results.

D. Classification using Extreme Learning Machines

Due to the difficulty of collecting data from bipolar patients, the BD dataset has only 164 data points in total, and it is crucial to pay attention to getting accurate predictions while avoiding overfitting. The performance measure used here is Unweighted Average Recall (UAR), which is the average of recall performances across the classes (the same measure as in the AVEC 2018 challenge [3]). This measure has a chance level performance of $1/K$ for K-class classification.

In our experiments, we employ a Kernel Extreme Learning Machine (Kernel ELM) classifer [56]. The initially introduced Basic ELM is a simple and robust machine learning method that contains a single hidden layer. Input weights are randomly generated and they are not tuned. The weights between the hidden layer and the output layer are analytically calculated by

a pseudo inverse operation. While multi-layered deep learning approaches are used in many problems, on such a small dataset they will easily overfit, and simpler models should be preferred.

In a single hidden layer ELM, the hidden layer output matrix is $\mathbf{H} \in \mathbb{R}^{N \times h}$, the weight matrix between the hidden layer and the output layer is $\beta \in \mathbb{R}^{h \times 1}$ and the output layer matrix is $\mathbf{T} \in \mathbb{R}^{N \times 1}$, where N is the number of training samples and h is the number of hidden layer nodes. The output weight matrix β is calculated using least squares solution of $\mathbf{H}\beta = \mathbf{T}$ as $\beta = \mathbf{H}^{\dagger}\mathbf{T}$. $\mathbf{H}^{\dagger}$ represents the Moore-Penrose generalized inverse [57], which minimizes L_2 norms of both $\|\mathbf{H}\beta - \mathbf{T}\|$ and $\|\beta\|$. For increased generalization and robustness, the kernel trick and a regularization coefficient C is used. The set of weights is calculated as:

$$\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{K}\right)^{-1} \mathbf{T}, \tag{1}$$

where $\mathbf{I}$ is an identity matrix, and $\mathbf{K}$ is a kernel (i.e. similarity matrix) obtained from the training dataset. We use a radial basis function (RBF) kernel $\mathbf{K}$, as suggested in [58].

In weighted ELM [34], which is a variant that is used with class imbalanced problems, we define a $N \times N$ diagonal weight matrix $\mathbf{W}$. Each diagonal element stores the multiplicative inverse of the number of training samples N_i of the corresponding class i. Integrating $\mathbf{W}$ into the formula, β is calculated as:

$$\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{W}\mathbf{K}\right)^{-1} \mathbf{W}\mathbf{T}. \tag{2}$$

There is a trade-off between weighted and unweighted models, where the former favors minority classes (better UAR), while the latter favors majority classes (better accuracy). To reach the best performance, we investigate a weighted decision level fusion approach:

$$\mathbf{P}_{fusion} = \alpha \mathbf{P}_{unweighted} + (1 - \alpha)\mathbf{P}_{weighted}, \tag{3}$$

where $\mathbf{P}$ is an $N \times t$ matrix that contains the class probabilities of each sample. α is a coefficient in [0, 1] range. The best α is chosen according to the UAR score of $\mathbf{P}_{fusion}$ on the development set.

E. Modality Fusion

The expert assesses patient's speech patterns (e.g. rate, amount of speech), visual appearance, gestures, motor activity, as well as expressed sentiments and ideas, and the content of speech during the interviews. All of these indicators are informative, and instrumental in deciding the patient's YMRS score and to diagnose BD episodes. We investigate how an automatic system can best extract indicators from each modality, and how they complement each other by providing context or including more discriminative information.

We firstly experiment with audio, speech, and text modalities separately. Our experiments show that the audio modality gives a better score overall, while the hypomania class is not classified well. The linguistic modality generally gives lower UAR than audio modality, but all three levels of mania are classified with similar performance. For the multimodal systems, we use the best performing unimodal models and features.

First, we consider two late fusion methods, namely, majority voting and weighted sum. Majority voting outputs the mostly seen label (mode) for a sample, and uses the audio modality for tie breaking. Weighted sum combines two or three modalities. When fusing two modalities, the probability vectors from each model are given as input and the final probabilities are obtained as:

$$\mathbf{P}_{fusion} = \alpha \mathbf{P}_{model_1} + (1 - \alpha)\mathbf{P}_{model_2}, \tag{4}$$

where $\mathbf{P}$ is an $N \times t$ matrix that contains the class probabilities of each sample, N is the number of samples and t is the number of classes. α is a coefficient in the [0, 1] range, optimized according to the UAR score obtained from P_{fusion}. For the fusion of three modalities, we apply a variant of Equation 4.

$$\mathbf{P}_{fusion} = \sum_{k=1}^{K} \alpha_k \mathbf{P}_{model_k} \tag{5}$$

In Equation 5, the alpha values are the elements of the vector drawn from a Dirichlet distribution. We sample 500 times randomly and find the values that maximize the UAR of the final fusion model. A probability density function of a Dirichlet distribution of order $N \geqslant 2$ with parameters $\alpha_1, ..., \alpha_n > 0$ is

$$\frac{1}{B(\alpha)} \prod_{i=1}^{N} x_i^{(\alpha_i - 1)}, \tag{6}$$

where $B(\alpha)$ is a normalizing factor given in terms of multivariate beta function, and $x_i \in (0, 1)$ and $\sum_{i=1}^{N} x_i = 1$.

