# Multi-modality Hierarchical Recall based on GBDTs for Bipolar Disorder Classification

Xiaofen Xing
South China University of Technology(SCUT)
Guangzhou, China
xfxing@scut.edu.cn

Bolun Cai
South China University of Technology(SCUT)
Guangzhou, China
caibolun@gmail.com

Yinhu Zhao
South China University of Technology(SCUT)
Guangzhou, China
zhaoyinhu0502@gmail.com

Shuzhen Li
South China University of Technology(SCUT)
Guangzhou, China
ee_17szli@mail.scut.edu.cn

Zhiwei He
South China University of Technology(SCUT)
Guangzhou, China
zhiwei.he96@gmail.com

Weiquan Fan
South China University of Technology(SCUT)
Guangzhou, China
weiquan.fan96@gmail.com

## ABSTRACT

In this paper, we propose a novel hierarchical recall model fusing multiple modality (including audio, video and their combination) for bipolar disorder classification, where patients with different mania level are recalled layer-by-layer. To address the complex distribution on the challenge data, the proposed framework utilizes multi-model, multi-modality and multi-layer to perform domain adaptation for each patient and hard sample mining for special patients. The experimental results show that our framework achieves competitive performance with Unweighed Average Recall (UAR) of 57.41% on the test set, and 86.77% on the development set.

## CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*; • **Applied computing** → *Psychology*;

## KEYWORDS

Bipolar classification, Hierarchical recall, GBDTs, Multi-modality

**ACM Reference Format:** Xiaofen Xing, Bolun Cai, Yinhu Zhao, Shuzhen Li, Zhiwei He, and Weiquan Fan. 2018. Multi-modality Hierarchical Recall based on GBDTs for Bipolar Disorder Classification. In 2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea. ACM, NY, NY, USA, 7 pages. https://doi.org/10.1145/3266302.3266311

## 1 INTRODUCTION

Bipolar disorder (BD) [1], previously known as manic depression, is a mental disorder that causes periods of depression and abnormally elevated mood. BD is a highly prevalent mental disorder in young adults, and ranks top-10 disorder by disability-adjusted life year (DALY) [18, 27]. Therefore, early and accurate recognition of BD episodes through machine learning techniques is greatly significant. The Bipolar Disorder Sub-challenge (BDS) of Audio Visual Emotion Challenge (AVEC) 2018 [10] requires participants into remission, hypo-mania and mania depending on audio and visual analysis.

The AVEC 2018 BDS is a new task in the scope of mental disorder analysis while AVEC 2016[28] and AVEC 2017[23] focus on the depression recognition. Both BD and Depression belong to mood disorder (MD) related with psychical analysis. Therefore, although BD classification and depression detection are different tasks, features and models can still be referenced from each other.

Recently, behavioral signal processing and machine learning have been utilized for automatic MD recognition, where feature selection plays an important role. A various of primary features [23, 28], such as Action Units (AUs) and Mel-Frequency Cepstral Coefficients (MFCCs), are extracted and applied in mental disorder analysis. To obtain more robust information, statistical analysis is applied to merge these frame-level features into high-level descriptors. There are two kinds of information integration strategies, including subject-level [25, 30, 31] and topic-level [12], respectively. In subject-level processing, statistical functions or regression functions are performed over an interview[25, 30, 31]. In topic-level processing, statistical operations are implemented in specific topics separately[12]. These topic based features retain fine-grained features and play an important role in MD recognition. To eliminate overfitting, feature selection is used further to discard redundant features or select more discriminative features[12, 25]. In [12], a feature selection algorithm called correlation-based feature subset selection (CFS) is introduced to select informative features. In [25], text features combination is optimized by recurrently removing one or two features and evaluating the rest features via a machine learning model.

Based on feature engineering, multi-modality fusion is commonly used in previous works. There are two kinds of fusion. One is feature-level fusion[12, 26], the other is decision-level fusion[5, 25, 30, 31]. Because different modalities have different domains, it is difficult for feature-level fusion to take the advantages of all modalities simultaneously. Decision-level fusion easily takes the characteristic of each modality into consideration. The key of decision-level fusion is the fusing structure. In [5, 31], the predictions of different

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*AVEC'18, October 22, 2018, Seoul, Republic of Korea*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266311

modalities are fused via linear weighted summation. Bo Sun et al [25] concatenates predictions of different modalities and use Random Forest to output integrated results. In [30], predictions of three modalities are fused using a DNN. It is worth noting that in above methods, the final prediction for every instance is made at the same layer, which assumes different patients have same domain in each modality. However, with regard to the particularity of the given BD corpus, our experiments show that different patients, especially those who are difficult to identify, have different domains in multi modalities. For instance, with respect to the same mania level, some patients have loud voice but calm expression while some have low voice but exaggerated expression.

