Noname manuscript No.
(will be inserted by the editor)

Multimodal temporal machine learning for Bipolar Disorder and Depression Recognition

Received: date / Accepted: date

Abstract Mental disorder is a serious public health concern that affects the life of millions of people throughout the world. Early diagnosis is essential to ensure timely treatment and to improve the well-being of those affected by a mental disorder. In this paper, we present a novel multimodal framework to perform mental disorder recognition from videos. The proposed approach employs a combination of audio, video and textual modalities. Using recurrent neural network architectures, we incorporate the temporal information in the learning process and model the dynamic evolution of the features extracted for each patient. For multimodal fusion, we propose an efficient late fusion strategy based on a simple feed forward neural network that we call *adaptive non-linear judge classifier*. We evaluate the proposed framework on two mental disorder datasets. On both, the experimental results demonstrate that the proposed framework outperforms the state-of-the-art approaches. We also study the importance of each modality for mental disorder recognition and infer interesting conclusions about the temporal nature of each modality. Our findings demonstrate that careful consideration of the temporal evolution of each modality is of crucial importance to accurately perform mental disorder recognition.

Keywords multimodal · mental disorder · recurrent neural network · machine learning

1 Introduction

Mental disorder affects the well-being of millions of people throughout the world. The exact number of individuals who suffer from some form of it is

Address(es) of author(s) should be given

hard to quantify, but the World Health Organization (WHO) [1] reported that almost 800 million people lived with a mental health disorder in 2017. Mental disorders are generally associated with high mortality as well as cardiovascular and respiratory diseases, diabetes, and some forms of cancer [2,3]. They are also reported to be the primary drivers of disability worldwide and they negatively impact the length and quality of life [4]. Early detection is of crucial importance in improving the overall well-being of those affected by a mental disorder [5, 6]. Unfortunately, despite the existence of several Mental Health Assessment Protocols (MHAPs), there is a significant delay in the diagnoses [7–9]. Such delays are highly associated with negative health outcomes and they usually lead to sub-optimal treatments. The use of computational models for the identification of mental disorders has gathered significant attention in recent years [10–13]. These tools could be used to assist experts in the diagnostic process, hence increasing the possibility to detect disorders at their onset. However, some important aspects have not been fully addressed in the current literature: 1) the natural language spoken by the interviewed subjects is rarely incorporated in the final diagnosis, 2) the sequential (i.e temporal) information provided in the video recordings is not fully exploited and 3) current approaches are rarely validated on multiple datasets. In this paper, we handle the aforementioned limitations by proposing a multimodal temporal framework exploiting audio, video and textual modalities. The main contributions of the paper can be summarised as follows:

1. A novel multimodal framework for human behavior analysis capable of accurately performing bipolar disorder and depression recognition. The proposed approach aims at modelling the temporal evolution of the participants' behaviours using recurrent machine learning models.

2. Experimental evaluation on two different datasets of mental disorders demonstrating that the proposed framework outperforms other state-of-the-art models.

3. A detailed analysis of the importance of each modality using a formulation based on a classical question in computer science and combinatorics: the Set Cover Problem.

2 Related Work

In this section, we present a literature review of the most common techniques used for multimodal fusion. We also review the frameworks proposed in the AVEC2018 [14] challenge which address the problem of mental disorder recognition.

2.1 Multimodal fusion

Generally, a modality indicates the way in which something is perceived or experienced [15]. For instance, we are able to characterise the world around us by hearing sounds, seeing objects, smelling odors, and so on. Similarly, multimodal machine learning [16] tries to fuse together, in a coherent and efficient way, information originated from different modalities. There are different ways in which multimodal fusion of data can be achieved: feature fusion (i.e vector concatenation), decision fusion (i.e majority voting), hybrid fusion which exploits both, and deep learning fusion. In the context of deep learning fusion, there are two main schemes: early fusion and late fusion. The former is defined as a fusion scheme that integrates the individual modalities before learning the concept. The latter is instead a fusion scheme where the modalities are first reduced to separate concept scores, which are then integrated to learn the final concept [17]. For instance, Song *et al.* [18] combined body movements, facial micro-expressions, and audio signals to perform emotion recognition. They compared three fusion schemes: early fusion, early fusion with kernel CCA [19], and late fusion with voting. In their work, the late fusion approach consistently outperformed the two early fusion approaches. Dibeklioğlu *et al.* [20] achieved audio-visual fusion by performing feature concatenation. Min-Redundancy Max-Relevance algorithm [21] was then used on the concatenated features to select the most relevant ones based on mutual information. Alghowinem *et al.* [22] used hybrid fusion to combine features from different modalities. In detail, they first concatenated feature vectors and then performed majority voting. Two SVM classifiers [23] were used for single modality classification and one SVM classifier for feature fusion classification. Huang *et al.* [24] fused features from different modalities by separately training long-short-term memory (LSTM) models and concatenating estimates from different feature sets using Support Vector Regression (SVR) [25]. Their method was able to obtain promising results in the Audio-Visual Emotion Challenge [26].

2.2 Mental disorder recognition

Several mental recognition pipelines have been presented in the AVEC2017 [26] and AVEC2018 [14] challenges. Yang *et al.* [13] proposed two novel features: a histogram based arousal feature to characterize the mood swings typical of bipolar patients, and a Histogram of Displacement (HDR) which describes the speed of the upper body during movements. Along with these novel features, they also employed Action Units and Geneva minimalistic acoustic parameter set (GeMAPS) descriptors. Each feature was fed into a Deep Neural Network (DNN) model where the output layer was discarded to concatenate all the last hidden layers' outputs into a single representation. After feeding the concatenated features into multiple tree-based classifiers, the final outcome was obtained through a majority vote. Unfortunately, this approach fails to take