We also experiment with early (feature level) fusion methods. In our approach, the features from different modalities are summarized via LLDs, concatenated and normalized into a single feature vector before the classification.

We evaluated only a small number of models on the test set to prevent overfitting. While selecting the fusion models to test, we considered the Multimodal 1 (MM1) metric [59], which measures the improvement in the final fusion model:

$$MM1 = \frac{UAR_{fusion} - max(UAR_1, UAR_2, UAR_3)}{max(UAR_1, UAR_2, UAR_3)}, \tag{7}$$

where UAR_{fusion} is the UAR score of the fusion model, UAR_i are the UAR scores of the unimodal models. While calculating the MM1 score, we use 4-fold cross-validation scores, since it gives more robust results. After getting the test set results for the selected fusion models, we calculate MM1 scores using test set UARs.

F. Shapley Additive Explanations

To further investigate the contribution of features for classification, we use the SHAP method [60], which aims to provide an explanation to a particular prediction by means of additive feature attributions based on cooperative game theory [61]. Given a model f and input feature vector x with D features, SHAP assigns each feature i an importance weight $\phi_i(f, x)$

measuring its marginal contribution to the model output, where marginal contribution is calculated as a (weighted) average over a range of 'observed' feature subsets:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|(D-|z'|-1)!}{D!}(f_x(z') - f_x(z' \setminus i)), \quad (8)$$

where $x' \in \{0,1\}^D$ is the simplified feature representation, $|z'|$ is the number of non-zero entries in z' and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x' [60]. Observing means that a subset of features are seen by the model with their original values, while others are missing (represented with a 0 in z'). Since not all models are inherently capable of handling missing values, the authors propose an approximation to missing values, assuming model linearity and feature independence [60].

IV. Experiments and Results

We report both unimodal and multimodal experimental results in this section. In all tables, we denote tree-based feature selection with a star.

A. Unimodal Experiments

1) Audio Classification: For the clip level audio classification, we used eGeMAPS and IS10 feature sets, which are extracted using the openSMILE feature extraction toolkit [32]. The LLD features are summarized using 10 functionals proposed in [14]. We also used original eGeMAPS feature set [33]. Z-normalization is applied to each feature separately. After that, L_2 normalization is applied to the feature vector of each clip. The decision level fusion of weighted and unweighted RBF Kernel ELMs is used for the classification. In this setup, the best result is achieved on eGeMAPS10 features with 63.7% UAR on the development set (see Table III).

TABLE III
UAR scores for single modality experiments.

Modality	Features	Dimension	Dev.	4F-CV
Audio	IS10	760	55.2%	**56.8%**
	eGeMAPS10	230	**63.7%**	53.1%
	eGeMAPS10*	98	60.8%	52.2%
	eGeMAPS	88	52.9%	53.8%
Text	LIWC	93	**53.7%**	**57.3%**
	TF-IDF 500 Bigram	500	49.4%	57.3%
	Polarity	35	48.9%	42.5%
Visual	GEO - Mean	23	**57.1%**	59.2%
	GEO - Mean, Std.	46	55.8%	**60.7%**
	FAU - Mean, Std.	32	55.8%	56.0%
	VGG - Mean, Std.	49	41.2%	52.2%

2) Text Classification: Text level BD classification is performed with the same configuration used in the audio experiments. Since better feature extraction tools are available for the English language compared Turkish [50], we use Google Translation engine to process the text in English. We experiment with LIWC, TF-IDF and polarity features (see Section III-A2) for the text classification.

Table III shows the results on the text features obtained from the entire clip, with late fusion of weighted and unweighted Kernel ELMs. LIWC features give the best results for both development set and the cross-validation experiment, with 53.7% and 57.7% UAR, respectively.

3) Video Classification: For the visual experiments, FAUs, geometric features, and appearance descriptors, obtained from the pre-classification layer of a convolutional neural network (VGG), are used (see Section III-A3). Table III shows the best results achieved on the visual modality. All feature sets are normalized using Z- and L_2 normalization, as explained in Section III-B. The 4,096-dimensional VGG features are extracted from the pre-trained network, then summarised with mean and standard deviation functionals, which creates an 8,192-dimensional feature vector for each clip. We then reduce the dimensionality using PCA (retaining 99% variance) to 49 dimensions, and apply tree-based feature selection. The ExtraTreeClassifier method from the scikit-learn library [62] is used for feature selection. For the VGG feature set, using only PCA gives the best result. The fourth row in the Visual part of Table III shows the results obtained with a 49-dimensional feature vector after applying PCA to the VGG feature vector. On the development set, the best result (57.1% UAR) is achieved using geometric features summarized using the mean functional. Using 4-fold cross-validation (4F-CV), 60.7% UAR is achieved on geometric features summarized with mean and standard deviation.

B. Fusion of Modalities

After investigating unimodal performances, we perform multimodal fusion experiments using weighted sum, majority voting, and feature fusion methods as explained in Section III-E. Mainly, we select the feature sets that performed well in the unimodal experiments and use them in multimodal fusion. For the acoustic modality, we select eGeMAPS10 and eGeMAPS feature sets, since the eGeMAPS feature set is created specifically for the affective paralinguistic tasks, and may provide better intelligibility compared to the IS10 feature set. For the linguistic modality, LIWC features and for the visual modality, FAU and geometric features are used for the fusion experiments.