To build a classification model for the BD corpus, one intuition is to perform different decision strategies for different patients instead of classifying all patients at same layer. Therefore, targeting to AVEC 2018 BDS, a novel hierarchical recall framework is proposed in this paper for multi-modality fusion. In order to realize domain adaptation and hard sample mining, the predictions of patients are made layer-by-layer, where patients with high confidence level are first recalled while patients with low confidence level are delivered to next layer to perform further judgment. This hierarchical model is symmetric and take the advantages of different modalities in each layer. To the best of our knowledge, this is the first time that hierarchical recall framework is used in BD classification.

The paper is organized as follows. In Section 2, the topic modeling based on multi-modality features are discussed in details. Section 3 describes our multi-modality hierarchical recall framework based on GBDTs. Section 4 analyzes the experimental results, and the conclusion is given in Section 5.

## 2 BIPOLAR DISORDER FEATURE ANALYSIS

### 2.1 Topic Modeling

According to the BDS introduction, each patient was required to complete several tasks in front of video camera. According by [12], fine-grained features extracted by topic-level models were proven effective for performance improvement. However, the corpus provided by AVEC 2018 BDS does not contain the textual transcripts of interviews. Therefore, we perform automatic topic segmentation to conduct context analysis and topic segmentation. First, we use Automatic Speech Recognition [1] and Neural Machine Translation [2] of Google Cloud Platform (GCP) to transform Turkish audio into English text with word time offsets. As our observation on these translated text, most of interviews include 3 fixed topics shown as table 1. Then, an automatic topic segmentation was implemented via detection for counting numbers. That means we detected two time points at which the counting numbers occur and disappear respectively. Because there is a small portion of transcriptions that don't contain counting numbers, each of their corresponding videos is divided into three equal segments. Finally, each video is divided into three topics.

Therefore, topic modeling, where audio, video and text features are generated for each topic segment respectively, can be performed in this corpus. There are two inherent advantages of topic modeling:

[1] https://cloud.google.com/speech/
[2] https://cloud.google.com/translate/

**Table 1: The list of 3 topics in BD corpus**

| Ind. | Topic Abbr. | Required Task |
|---|---|---|
| 1 | negative_task | Describe why you come here<br>Depict Van Gogh's *Depression*<br>Describe the worst memory |
| 2 | neutral_task | Count 1-30<br>Count 1-30 again (often faster) |
| 3 | positive_task | Depict Dengel's *Home Sweet Home*<br>Describe the best memory |

(1) **Detailed information of each topic is retained.** It easily makes sense that applying statistical functions to short-term features over entire interview may loss detailed features.
(2) **Each topic is characterized by different features.** With feature selection analysis, we can focus on different features under each topic context such as the crying when describing the worst memory, and the smiling when describing the best memory.

### 2.2 Features Extraction

*2.2.1* ***Audio Features.*** The audio features provided by AVEC 2018 consist of low-level descriptors, the turn timings of speech and the timings between each sound separator. LLDs are extracted with openSMILE [24], including MFCCs [20] (0-12, delta, delta-delta), and eGeMAPS [9], which are common features for automatic audio analysis. The timings between each sound separator are obtained by template matching of the sound event, recoding the start time and end time of each answer to the questions. We concatenate these two as low-level feature extracted by openSMILE $f_{smile} \in \mathbb{R}^{62}$, composed of 23 eGeMAPS features and 39 MFCCs features. At first, we simply perform four statistic functions in the low-level features, including maximum, minimum, average and standard deviation, which are respectively defined as

$$\begin{cases} f_{smile}^{\max} = \max\limits_{t\in\tau} f_{audio}(t) \\ f_{smile}^{\min} = \min\limits_{t\in\tau} f_{audio}(t) \\ f_{smile}^{mean} = \frac{1}{\tau}\sum\limits_{t\in\tau} f_{audio}(t) \\ f_{smile}^{\text{var}} = \sqrt{\frac{1}{\tau}\sum\limits_{t\in\tau}\left(f_{smile}(t) - f_{smile}^{\text{mean}}\right)^2} \end{cases}, \quad (1)$$

where $\tau$ is the time sequence. Then, we further concatenate the features $F_{smile} \in \mathbb{R}^{248}$ for each topic: $F_{smile} = \{f_{smile}^{\max}, f_{smile}^{\min}, f_{smile}^{\text{mean}}, f_{smile}^{\text{var}}\}$. To describe global and local information, we respectively extract the global feature $F_{smile}^{all} \in \mathbb{R}^{248}$ and topic-level feature $F_{smile}^{1-3} \in \mathbb{R}^{248\times3}$.

Based on the turn timings of speech, we utilize 5 statistic functions (maximum, minimum, average, sum, and standard deviation) to extract the duration time $f_{duration} \in \mathbb{R}^5$ of each sentence and the pause time $f_{pause} \in \mathbb{R}^5$ between sentences, and concatenate them as time feature $F_{time} = \{f_{duration}, f_{pause}\} \in \mathbb{R}^{10}$.

That is, for each subject, audio features with $(248+248\times3+10) = 1002$ dimensions are extracted.