into account the temporal information of the inputs; while the histograms are able to correctly model the distribution of arousal and gestures for the single frames, they do not capture the temporal relationships between frames. In addition, the lack of any cross validation and the drop of performances on the test set of this framework are clear symptoms of overfitting. Inspired by the successes of DepAudioNet [27] and Inception Networks [28, 29], Du *et al.* [30] proposed a new architecture, named IncepLSTM, for bipolar disorder recognition. By applying kernels of different sizes (1, 3 and 5) on the temporal audio sequences, they were able to demonstrate the effectiveness of IncepLSTM. However, by only taking into account the audio modality, the proposed approach did not fully incorporate the multimodal information provided by the audio-visual recordings. Xing *et al.* [31] proposed a multimodal hierarchical recall framework. It is composed of three layers, from "easy" to "hard". That is, starting in the first layer, subjects whose classification confidence is higher than a predetermined threshold are directly assigned to the corresponding category. Alternatively, the sample, named "unrecall sample", is sent to the next layer for further judgement. Each layer contains a Gradient Boosted Decision Tree (GDBTs) [32] that uses different subsets of all features. In this approach, all the modalities, audio-visual-textual, are used. Despite the high accuracy on the training set, the poor performances on the testing set seem to suggest that the model suffers from overfitting and it is unable to generalize well to unseen data. Syed *et al.* [33] introduced the concept of "turbulence features". Turbulence features are used to capture the sudden, erratic changes in the behaviour of individuals with bipolar disorder [33]. As input signals, both visual and audio modalities were employed. They used Fisher Vector (FV) to create descriptors able to provide global information about the recordings and fed these descriptors into a Weighted Extreme Learning Machine (WELM) classifier [34]. Despite the use of features able to capture the evolution of different modalities, the proposed approach scored the lowest accuracy in the AVEC2018 challenge. Finally, Zhang *et al.* [35] proposed a deep learning multimodal framework based on early fusion strategy. Specifically, they used a Multimodal Deep Denoising Autoencoder in order to learn a shared representation of audio-visual modalities. As in [33], Fisher Vector was used in order to produce global descriptors for each video and Paragraph Vector (PV) was employed to embed the natural language spoken by the patients during the interviews. Final classification was obtained using a Multitask Deep Neural Network capable of integrating bipolar disorder stage classification with Young Mania Rating Scale (YMRS) regression prediction. Despite obtaining promising results, the framework proposed in [35] does not explicitly model the temporal evolution of the frames in each video. Furthermore, the use of early fusion strategy increases overfitting and reduces the generalisability of the proposed framework.

BD State	Number of recordings in Train/Development Sets	Average Time (s)	Standard Deviation
Mania	41 / 21	276.4	246.3
Hypo-mania	38 / 21	221.1	171.4
Depression	25 / 18	151.9	65.4

Table 1 Statistics for the Train and Development sets of the Bipolar Corpus.

3 The Datasets

This section describes the two datasets used in this paper: the Bipolar Disorder Corpus [12] and the Well-being dataset [36].

In [12], the Audio-Visual Bipolar Disorder (BD) Corpus was introduced. The dataset, which was collected to shed light upon the personalized treatment of BD patients, contains audio-visual recordings of patients with BD as well as and healthy controls. In detail, 35 male and 16 female patients were recruited from the mental health department of a hospital [12]. Clinical information, including identity, age, disease severity, and used treatments, were collected using semi-structured interviews. The recordings are annotated for BD state (mania, hypo-mania, depression) as well as for Young Mania Rating Scale (YMRS) [37] by psychiatrists. During hospitalization, in every follow up day (0^{th}- 3^{rd}- 7^{th}- 14^{th}- 28^{th} day), the presence of BD was annotated and interviews recorded. In each video, the participant is asked to explain the reasons to participate in the activity, to describe happy and sad memories, and to interpret the emotions evoked by two paintings: one painting meant to inspire sadness (Van Gogh's Depression) and one meant to inspire happy feelings (Dengel's Home Sweet Home). Table 1 shows the number of samples, for each BD state, in the training and development sets [1].

Orton [36] collected a non-clinical dataset designed to enable investigation of the body modality for mental distress recognition named Well-being dataset. Participants, recruited through the University of Cambridge, were interviewed in face-to-face sessions by one researcher. To ensure a natural behaviour, participants were not aware of the main research question of the study and they were instead told that the recordings would be used for building models that can help in mental well-being. Labels were assigned using self-evaluation questionnaires. In detail, the questionnaires used were: the PHQ-8 [38,39] for depression, GAD-7 [40] for anxiety, SSS-8 [41] for somatic symptoms, and the PSS [42] for perceived stress. The dataset contains facial expressions, body motion information, gestures, and audio recordings for a total of 35 interviewed subjects.

[1] the test set is not accessible as it is reserved for evaluation in the AVEC 2018 Workshop and Challenge [14].

4 Methodology

In this section, we describe our proposed multimodal temporal framework for mental disorder recognition. In the following subsections, audio and visual modalities are grouped together as they are extracted at *frame-level* while the text modality is processed on *session-level*. Figure 1 shows an overview of the proposed framework.

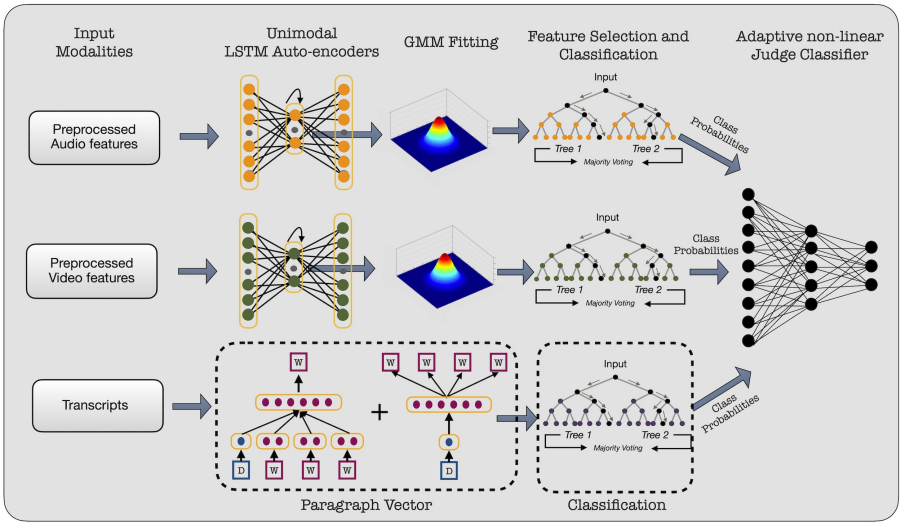

Fig. 1 Overview of the proposed framework. The audio and visual modalities are encoded using bidirectional LSTM models. The descriptors for the whole videos are generated using Fisher Vector. For the textual modality, Paragraph-Vector is proposed. A final deep classifier is used to combine the unimodal predictions.