Many previous works on this dataset used a validation set to optimize their models, but such optimization did not correlate well with the results obtained on the test set, and often, these models did not perform better than the baseline test set performance. We use 4F-CV and MM1 scores to select the fusion systems that will be used for the limited test set probes.

TABLE IV
The best majority voting fusion results.

Acoustic	Visual	Linguistic	UAR 4F-CV	MM1 4F-CV	UAR Test	MM1 Test
eGeMAPS10*	FAU	LIWC	**65.8%**	0.15	61.1%	0.10
eGeMAPS	FAU	LIWC	65.1%	0.14	57.4%	0.0
eGeMAPS10	FAU	LIWC	64.9%	0.13	**64.8%**	0.09

Ranking the fusion systems with respect to 4F-CV UAR performance, the top systems are observed to use majority voting, which shows the effectiveness of this approach. Feature fusion method is not as successful as majority voting in improving the unimodal performances, since after concatenating the feature sets, the newly generated feature vector has a higher dimension, which requires more data for a robust training [63]. Regarding the visual modality, the FAU features are observed to contribute more in the multimodal setup, but the geometric features perform better in the unimodal system.

The best 4F-CV UAR score (65.8%) is achieved using eGeMAPS10 with tree feature selection, LIWC, and FAU features fused with the majority voting method. The MM1 score shows that fusion of the modalities increases the maximum unimodal performance by 15%, which is the highest MM1 score achieved on the 4F-CV results as well.

The final test set experiments are done using the top performing three multimodal fusion systems for a fair comparison with the literature. We also limit the test set probes to 10, including the unimodal constituents. We also obtain the test set results of the constituent unimodal models (i.e. eGeMAPS10, eGeMAPS10 with tree feature selection, eGeMAPS, LIWC, and FAU) in order to report their MM1 scores on the test set.

Table IV shows the test set results for the top three multimodal systems. The best test set result is achieved using the eGeMAPS10, LIWC and FAU feature sets with the majority voting method. With this setup, we achieve 64.8% UAR score, which is 5.5% (absolute) higher than the best result published on this challenging corpus (see Table VI).

We further analyzed the contributions of each modality/feature set on the fusion performance of the top three systems in the 4F-CV setting with a randomization test. For this, we randomly generated predictions for each modality and combined them with the true predictions of from the other modalities via majority voting. This was repeated a hundred times and an average UAR was calculated. The drops in UAR performance with respect to the original systems are shown in Figure 3. We observe that within each system, the contributions of the modalities are similar, however the ranking changes per system. In two systems (Systems 1 and 3), the contribution of the visual (FAU) model is the highest and the linguistic modality ranks the second. In System 2, the linguistic modality has the highest contribution.

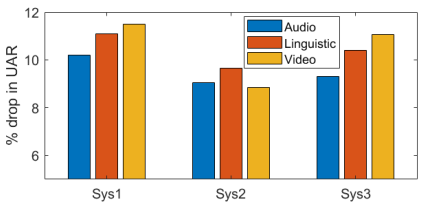


Fig. 3. Impact of each modality on the fusion UAR in the 4F-CV setting in terms of performance drop due to randomization of the corresponding predictions.

C. Choice of classifier

In earlier sections, we have compared the unimodal performances of the acoustic, linguistic and visual features used in the proposed pipeline. In order to motivate the choice of Kernel ELM in the proposed system, we compare it with Ordinal Multi-Class SVM (OMSVM) [64] classifier with the same set of features and optimization procedure as in the Kernel ELM. The choice of OMSVM is motivated by the ordinal nature of the mania levels under investigation and recent trends to use ordinal machine learning methods, as opposed to classification and regression in ordinal affective computing [65], [66]. We report the results with OMSVM at the end of Table VI. Unlike Kernel ELM, OMSVM does not model FAU and LIWC features well for mania level prediction, and the development and test set UAR performances remain below 40%. Subsequently, the majority voting of the three modalities fall below the performance of eGEMAPS features alone, which gives consistent performance across the development and test sets, but with dramatically lower performance compared to the proposed system.

TABLE V
UNIMODAL UAR (%) PERFORMANCES OF THE FEATURES FOUND IN THE BEST PERFORMING FUSION SYSTEMS.

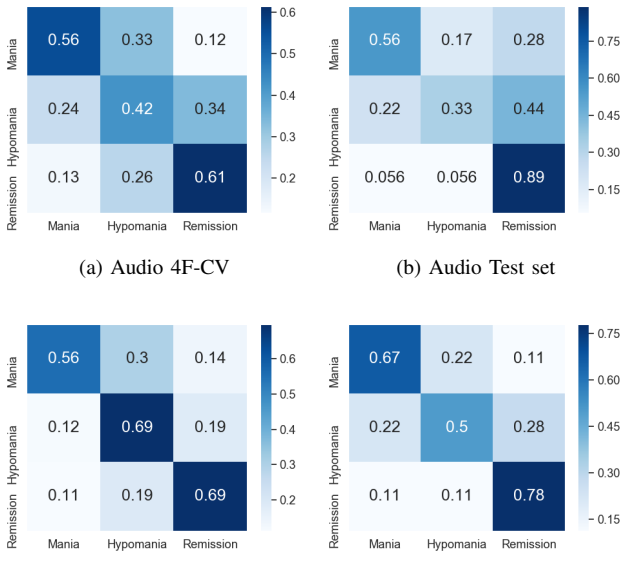
Modality	Feature Set	4F-CV	Test
Acoustic	eGeMAPS10*	52.2	55.5
Acoustic	eGeMAPS	53.8	57.4
Acoustic	eGeMAPS10	53.1	59.2
Linguistic	LIWC	57.3	51.8
Visual	FAU	56.0	51.8

Analyzing the confusion matrices of the best unimodal and multimodal systems (see Figure 4), we observe that the multimodal system improves the recall performance of the hypomania class on both 4F-CV setting and the test set. Furthermore, the recall of the mania class is dramatically improved with multimodality on the test set, while the recall of the remission class is markedly higher (0.89) with the acoustics-based unimodal model, compared to the trimodal fusion system (0.78).

Table V shows the test set results of the constituent unimodal systems obtained using the feature sets that perform the best on the multimodal fusion experiments. The test set results are obtained using the model trained on the combination of the training and the development sets, with the parameters optimized on the 4F-CV experiments. eGeMAPS10 gives the highest unimodal UAR score (59.2%), which is on par with the state-of-the-art multimodal test set performance reported on this dataset (see Table VI).

D. Interpretability Analysis

In order to provide insights into the decision making process of the multimodal system, we carried out two sets of experiments. The first set of experiments are conducted using a state-of-the-art explainable machine learning method, Shapley Additive Explanations [60], to obtain feature importance attributions for each modality-specific model that are fused via majority voting. The second set of experiments are conducted to predict *activity* in each of the eleven items of



(a) Audio 4F-CV (b) Audio Test set

(c) MM Fusion 4F-CV (d) MM Fusion Test set

Fig. 4. Confusion matrices of the best unimodal system using eGEMAPS10 features (a-b) and the top multimodal fusion system (c-d).

TABLE VI
UAR (%) COMPARISON OF WORKS USING BD WITH AVEC 2018 CHALLENGE PROTOCOL. MV: MAJORITY VOTING OF UNIMODAL MODELS.

Paper	Validation	Test
Ringeval *et al.* [3] (AVEC challenge baseline)	55.0	57.4
Yang *et al.* [23]	78.3	40.7
Xing *et al.* [24]	86.7	57.4
Syed, Sidorov, Marshall [25]	55.0	48.2
Ebrahim, Al-Ayyoub, Alsmirat [26]	59.2	44.4
Amiriparian *et al.* [27]	46.2	45.5
Ren *et al.* [28]	61.6	57.4
AbaeiKoupaei, Al Osman [29]	60.6	57.4
AbaeiKoupaei, Al Osman [30]	64.0	59.3
eGEMAPS with OMSVM	54.8	53.7
MV(eGEMAPS, FAU, LIWC) with OMSVM	45.5	51.9
Proposed system with KELM	**64.0**	**64.8**

Young Mania Rating Scale (YMRS), which served as ground truth annotations in the present study.

1) SHAP based interpretability analysis: Since the top multimodal systems are based on decision fusion rather than feature fusion, we analyse each modality-specific model separately using Kernel SHAP [60], a model-agnostic explanation method. The ELM models do not provide a straightforward way to read the relative feature importances. SHAP gives an 'explanation' for each decision, and this amounts to the impact of each feature on the output. For D features and $t > 2$ classes, SHAP will output a $D \times t$ matrix of feature attributions for each prediction. A common approach for explaining an instance is to visualize the vector for the class having the highest posterior probability. To provide insight into our trained models, we averaged the absolute SHAP feature attributions corresponding to the predicted class over the test set. The resulting feature importances are sorted and the top 20 features for each feature set are shown in Figure 5.

Analyzing the plots for eGEMAPS related acoustic features, we observe a large difference in the top features, due to the additional functionals used in the proposed eGEMAPS10 set over the set of LLDs (see Table II). However, in both plots we observe the dominance of formant related supra-segmental features among the most influential features. In the affective computing literature, formats (resonant frequencies of the vocal tract filter), specifically the first three, are known to carry affect related information [67], [68]. Thus, this outcome is very intuitive. The F_2 descriptor appears at the top of both rankings, albeit with different functionals; the curvature and slope for eGEMAPS10, and the mean for eGEMAPS.

Another observation in the acoustic models is about prosody and voice quality related features. High level of mania is associated with high vocal and kinesthetic energy, thus it is not surprising to see functionals of loudness among the most influential features. In both plots, we observe Harmonics-to-Noise Ratio (HNR), although summarized with different functionals. Jitter and shimmer, two other voice quality features, are also commonly associated with mood disorders, mainly with depression [4], and we observe them among the influential features.

When we look at the FAU features in Fig. 5, we see that AU12 (lip corner puller) is the most prominent facial feature, which is an important feature for capturing mirth, as it is activated in smiling and laughter. The second important feature is AU17 (the chin raiser), which could indicate rapid speech and excitation.