4.1 Audio-Visual modalities

Bidirectional Long Short-Term Memory autoencoders are proposed for the audio and visual modalities. Since a video is described by a sequence of frames, each of these frames is highly correlated to the previous and next ones. As such, LSTM autoencoders (and even more bidirectional LSTM autoencoders) are a suitable solution to reduce the dimensionality of these modalities while taking into account the temporal/sequential nature of the inputs. To avoid normalisation which could wrongly corrupt the input modalities, the LSTM autoencoders were trained using mean absolute error (MAE) and linear activation function in the output layer (Figure 2). Before feeding the audio and visual modalities to the LSTM autoencoders, we propose two processing steps: dynamic computation and feature serialisation.

4.1.1 Dynamic computation

Firstly, inspired by [43] and [35], we compute the dynamic changes between subsequent frames. Let us consider a modality represented as a matrix $H \in \mathbb{R}^{n \times d}$, where n is the number of frames and d the feature dimension. Denoting each column in H as H_i with $i \in \{1, 2, \ldots, d\}$, we compute the first order dynamic, V, as $V_i = \frac{dH_i}{dt} \approx H_i - H_{i-1}$. Hence, V represents the velocity of change between subsequent frames. The use of dynamic changes rather than static features allows to emphasise the temporal evolution of audio and visual modalities.

4.1.2 Feature serialisation

In 1992, Ambady et al. [44] introduced the concept of *thin-slicing*. Thin-slicing is an established concept in psychology and philosophy which describes the ability of identifying patterns or extracting relevant information in thin slices of experience. Thin slices of individuals' behaviours could reveal important aspects of their personalities like cognitive ability [45], sexual preference [46] and personality disorders [47]. Using the idea of thin-slicing, in the next preprocessing step, we serialise the (dynamic) audio and visual features using a stride of one and experimenting with several sliding windows. Focusing on smaller segments of the videos (thin slices) allows for more accurate identification of patterns and permits to study the extent to which different modalities have to be observed to produce accurate predictions.

4.1.3 Fisher Vector

The encoded representation learned by the dynamic autoencoders only provides a *per-frame* description. However, patient labels (i.e presence of depression, bipolar state, and so on) are usually provided for an entire video. In order to unify these *per-frame* representations into a coherence *whole-video* descriptor, we propose the use of Fisher Vector (FV) [48]. FV characterises a sample by its deviation from a generative model of the data (in most cases a Gaussian Mixture Model). The deviation is defined as the gradient of the sample log-likelihood with respect to the parameters of the generative model [48]. In addition, inspired by [49], we implement power normalization and L2 normalization to generate the Improved Fisher Vectors (IFVs) as the session-level descriptors in the proposed framework. Using the information gain given by Equation (1) and a tree-based model (Random Forest), we perform feature selection to reduce the redundancy of each modality.

Finally, we use a RF model on the audio and visual features (separately) for the final classification.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{1}$$

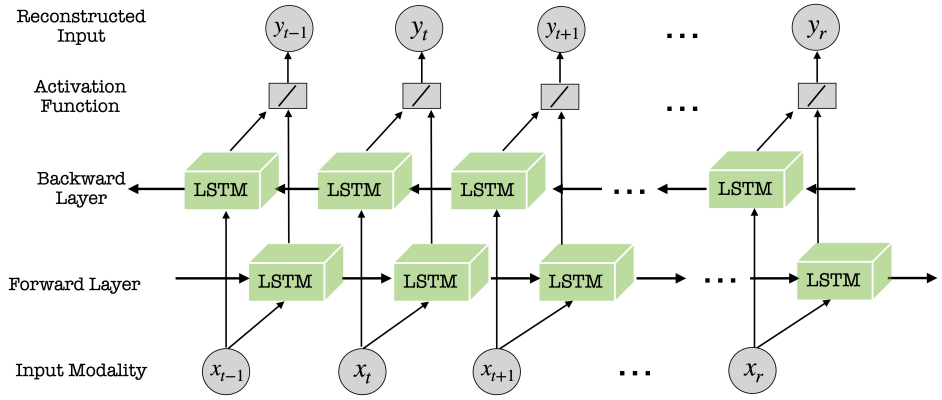

Fig. 2 Bidirectional LSTM. Two layers, one backward and one forward, are used to exploit information from past and future states simultaneously.

Parameters	Model	
	PV-DBOW	PV-DM
Aggregation method	–	{average, concatanation}
Vector size	{25, 50, 75, 100}	{25, 50, 75, 100}
Window size	–	{5,10}
Negative words	{5, 10}	{5, 10}
Hierarchical sampling	{yes, no}	{yes, no}

Table 2 Parameters explored for the Paragraph Vector models.

4.2 Textual modality

Using the Speech-to-Text Google API [2], we convert the audio of the interview sessions to transcripts. Paragraph Vector (PV) [50] is proposed to learn a fixed-length representation from these variable-length pieces of texts. PV consists of two main architectures: Distributed Memory Model Paragraph Vector (PV-DM) and Distributed Bag of Words Paragraph Vector (PV-DBOW). The former aims at predicting the current word using its surrounding word and its paragraph vector, while the latter tries to predict randomly sampled words from that document given the document ID as input. Experimental results in [50] demonstrated that PV-DM works well for most of the tasks and it consistently outperforms PV-DBOW. Moreover, the authors suggested that a combination of PV-DBOW and PV-DM might reach better performances on some tasks. The hyperparametrs explored for the textual modality are reported in Table 2. The PV embeddings are fed to a Random Forest Classifier to perform mental disorder classification.