For the LIWC features, we have an interesting pattern, with the "Religion" category showing a large feature importance. When we investigate the transcripts that show high value for this category, we notice several manic patients producing an incoherent discourse intermingled with heavy religious terminology. In the literature, religious fervor is listed as a possible indicator of the manic stage [69], and the 8th item of the YMRS measures hyper-religiosity explicitly. Our patients indeed were exhibiting this symptom. Significantly, our experimental setting does not naturally lead to the recounting of a religious experience; the two paintings shown to the patients do not have religious overtones, and the happy/sad memories also do not suggest anything in this direction. Our conclusion is that LIWC-based automatic analysis is powerful enough to spot these instances, and such a cue is not detectable via face analysis, or voice paralinguistics. We note that such symptoms are not seen in each patient, and must be treated cautiously. By themselves, religious fervor or discourse obviously do not immediately suggest mania; it is by the combination of several symptoms that a clinical diagnosis can be made.

2) YMRS Item Activity Analysis: YRMS scores are composed of eleven items that are summed up and thresholded for mania level classification in the present study. These items assess the elevated mood, increased motor activity-energy, sexual interest, sleep, irritability, speech rate and amount, language-thought disorder, content, disruptive-aggressive behavior, appearance, and insight.

A detailed description of the levels of each item can be

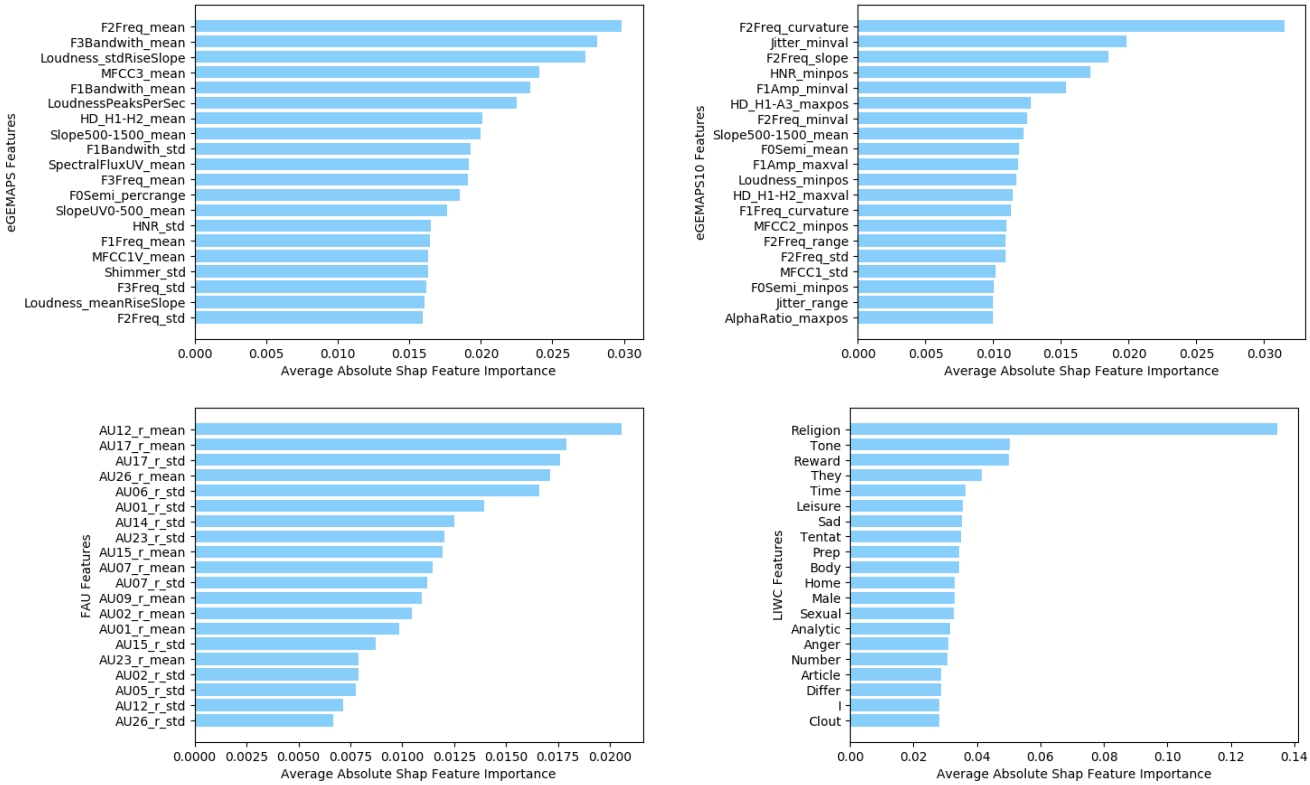


Fig. 5. SHAP based feature importance plots for unimodal constituent models of the proposed mania level recognition system.

found online[5]. While some of these items are can be predicted from acoustic, linguistic and visual features/correlates, some involve evaluation of the clothing choice (YMRS10 - Appearance) or medical expert's insight (YMRS11 - Insight), which we cannot model directly using the modalities and features in question. Therefore, we set up the following experiment in order to have an insight on the extent a modality / feature set used in the proposed pipeline can predict activity in these items. For the sake of uniformity of treatment and ease of interpretation, we binarized each YMRS item score y_i at a threshold of 0 to indicate the existence of activity a_i in the respective item: $a_i = I(y_i > 0)$, where $I()$ is the indicator function. We then trained models to predict these activities using the constituent feature sets in our system.

The results of the YMRS item activity recognition experiments are summarized in Table VII. All experiments are conducted in the same manner as in the three level mania classification.

Looking at the 4F-CV UAR performances, on the overall we observe the dominance of acoustic models' recognition of item activity (seven out of 11 items). A similar observation is that FAU and LIWC based models do not rank at the top in any of the YMRS items, while their contribution is observed in four YMRS items, where the multimodal fusion performs the best. As observed before, FAU based visual model contains cues for measuring the "Speech rate and amount" as scored in YMRS6, and gives a higher UAR performance than LIWC.

[5]https://dcf.psychiatry.ufl.edu/files/2011/05/Young-Mania-Rating-Scale-Measure-with-background.pdf

Observing the held-out test set performances, it is not surprising to see that the best model setting on 4F-CV performs worse than chance level on YMRS10 (Appearance) and YMRS11 (Insight). We also see that YMRS5 (Irritability) activity is also a difficult task to generalize, where the multimodal system performs at chance level on the test set. From the remaining eight items, seven of them have higher than 60% UAR, of which four could reach higher than 68% UAR, motivating a further study in this direction for interpretable modeling based on symptoms. While the test set performance of YMRS2 (Increased motor activity-energy, 68.6% UAR) correlates with our expectations, observing the best overall test set performance (74.3%) on YMRS3 (Sexual interest) is positively surprising and motivating. We attribute this to the acoustic features that capture arousal that also correlate with sexual arousal. We note that this aspect addresses a long standing source of violence on psychiatry nurses [70]. Considering the potential for preventing violence on nurses alone, this item deserves further study for accurate automatic prediction.

When we use the ground truth binarized YMRS item activity labels to predict mania levels, we obtain a 4F-CV UAR performance of 83.5% and a corresponding test set UAR performance of 72.2%. The test set performance is dramatically higher compared to the state-of-the-art reported here and has room for improvement, since a simple discretization to two levels with thresholding at zero causes loss of valuable intensity information. To the best of our knowledge, this is the first study to conduct item-wise YMRS activity prediction.

TABLE VII
4F-CV AND TEST SET (LAST ROW) UAR (%) SCORES OF YMRS ITEM ACTIVITY PREDICTION MODELS. MV: THREE MODAL MAJORITY VOTING.

Feature/System	YMRS1	YMRS2	YMRS3	YMRS4	YMRS5	YMRS6	YMRS7	YMRS8	YMRS9	YMRS10	YMRS11
eGEMAPS10	66.3	65.4	63.4	**56.9**	65.6	76.4	**72.5**	63.5	**67.4**	68.2	63.6
eGEMAPS	**74.1**	66.2	**69.5**	55.0	68.4	76.4	69.5	**67.9**	61.6	62.3	**64.9**
FAU	68.6	67.4	56.8	56.1	64.8	73.5	60.6	61.6	58.0	69.4	56.6
LIWC	65.7	63.1	61.0	53.3	62.3	69.8	64.8	63.8	56.1	62.5	57.9
MV (eGEMAPS10)	68.5	69.2	61.9	56.2	68.3	**78.1**	70.2	65.8	64.4	**73.8**	61.8
MV (eGEMAPS)	70.1	**69.2**	65.2	55.0	**70.7**	77.8	64.6	66.6	60.6	66.9	59.8
Test Set Performance	64.1	68.6	74.3	62.6	51.7	69.0	67.0	72.0	57.4	46.5	46.5

However, we are cautious not to over-interpret the results, due to the low number of samples, which limits the successful application of regression on original YMRS item scores.

V. DISCUSSION AND CONCLUSIONS

In this paper, we investigated mania-level classification (mania, hypomania, remission) of bipolar disorder (BD) patients using the Turkish Audio-Visual BD dataset, and proposed a trimodal architecture. We have performed a comprehensive analysis of fusion of modalities for predicting mania levels. The results showed that multimodality improves the classification of bipolar disorder. The acoustic, textual, and visual modalities complement each other and using all three modalities gives the best performance. A fusion model of just the linguistic and acoustic modalities still performs well, while requiring less information. This may be important, in case a camera is considered to be intrusive in the assessment sessions.

The best performing system combines audio, video and linguistic modalities using modality-specific weighted score fusion of weighted and unweighted Kernel ELMs, decisions of which are finally fused using majority voting. We achieve 64.8% test set UAR on this configuration, which advances the state-of-the-art on the BD dataset. The unimodal test performance breakdown of the top multimodal systems confirm the robustness of acoustic eGeMAPS descriptors, which deserves further research in depression studies.

The accuracy results we have obtained are not high enough to use the proposed system in a real-world clinical application as a decision support system for the clinician. But this may partly be due to the small size of the training corpus. There are 25, 38, and 41 clips in the training set for the remission, hypomania, and mania classes, respectively, which is not enough to generalize with a high certainty. On the positive side, the dataset is collected in a real-life scenario, and has a high level of ecological validity. It contains background noise, and in some cases, the voice of the clinician to explain points related to the questions. These issues are expected to be present if a real-life application is created, and the natural recording setup makes this database valuable. Another difficulty stems from missing information in some clips, where patients do not answer some of the questions. In one of the test case clips, the patient does not answer any questions at all. For the clinician, this may inform the diagnosis, but for an automatic system, it is difficult to take such features into account, and for the standard assessment methodology we use in this paper to ensure comparability, these cases cause issues.