[2] https://cloud.google.com/speech-to-text

4.3 Multimodal fusion

For the multimodal fusion, we develop a novel temporal framework based on **late fusion**. Given the predictions of each modality, a simple feed forward neural network (NN) is employed for multimodal integration. This network takes as input the class probabilities predicted from each modality using the Random Forest models. Specifically, the probabilities estimated by each modality for the same sample are concatenated and they form the representation for that sample in the NN. As such, the dimensionality of the input matrix to the NN is $(N \times P)$, where N is the number of samples, and P is the product between the number of classes of the target variable and the number of modalities used. This formulation allows to represent each sample with few, albeit useful, features. Using adaptive learning of the weights, the NN is able to learn more complex non-linear functions that map the input to the output. We call this NN, which integrates the different modalities to produce the final prediction, *adaptive non-linear judge classifier*.

5 Experimental evaluation

In this section, we present the results of the proposed framework on the Bipolar Disorder Corpus and on the Well-being dataset. For every dataset, we present the features extracted and the evaluation metrics used. We firstly evaluate our model using every modality separately then combining all modalities using the late fusion strategy proposed in 4.3. We also analyse the dynamics of each modality in the different datasets and study the relevance of each modality for mental disorder recognition.

5.1 Bipolar Disorder Recognition

5.1.1 Features

The BD Corpus contains 104 recordings for training purpose, 60 recordings for development (validation) purpose, and 54 recordings for testing. The label for each patient is his/her bipolar status: remission, hypomania, or mania. Since the labels for the 54 testing recordings are not publicly available, we only focus on the training and validation samples. Specifically, while the training set is used (as usual) for the estimation of parameters, the development set is employed for testing. For each video, facial landmarks, head pose, eye gaze, and action units are extracted as visual features using OpenFace 2.0 [51] at a rate of 30 frame-per-second (fps). Mel Frequency Cepstral Coefficients (MFCCs) and extended Geneva minimalistic acoustic parameter set (eGeMAPS) are extracted as separate acoustic features using openSMILE [52] at 10 fps. Due to the different nature extraction of MFCCs and eGeMAPS, we hypothesised that these acoustic features could provide complementary, useful information for our task. For more insight on these features, the interested reader is referred

Modality	Dimensionality
Facial Landmarks	136
Eye Gaze	6
Head Pose	6
Action Units	35
MFCC	117
eGeMAPS	69

Table 3 Features in Bipolar Disorder Corpus along with their dimensions.

to [53–55]. We aligned acoustic and visual modalities to make sure they are extracted from the same time window by concatenating three contiguous audio features (as they have the same duration as one visual feature). The audio and visual features in the BD Corpus, along with their dimensions, are reported in Table 3.

After computing the dynamic changes, we generated sequences of audio and visual features experimenting with different slides. Specifically, we implemented a moving window with a stride of one and slides of 10, 30 and 60 frames (corresponding to $\frac{1}{3}$ of a second, 1 second and 2 seconds, respectively). For the autoencoders, the amount of shrinkage depends on the size of the hidden layer(s) connecting the encoder and decoder (defined as *hidden ratio*). Given an input modality whose dimension is d, we experimented with three-layers autoencoders which produce an encoded representation of size $0.2d$ or $0.3d$. We choose to compute the Fisher Vector (FV) using 16 and 32 kernels in the Gaussian Mixture Models (GMM) as in [33, 35, 43]. Finally, for feature selection, 50 and 100 features have been empirically evaluated and five-fold cross validation was used for the final classification.

5.1.2 Metrics

We scored each classifiers using common multiclass metrics: accuracy, unweighted average recall (UAR), unweighted average precision (UAP), and $F1$ score. In the following tables, the model which exhibits the best performance, as an average of these four metrics, is shown in **bold**.

5.1.3 Unimodal results

Table 4 shows selected unimodal results for facial landmarks, eye gaze, head pose, action units, MFCC and eGeMAPS. Overall, the use of 32 kernels for the Fisher Vector resulted in better performances for all the visual modalities. On the other hand, the hidden ratio and the number of features were usually influenced by the length of the input. Longer sequences of frames usually required more features and a bigger hidden ratio in order to be accurately reconstructed. However, the most interesting finding is probably related to the

Index	Modality	Timesteps	Hidden ratio	GMM kernels	Feature number	Accuracy	UAR	UAP	F1
(1)	Landmarks	30	0.3	32	50	0.5333	0.5370	0.5489	0.5321
(2)	Landmarks	30	0.3	32	100	**0.6167**	**0.6190**	**0.6226**	**0.6173**
(3)	Eye Gaze	30	0.2	32	50	**0.5833**	**0.5847**	**0.5873**	**0.5826**
(4)	Eye Gaze	30	0.2	32	100	0.5333	0.5344	0.5417	0.5372
(5)	Head Pose	10	0.2	32	50	**0.5500**	**0.5423**	**0.5578**	**0.5405**
(6)	Head Pose	10	0.2	32	100	0.5000	0.4841	0.5429	0.4606
(7)	AUs	60	0.3	32	50	0.6000	0.5794	0.6022	0.5411
(8)	AUs	60	0.3	32	100	**0.6500**	**0.6323**	**0.6972**	**0.6167**
(9)	MFCC	60	0.2	16	50	0.6167	0.6032	0.6760	0.5953
(10)	MFCC	60	0.3	32	100	**0.8000**	**0.7989**	**0.8151**	**0.8040**
(11)	eGeMAPS	10	0.2	16	50	0.5500	0.5397	0.5761	0.5358
(12)	eGeMAPS	10	0.3	16	100	**0.6167**	**0.6005**	**0.6728**	**0.5833**

Table 4 Selected results for the evaluation of audio-visual modalities. Different timesteps result optimal for the different visual modalities suggesting diverse temporal intervals required for each modality to display. Facial Action Units (AUs) and MFCC exhibited the best performance as visual and audio modalities, respectively.

nature of the modalities themselves. Experimenting with different sequences of frames and evaluating their performances suggested that different visual features are displayed in different temporal intervals. For instance, since the strongest association between facial landmarks and bipolar disorder was found when using landmark sequences of 1 second (30 frames), this temporal interval is probably the accurate window for a "landmark action" to take place [3]. While 1 second is also optimal for capturing an "eye gaze action", in our experiments, only $\frac{1}{3}$ of a second (10 frames) was needed for a "head pose action" and 2 seconds (60 frames) for an "AU action". Similarly, 2 seconds were needed to identify an "MFCC action" and $\frac{1}{3}$ of a second (10 frames) for an "eGeMAPS action".