Experimental results have shown that the linguistic modality contributes to the performance. We note two limitations related to this modality, which can be tackled in future studies. First, we use automatic transcription, which is prone to errors, as Turkish is not a well-studied language for automated speech recognition. Note that a fully automated recognition system was a requirement in the AVEC Challenge. For a fair and direct comparison with the works presented in the challenge, we have strictly adhered to the challenge protocol in this work, and did not use manual translation. We have, in a preliminary experiment, manually transcribed one task to assess the performance of the automatic translation, and verified that it was producing comparable results with the manual translation.

Our final model contains information from three different modalities, and each modality is represented using feature vectors with various sizes. It is especially important to create explainable [71], and more preferably, interpretable [72] models in the medical domain, but model complexity poses challenges from this perspective. In this study, we opted for a compact set of interpretable features in each modality and we analyzed the models to gain insights into the influential features in the decision-making process. The most important features in each modality correlate well with the domain knowledge and have complementary information. While the top ranking features in each modality are not individually sufficient for diagnosis, collectively they contain information that correlates with observable symptoms and have a high potential to be used in a clinical decision support system. To provide further insights into the capability of the used feature sets in symptomatic mania classification, we conducted item-wise YMRS activity modeling. In line with our expectations, this analysis showed which YMRS items cannot be accurately modeled with the used feature sets, while some items, such as ‘Sexual interest’, ‘Speech rate and amount’ as well as ‘Content’, provided promising results for future studies.

It is crucial to note that AI systems similar to the one we have proposed in this paper use a very limited set of sources in their assessment compared to the clinician, and are primarily statistical (as opposed to causal) in their nature. These limitations should be very clear in the reporting of the results, and the support offered by automatic tools should not be over-estimated. The Turkish Audio-Visual BD dataset we have opened to the research community is the first dataset including audio, visual, and text modalities in this area, and we hope it will foster the development of richer analysis tools for helping clinicians.

REFERENCES

[1] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Transactions on Affective Computing*, 2020.

[2] M. Arif, A. Basri, G. Melibari, T. Sindi, N. Alghamdi, N. Altalhi, and M. Arif, "Classification of anxiety disorders using machine learning methods: A literature review," *Insights Biomed Res*, vol. 4, no. 1, pp. 95–110, 2020.

[3] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. AVEC*, 2018, pp. 3–13.

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

[5] A. A. Salah, "Designing computational tools for behavioral and clinical science," in *Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 2021, pp. 1–4.

[6] C. C. Nuckols and C. C. Nuckols, "The diagnostic and statistical manual of mental disorders, (DSM-5)," *Philadelphia: American Psychiatric Association*, 2013.

[7] K. R. Merikangas, H. S. Akiskal, J. Angst, P. E. Greenberg, R. M. Hirschfeld, M. Petukhova, and R. C. Kessler, "Lifetime and 12-month prevalence of bipolar spectrum disorder in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 64, no. 5, pp. 543–552, 2007.

[8] World Health Organization, *The global burden of disease: 2004 update*. World Health Organization, 2008.

[9] A. F. Carvalho, J. Firth, and E. Vieta, "Bipolar disorder," *New England Journal of Medicine*, vol. 383, no. 1, pp. 58–66, 2020.

[10] K. Van Til, M. G. McInnis, and A. Cochran, "A comparative study of engagement in mobile and wearable health monitoring for bipolar disorder," *Bipolar disorders*, vol. 22, no. 2, pp. 182–190, 2020.

[11] 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, no. 5, pp. 429–435, 1978.

[12] R. Voleti, S. Woolridge, J. M. Liss, M. Milanovic, C. R. Bowie, and V. Berisha, "Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder," in *Proc. Interspeech*, 2019, pp. 1433–1437.

[13] K. Matton, M. G. McInnis, and E. M. Provost, "Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder," in *Proc. Interspeech*, 2019, pp. 1438–1442.

[14] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, "The Turkish audio-visual bipolar disorder corpus," in *ACII Asia*. IEEE, 2018, pp. 1–6.

[15] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech," *Cognitive Computation*, pp. 1–14, 2021.

[16] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. AVEC*, 2016, pp. 43–50.

[17] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2016.

[18] H. Kaya and A. A. Salah, "Eyes whisper depression: A CCA based multimodal approach," in *Proc. ICMI*. ACM, 2014, pp. 961–964.

[19] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proc. AVEC*, 2019, pp. 27–35.

[20] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 239–253, 2021.

[21] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[22] P. Baki, H. Kaya, E. Çiftçi, H. Güleç, and A. A. Salah, "Speech analysis for automatic mania assessment in bipolar disorder," in *Proc. IEEE 28th Signal Processing and Communications Applications Conf.*, 2020.

[23] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, "Bipolar disorder recognition with histogram features of arousal and body gestures," in *Proc. AVEC*, 2018, pp. 15–21.

[24] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality hierarchical recall based on GBDTs for bipolar disorder classification," in *Proc. AVEC*, 2018, pp. 31–37.

[25] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proc. AVEC*, 2018, pp. 39–45.

[26] M. Ebrahim, M. Al-Ayyoub, and M. Alsmirat, "Determine bipolar disorder level from patient interviews using Bi-LSTM and feature fusion," in *Proc. SNAMS*. IEEE, 2018, pp. 182–189.

[27] S. Amiriparian, A. Awad, M. Gerczuk, L. Stappen, A. Baird, S. Ottl, and B. Schuller, "Audio-based recognition of bipolar disorder utilising capsule networks," in *Proc. IJCNN*, July 2019, pp. 1–7.