Compared to audio and visual performance, the results in Table 5 suggests the textual modality is less predictive of bipolar disorder. The poor performance of the textual modality for bipolar disorder recognition is probably linked to the limited size of the BD corpus. Although the natural language spoken by the patients could be strongly associated to mental health disorders, pre-training on external, large-scale textual resources is usually deemed necessary for improving the model performances [56].

5.1.4 Multimodal results

Table 6 (1^{st} and 2^{nd} rows) shows the results of late fusion aggregation using the 7 modalities (landmarks, eye gaze, head pose, facial action units, MFCC,

[3] Analogously to the concept of thin-slice, we used the word "action" to refer to a piece of relevant information – or change – about a modality (for instance an eyebrow raise or an head shake) which can be captured in a small fragment of a video. We refer to "temporal interval" as the (minimum) amount of time the video fragment has to last for in order to capture that information.

Index	Text Model	Vector size	Window size	Negative words	Accuracy	UAR	UAP	F1
(1)	PV-DBOW	50	–	10	0.4333	0.4206	0.4500	0.4014
(2)	PV-DBOW	75	–	hs	**0.5333**	**0.5212**	**0.5575**	**0.5125**
(3)	PV-DM (*av*)	75	5	hs	0.4833	0.4788	0.4799	0.4785
(4)	PV-DM (*av*)	100	5	10	**0.5333**	**0.5265**	**0.5333**	**0.5254**
(5)	PV-DM (*conc*)	25	10	hs	**0.5167**	**0.5000**	**0.4952**	**0.4730**
(6)	PV-DM (*conc*)	50	5	10	0.3667	0.3492	0.2424	0.2855
(7)	DBOW + DM (*av*)	50	10	10	**0.5333**	**0.5185**	**0.5487**	**0.5005**
(8)	DBOW + DM (*av*)	100	10	hs	0.4833	0.4815	0.4861	0.4821
(9)	DBOW + DM (*conc*)	50	10	10	0.4000	0.3810	0.2775	0.3182
(10)	DBOW + DM (*conc*)	100	10	hs	**0.5167**	**0.5026**	**0.5556**	**0.4860**

Table 5 Selected results for the evaluation of textual modality. hs = hierarchical softmax. *av* and *conc* refer to the type of aggregation method: average or concatenation.

Integration Method	Accuracy	UAR	UAP	F1 score
Mean	0.8667	0.8545	0.8984	0.8562
Majority-voting	0.8667	0.8624	0.8722	0.8649
Adaptive Classifier	**0.9167**	**0.8836**	**0.8857**	**0.8831**

Table 6 Late fusion aggregation with *mean*, *majority-voting*, and *adaptive non-linear judge classifier*. While *mean* and *majority-voting* exhibit similar performance, the *adaptive non-linear judge classifier* is able to learn more complex mapping functions which result in better performances.

eGeMAPS and textual embeddings) and two simple fusion methods: mean and majority voting. These two non-adaptive aggregation methods show similar performances on all the metrics. However, these methods are unable to dynamically assign weights to the input modalities according to their relevance for the target predictions. As shown in Table 6 (3$^{\text{rd}}$ rows), the *adaptive non-linear judge classifier* neural network (NN) is able to learn more efficient, albeit more complex, mapping functions which exhibit better performance. By allocating appropriate parameters during the learning phase, the NN selects the modalities to "trust" more and shows good predictive power of bipolar status.

5.1.5 Modality importance

To study the dependencies of the modalities in the BD Corpus and their relevance for prediction, we employed a simple formulation based on a classical question in computer science and combinatorics: the **Set Cover Problem** (SC). The problem is simple: given a set of elements $\mathcal{U} = \{1, \ldots, n\}$, called *universe*, and a collection $\mathcal{S}$ of m sets whose union is equal to the universe, find the smallest sub-collection of $\mathcal{S}$ which equals the universe. Formally:

$$\begin{aligned} \text{minimize} \sum_{S \in \mathcal{S}} x_S & \quad \text{(minimize the number of sets)} \\ \text{subject to} \sum_{S:e \in S} x_S \geq 1, \forall e \in \mathcal{U} & \quad \text{(cover every element of the universe) and} \\ x_S \in \{0,1\}, \forall S \in \mathcal{S} & \quad \text{(every set is either in the set cover or not)} \end{aligned} \tag{2}$$

For our classification problem, we set the universe to be the union of all the samples which were correctly classified by each modality and employed a brute-force approach to find all the optimal solutions. Table 7 shows the sets of modalities that have been identified as solutions for the SC problem. Out of the 7 modalities, the minimal subset which allows for the correct classification of all the samples in the universe only requires 4 modalities (although different combinations of these are possible). From Table 7 it is evident that facial landmarks allow for correct classification of samples for which all the other modalities fail. As such, landmarks have been identified as a necessary modality in all the SC solutions. For audio modality, MFCC is chosen in 8 out of the 9 optimal solutions. The importance of MFCC is a reflection of the good classification performance reported in Table 4. It is also interesting to notice that one of the solutions only includes visual modalities (1^{st} row of Table 7). This suggests that visual features are complementary and their diversity allows to capture useful, non-redundant information. Since 4 of the solutions in Table 7 employed an audio-visual combination, and other 4 employed an audio-visual-textual combination, it is possible to conclude that aggregation of features originated from different communication channels usually benefits the performance of prediction. Nonetheless, our evaluation suggested that a careful combination of modalities is an important step to reduce redundant information and lower the computational burden of processing non-informative modalities.

5.1.6 Comparison with state-of-the-art

We evaluated the proposed model against the baseline of the AVEC2018 Challenge [14] as well as the frameworks in Du et al. [30], Yang et al. [13], Xing et al. [31], Syed et al. [33] and Zhang et al. [35]. All the aforementioned frameworks, including the proposed one, were scored on the same test set, suggesting the fairness of the following comparisons.

Table 8 shows that the proposed approach outperforms all the previous state-of-the art methods reaching an accuracy of 0.916 and UAR of 0.883. From Table 8, it is clear that the best performing frameworks employ an audio-visual-textual combination. For instance, the works in [30] and [33] made use of audio and audio-visual features respectively and they scored the lowest UARs in the AVEC 2018 Challenge [14]. However, differently from the proposed model, other frameworks only employed static features and failed to capture

Set Cover Solutions	
Landmarks, Action Units, Gaze, Pose	video
Landmarks, Action Units, Gaze, MFCC	video+audio
Landmarks, Action Units, Pose, MFCC	video+audio
Landmarks, Pose, MFCC, eGeMAPS	video+audio
Landmarks, Gaze, MFCC, eGeMAPS	video+audio
Landmarks, Action Units, MFCC, Textual	video+audio +text
Landmarks, Gaze, MFCC, Textual	video+audio +text
Landmarks, Pose, MFCC, Textual	video+audio +text
Landmarks, MFCC, eGeMAPS, Textual	video+audio +text

Table 7 Set Cover Problem solutions. One of the solutions (1st row) only uses facial modalities, suggesting that they contain complementary, useful information. While Landmarks is present in all the solutions as a visual modality, MFCC is selected 8 out 9 times. Overall, since 4 of the solutions employed an audio-visual combination, and other 4 employed an audio-visual-textual combination, it is possible to state that fusion of features originated from different modalities benefits the detection of mental illness.

Model	Architecture	UAR	Accuracy
Baseline [14]	audio-visual SVM	0.635	NA
Du et al. [30]	audio LSTM	0.651	0.650
Yang et al. [13]	audio-visual DNN	0.714	0.717
Xing et al. [31]	audio-visual-textual GDBT	0.868	NA
Syed et al. [33]	audio-visual WELM	0.635	NA
Zhang et al. [35]	audio-visual-textual DDA	0.709	0.717
Proposed	audio-visual-textual LSTM	**0.883**	**0.916**

Table 8 Comparison with the state-of-the-art on the BD Corpus. SVM = Support Vector Machine. LSTM = Long Short Term Memory. DNN = Deep Neural Network. GDBT = Gradient Boosted Decision Tree. WELM = Weighted Extreme Learning Machine. DDA = Deep Denoising Autoencoder. The proposed approach exhibits better accuracy and UAR compared to all the other frameworks on the same test set.

the temporal evolution of the modalities. Due to the lack of cross validation and the use of early fusion schemes, the models proposed in [35] and [31] are prone to overfitting and they poorly generalise to new data. Compared to the AVEC2018 Challenge baseline [14], our model improves upon UAR by 26.84%. The good classification performance of the proposed framework suggests that modelling the dynamic and temporal information of the video recordings and using an adaptive late fusion strategy can be successfully used for predicting bipolar disorder status.

5.2 Depression Recognition

5.2.1 Features

For the 35 participants in the Well-Being dataset, we used the same features as extracted by [57]. Table 9 shows the modalities and their dimensions. The target variable for this dataset is a continuous value of self-reported depression (in the range $[0, 19]$). Similarly to [57], we converted these continuous values into binary classes using a threshold of 7. The textual modality is not used to fairly compare the proposed framework and the work in [57].

Modality	Dimensionality
Fidget	9
Gaze	8
Action Units	35
MFCC	13

Table 9 Features in the Well-Being dataset and their dimensions.

5.2.2 Metrics

We evaluated each classifiers using common binary metrics: accuracy, recall, precision, and $F1$ score. All results on the Well-Being dataset are calculated as the mean of three-fold cross validation results.

5.2.3 Unimodal results

Overall, all the modalities exhibited good predictive performance for the binary depression label. From Table 10, it is evident that the performances are generally worse when the number of selected features is high ($\geq$ 200). Indeed, all the modalities reached the best performance when using 50 or 100 features. This is probably due to the overfitting of the models when many non-informative features are used. Similarly to the results on the Bipolar Corpus, the use of more GMM kernels was beneficial for the predictions. All the highest scores in Table 10 were obtained with 32 kernels. More interestingly, action units, eye gaze, and MFCC exhibited best performances with window sizes of 60, 30, and 60 frames respectively. These findings match the results obtained on the Bipolar Corpus for the same modalities (§5.1.3). This further indicates that different features are displayed in different temporal intervals, and careful consideration of such intervals is crucial for good predictions. Moreover, these results have allowed identifications of different quantitative intervals which seemed to be optimal for each modality. No existing work, that we are aware of, analytically quantified the time required for distinctive features to display, especially in the mental health domain.

Index	Modality	Timesteps	Hidden ratio	GMM kernels	Feature number	Accuracy	Recall	Precision	F1
(1)	AUs	60	0.4	32	50	**0.7489**	**0.6136**	**0.7350**	**0.7818**
(2)	AUs	60	0.4	32	100	0.7027	0.7652	0.7474	0.8056
(3)	Fidget	10	0.4	32	50	0.6724	0.6162	0.8197	0.6938
(4)	Fidget	10	0.4	32	100	**0.7359**	**0.5934**	**0.8463**	**0.6431**
(5)	Eye Gaze	30	0.4	32	50	**0.6869**	**0.7980**	**0.6379**	**0.7621**
(6)	Eye Gaze	30	0.4	32	100	0.6241	0.5934	0.7167	0.6551
(7)	MFCC	60	0.4	32	50	**0.7179**	**0.7096**	**0.6530**	**0.7652**
(8)	MFCC	60	0.4	32	100	0.6378	0.7071	0.7835	0.6858

Table 10 Selected results for the evaluation of modalities on the Well-Being dataset. Different timesteps result optimal for the different modalities suggesting diverse temporal intervals required for each feature to display. In line with the findings reported for the BD Corpus, action units (AUs), eye gaze, and MFCC exhibited best performances with window sizes of 60, 30, and 60 frames respectively. Among all the features, Facial Action Units (AUs) had the best performance.