[28] Z. Ren, J. Han, N. Cummins, Q. Kong, M. D. Plumbley, and B. W. Schuller, "Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data," in *Proc. 9th Int. Conf. on Digital Public Health*, 2019, pp. 79–83.

[29] N. AbaeiKoupaei and H. Al Osman, "A hybrid model for bipolar disorder classification from visual information," in *Proc. ICASSP*, 2020, pp. 4107–4111.

[30] ——, "A multi-modal stacked ensemble model for bipolar disorder classification," *IEEE Transactions on Affective Computing*, 2020.

[31] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proc. AVEC*, 2018, pp. 23–30.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACMMM*. ACM, 2010, pp. 1459–1462.

[33] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[34] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[35] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[36] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. NeurIPS*, 2017, pp. 3856–3866.

[37] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *Proc. FG*, 2020.

[38] S. A. Crossley, K. Kyle, and D. S. McNamara, "Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis," *Behavior research methods*, vol. 49, no. 3, pp. 803–821, 2017.

[39] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, 2010.

[40] D. M. Hilty, K. T. Brady, and R. E. Hales, "A review of bipolar disorder among adults," *Psychiatric Services*, vol. 50, no. 2, pp. 201–213, 1999.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[44] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[45] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[46] M. F. Porter *et al.*, "An algorithm for suffix stripping." *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[47] C. Gilbert and E. Hutto, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *8th Int. Conf. on Weblogs and Social Media (ICWSM-14)*, 2014.

[48] S. Loria, "textblob documentation," *Release 0.15*, vol. 2, 2018.

[49] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[50] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *Proc. Interspeech*, 2020, pp. 2097–2101.
[51] P. Ekman and E. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
[52] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Proc. WACV*, 2016.
[53] 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. ICMI*, 2015, pp. 459–466.
[54] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
[55] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[56] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2011.
[57] C. R. Rao, S. K. Mitra *et al.*, "Generalized inverse of a matrix and its applications," in *Proc. Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California, 1972.
[58] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel ELM and CNN based facial age estimation," in *Proc. CVPRW*, 2016, pp. 80–86.
[59] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.
[60] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4768–4777.
[61] L. S. Shapley, *A value for n-person games. Contributions to the Theory of Games*, H. Kuhn and A. Tucker, Eds. Princeton Univ. Press, 1953.
[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[63] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 1, pp. 64–84, 2008.
[64] K.-j. Kim and H. Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach," *Computers & Operations Research*, vol. 39, no. 8, pp. 1800–1811, 2012.
[65] S. Jayawardena, J. Epps, and E. Ambikairajah, "Ordinal logistic regression with partial proportional odds for depression prediction," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.
[66] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction," *Frontiers in Computer Science*, vol. 3, 2021.
[67] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[68] E. Lyakso, O. Frolova, and A. Grigorev, "A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders," in *Proc. SPECOM*, 2016, pp. 43–50.
[69] T. D. Brewerton, "Hyperreligiosity in psychotic disorders." *Journal of Nervous and Mental Disease*, 1994.
[70] H. Nijman, L. Bowers, N. Oud, and G. Jansen, "Psychiatric nurses' experiences with inpatient aggression," *Aggressive Behaviour*, vol. 31, no. 3, pp. 217–227, 2005.
[71] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.
[72] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

Pınar Baki received her B.Sc. and M.Sc. degrees from the Computer Engineering Department of Bogazici University in 2018, and 2020, respectively. For her masters thesis, she worked on detecting the mood states of the bipolar disorder patients from multimodal data. She is also a machine learning engineer at Arcelik Innovation Center, studying age, gender, and emotion analysis from speech data.

Heysem Kaya is an assistant professor at Utrecht University, Department of Information and Computing Sciences. He was awarded PhD degree from Computer Engineering Department, Bogazici University in 2015. He has co-authored over 70 publications on machine learning, computational paralinguistics and affective computing. His works won six Computational Paralinguistics Challenge (ComParE) awards and four awards in multimodal affective computing challenges, including three ChaLearn competitions.

Elvan Çiftçi Dr. Ciftci has a background in neuroscience and clinical psychiatry. She is currently working as a psychiatrist responsible from inpatient unit, a specialty outpatient clinic in NP Brain Hospital and an association with Üsküdar University. Her research areas are related with neurobiology of schizophrenia and bipolar disorder, using brain imaging techniques, electrophysiological paradigms, clinical databases, and in vivo studies.

Hüseyin Güleç is an associate professor and chair of the Psychiatric Department and Sleep Center at the Psychiatric and Neurological Diseases Training and Research Hospital of Health Science University Turkey. He has co-authored over 100 publications in the fields of Consultation Liaison Psychiatry, Psychometry, Schizophrenia, Sleep Medicine, General Psychiatry, Personality, Violence and Cognition. He serves as an Editorial and Advisory Board member of several journals in his field.

Albert Ali Salah (M08, SM17) is professor and chair of Social and Affective Computing at the Information and Computing Sciences Department of Utrecht Univ. and adjunct professor at the Computer Engineering Department of Bogazici Univ. He works on pattern recognition, multimodal interaction, and computer analysis of human behavior. He serves in the Steering Boards of ACM ICMI and IEEE FG, and as an associate editor of journals including IEEE Trans. Affective Computing, Pattern Recognition, and Int. Journal on Human-Computer Studies.