Integration Method	Accuracy	Recall	Precision	F1
Mean	0.8438	0.8451	0.8438	0.8436
Majority-voting	0.7812	0.7882	0.7976	0.7804
Adaptive Classifier	**0.858**	**0.883**	**0.867**	**0.870**

Table 11 Late fusion aggregation with *mean*, *majority-voting*, and *adaptive non-linear judge classifier*. The *adaptive non-linear judge classifier* is the best performing feature aggregation method due to its ability to dynamically assign weights to each modality according to their relevance for the final predictions.

5.2.4 Multimodal results

As baseline, we first computed the multimodal results using majority-voting and mean as feature aggregation methods. Results are shown in Table 11 (1^{st} and 2^{nd} rows). Using the mean of the probabilities computed by the unimodal models seems to perform better compared to majority-voting. As before, we further evaluated the performance of the *adaptive non-linear judge classifier* neural network (NN) for multimodal fusion (Table 11 3^{rd} row). The use of an adaptive classifier exhibited better results compared to simple aggregation methods like mean and majority voting. With accuracy of 0.858, recall of 0.883, precision of 0.867 and $F1$ score of 0.870, this NN outperforms the other fusion approaches on all metrics.

5.2.5 Modality importance

As for the Bipolar Disorder Corpus, we studied the importance of modalities in the Well-Being dataset. However, since in this case we only employed 4 modalities, a simple visualisation allows to assess the relevance of each modality for distress classification. Figure 3 shows the percentage of samples that

each modality was able to correctly classify as a Venn diagram. It is clear that all the modalities capture complementary information and each modality uses its diversity to correctly classify samples which all the others fail to classify. From Figure 3, it appears that fidget features are the most successful (4.7%) in extracting information which the other modalities fail to associate to psychological distress. This emphasises the importance of features extracted from the body modality and it demonstrates the complementary of diverse modalities (facial, body and audio features).

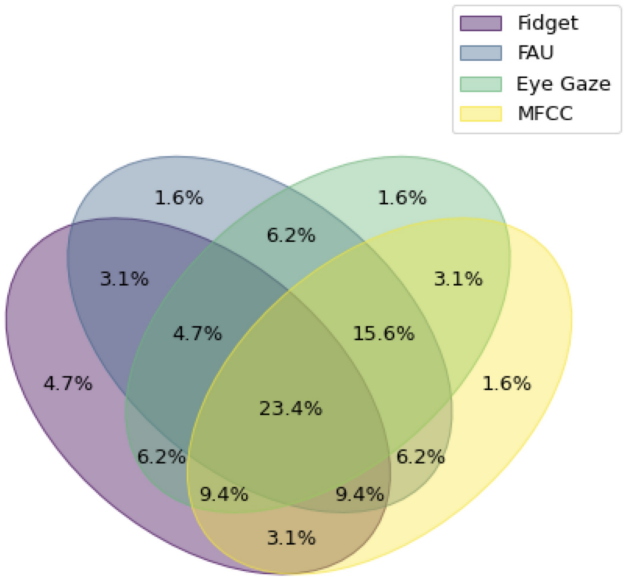

Fig. 3 Venn diagram for modality importance. Each modality captures useful information that other modalities fail to associate to mental disorder (i.e each modality has a non-overlapping percentage of samples which are correctly classified by that modality only).

5.2.6 Comparison with state-of-the-art

The current state-of-the-art approach is described in [57]. Since in [57] the only reported metric is $F1$ score, we include the performance of the proposed framework as measured by the same metric in Table 12 (for more complete information about the performances the reader is referred to Table 11). Also for distress recognition, with $F1$ score of 0.870, the proposed framework exhibited very good performances and it outperformed the state-of-the-art approach by 10.34%. Similar to [35], the framework in [57] uses early fusion and static features. By modelling the dynamic evolution of the modalities, the proposed model is able to extract more useful information and accurately perform depression recognition.

Model	Architecture	F1
Lin et al. [57]	audio-visual DDA	0.787
Proposed	audio-visual LSTM	**0.870**

Table 12 Comparison with the state-of-the-art on the Well-Being data. LSTM = Long Short Term Memory. DDA = Deep Denoising Autoencoder. The proposed approach exhibits better F1 score compared to the state-of-the-art model on the Well-Being data.

6 Conclusion and future work

In this paper, we presented a novel dynamic and multimodal framework to perform bipolar disorder and depression recognition from video recordings. Combination of audio, video and textual modalities is suggested to fully exploit all the information from the videos. By emphasising the dynamic context and using LSTM models, we aimed at including the temporal information in the learning process. Experimental evaluations on two different datasets showed that the proposed framework outperforms other state-of-the-art approaches. Since it successfully identified two types of mental disorder, bipolar disorder and depression, the proposed framework could be easily generalised to other datasets. Moreover, experimental evaluation allowed us to infer interesting temporal properties of each modality. Specifically, by exploiting feature serialisation with multiple timesteps, we identified temporal intervals in which each modality is most likely to be displayed. To the best of our knowledge, such findings were never reported in the literature and could demonstrate useful for future research in related fields.

In this paper, we experimented with simple three-layers LSTM autoencoders. However, it is worthwhile to explore the use of more sophisticated architectures like Attention-based LSTMs. Furthermore, since most datasets are composed of multiple video recordings of the same patient at different points in time, it would be useful to augment the final neural network with constraints that incorporate this domain knowledge. By informing the learning about "different samples being the same patient" and constraining the output to be relatively close for each of those samples could demonstrate beneficial for the final predictions.

7 Declarations

7.1 Funding

Part of this research is funded by King's College Cambridge.

7.2 Conflicts of interest/Competing interests

The authors declare no conflict of interest.

7.3 Availability of data and material

The Bipolar Disorder Corpus was part of the AVEC2018 challenge and it can be accessed by contacting the authors. The Well-Being dataset is a private dataset collected at University of Cambridge by Dr Marwa Mahmoud.

7.4 Code availability

Code can be found at https://github.com/cecca46/MentalDisorderRecogntion.

7.5 Ethics approval

Not applicable.

7.6 Consent to participate

Not applicable.

7.7 Consent for publication

Not applicable.

References

1. Hannah Ritchie and Max Roser. Mental health. *Our World in Data*, 2020. https://ourworldindata.org/mental-health.
2. Lisa Dixon, Leticia Postrado, Janine Delahanty, Pamela J Fischer, and Anthony Lehman. The association of medical comorbidity in schizophrenia with poor physical and mental health. *The Journal of nervous and mental disease*, 187(8):496–502, 1999.
3. Francine Cournos, Karen M McKinnon, and Greer Sullivan. Schizophrenia and comorbid human immunodeficiency virus or hepatitis c virus. *Journal of Clinical Psychiatry*, 66, 2005.
4. Alize J Ferrari, Fiona J Charlson, Rosana E Norman, Scott B Patten, Greg Freedman, Christopher JL Murray, Theo Vos, and Harvey A Whiteford. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS medicine*, 10(11), 2013.
5. Lucio Ghio, Simona Gotelli, Maurizio Marcenaro, Mario Amore, and Werner Natta. Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *Journal of affective disorders*, 152:45–51, 2014.
6. A Carlo Altamura, Bernardo Dell'Osso, Heather A Berlin, Massimiliano Buoli, Roberta Bassetti, and Emanuela Mundo. Duration of untreated illness and suicide in bipolar disorder: a naturalistic study. *European archives of psychiatry and clinical neuroscience*, 260(5):385–391, 2010.
7. Cheung Ricky, Madi Nawaf O'Donnell Siobhan, et al. Factors associated with delayed diagnosis of mood and/or anxiety disorders. *Health promotion and chronic disease prevention in Canada: research, policy and practice*, 37(5):137, 2017.
8. Alan E Kazdin and Stacey L Blase. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on psychological science*, 6(1):21–37, 2011.
9. Philip S Wang, Patricia Berglund, Mark Olfson, Harold A Pincus, Kenneth B Wells, and Ronald C Kessler. Failure and delay in initial treatment contact after first onset of mental disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):603–613, 2005.
10. James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48, 2013.
11. Heysem Kaya and Albert Ali Salah. Eyes whisper depression: A cca based multimodal approach. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 961–964, 2014.
12. Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.
13. Le Yang, Yan Li, Haifeng Chen, Dongmei Jiang, Meshia Cédric Oveneke, and Hichem Sahli. Bipolar disorder recognition with histogram features of arousal and body gestures. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 15–21, 2018.
14. Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13. ACM, 2018.
15. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.
16. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

17. Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.
18. Yale Song, Louis-Philippe Morency, and Randall Davis. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 237–244, 2013.
19. David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
20. Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310, 2015.
21. Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
22. Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. Cross-cultural detection of depression from nonverbal behaviour. In *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.
23. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
24. Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Zhengqi Wen, Minghao Yang, and Jiangyan Yi. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2017.
25. Mariette Awad and Rahul Khanna. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA, 2015.
26. Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.
27. Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42, 2016.
28. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
29. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
30. Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 23–30, 2018.
31. Xiaofen Xing, Bolun Cai, Yinhu Zhao, Shuzhen Li, Zhiwei He, and Weiquan Fan. Multi-modality hierarchical recall based on gbdts for bipolar disorder classification. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 31–37, 2018.
32. Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
33. Zafi Sherhan Syed, Kirill Sidorov, and David Marshall. Automated screening for bipolar disorder from audio/visual modalities. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 39–45, 2018.
34. Weiwei Zong, Guang-Bin Huang, and Yiqiang Chen. Weighted extreme learning machine for imbalance learning. *Neurocomputing*, 101:229–242, 2013.
35. Ziheng Zhang, Weizhe Lin, Mingyu Liu, and Marwa Mahmoud. Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020). IEEE*, 2020.

36. Indigo JD Orton. Vision based body gesture meta features for affective computing. *arXiv preprint arXiv:2003.00809*, 2020.
37. Robert C Young, Jeffery T Biggs, Veronika E Ziegler, and Dolores A Meyer. A rating scale for mania: reliability, validity and sensitivity. *The British journal of psychiatry*, 133(5):429–435, 1978.
38. Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.
39. Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.
40. Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.
41. Benjamin Gierk, Sebastian Kohlmann, Kurt Kroenke, Lena Spangenberg, Markus Zenger, Elmar Brähler, and Bernd Löwe. The somatic symptom scale–8 (sss-8): a brief measure of somatic symptom burden. *JAMA internal medicine*, 174(3):399–407, 2014.
42. Sheldon Cohen, T Kamarck, R Mermelstein, et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, 10, 1994.
43. Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 22(2):525–536, 2017.
44. Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.
45. Nalini Ambady and Heather M Gray. On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of personality and social psychology*, 83(4):947, 2002.
46. Nalini Ambady, Mark Hallahan, and Brett Conner. Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of personality and social psychology*, 77(3):538, 1999.
47. Jacqueline NW Friedman, Thomas F Oltmanns, and Eric Turkheimer. Interpersonal perception and personality disorders: Utilization of a thin slice approach. *Journal of Research in Personality*, 41(3):667–688, 2007.
48. Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.
49. Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.
50. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.
51. Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.
52. Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.
53. Jonathan T Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147. International Society for Optics and Photonics, 1997.
54. Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11, 2000.
55. Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.
56. Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.
57. Weizhe Lin, Indigo Orton, Mingyu Liu, and Marwa Mahmoud. Automatic detection of self-adaptors for psychological distress. In *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020). IEEE*, 2020.