*2.2.2* ***Video Features.*** Motivated by Young Mania Rating Scale (YMRS) [32], four kinds visual features (AUs, eyesight, emotion and body movement) related to YMRS are extracted.

Facial Action Units (FAUs) describe human facial expression, which can judge whether the subject suddenly laughs or cries without reason. Therefore, we use FAUs provided by AVEC2018, including the intensity of 17 key AUs in each frame. Inspired by Motion History Histogram (MHH) [17, 31], we extract the 17 key AUs MHH features $f_{au\_mhh} \in \mathbb{R}^{17\times5\times10=850}$ with 5 time intervals $M_k \in \{10, 20, 30, 40, 50\}$ and 10 equally spaced bins $R_b \in \{-5, -4, -3, ..., 3, 4, 5\}$. Moreover, we apply 16 statistic functions to reduce the feature dimension and generate the AUs statistic feature $f_{au\_static} \in \mathbb{R}^{17\times16=272}$ for each topic.

Eyesight features can show whether the subject is in hostility or dull state. We calculate 7 characteristics such as Mean, Variance, Covariance etc. for left and right eyes separately. Thus, there are $f_{eyesight} \in \mathbb{R}^{17\times1=17}$ statistic features for each topic.

Inspired by YMRS, many BD patients have higher emotions and they are infuriated easily. We extract emotion features from Face++ toolkit [16]. Face++ toolkit is a platform to detect and analyze the detected faces, including the analysis result of confidence scores for seven kinds of emotion: anger, disgust, fear, happiness, neutral, sadness and surprise. Using Face++ toolkit, we estimate the emotion in each frame. We extract seven frequency histograms of all emotions and four presences of four emotions, such as sadness, anger, happiness and surprise, to the visual feature vector for each topic. However, it is hard to measure the changes of emotions. Therefore, we obtain the values of valence and arousal respectively in each frame according to the rules for the map of seven emotion mentioned above to Valence-Arousal coordinate space [21]. Besides adding 13 statistic functions of valence values and arousal values respectively, we add 24 statistic functions of Euclidean distance of the consecutive frame in Valence-Arousal coordinate space to the visual feature vector, which can reflect the intensity of emotion's changes. Thus, we generate emotion feature $f_{emotion} \in \mathbb{R}^{11+13\times2+24=61}$ for each topic.

Mentioned by YMRS, limb and head movements when answering questions are also important features indicating the increasing energy and movements on patients. For head movements, we use the same method as eyesight. For limb movement, we count the significant pixel variations between frames, which reflects the patient's range of limb movement. This part provides body movement feature $f_{body\_movement} \in \mathbb{R}^{7\times1=7}$ as global feature.

Totally, 3607 visual features are extracted for three topics.

*2.2.3* ***Text Features.*** Text-based features have been proved effective in the work of [25, 29]. So we extracted text features as well. Using the suite of Linguistic Analysis Tools (SALAT)[13], text analysis was performed automatically on the transcripts of BD interview. The tools we used include Natural Language Processing Tool (siNLP)[6] and Sentiment Analysis and Cognition Engine (SEANCE)[7].

siNLP[6] extracts 14 linguistic features such as the number of words, sentences, unique words. SEANCE[7] contains eight kinds of sentiment indices. Ting Dang et al. [8] investigated the ANEW[2], EmoLex[19], SenticNet[3] and Lasswell[14] among these indices and showed their effectiveness in emotion prediction and depression recognition. Therefore, we used these four indices to extract word affect features for each topic among each transcript. In total, 786 dimensional text-based features were extracted.

**Table 2: Dimension of each feature category in each interview**

| Feature name | Dimension |
|---|---|
| MFCC-eGeMAPS-original | 248 |
| MFCC-eGeMAPS-3topics | 744 |
| Timing | 10 |
| AUs | 3366 |
| Emotion | 183 |
| Eyesight | 51 |
| Body movement | 7 |
| SiNLP | 42 |
| SenticNet | 90 |
| ANEW | 96 |
| EmoLex | 120 |
| Lasswell | 438 |
| Sum | 5395 |

## 2.3 Features Selection

According to Table 2, totally 5395-dimensional features are selected. The high-dimensional feature vector is too redundant to train a classification model with excellent performance. For avoiding overfitting in the case of few samples and train an effective classification model, it is essential to do the feature selection.

We apply Analysis of Variance (ANOVA)[22] algorithm for feature selection. ANOVA is a collection of statistical models, which is used to analyze the differences among group means in a sample, such as the variation among and between groups. So we use scikit-learn toolkit [15] to calculate the ANOVA F-value between label and feature, and select the effective features by the orders of the importance of ANOVA F-value.

# 3 THE PROPOSED FRAMEWORK

## 3.1 Gradient Boosted Decision Tree

Gradient Boosted Decision Tree [11] (GBDT) algorithm is one of the most mentioned algorithms in recent years, mainly due to its superior performance in various data mining and machine learning competitions. It can be used for classification and regression tasks with automatic feature selection.

GBDT, like the Random Forest algorithm, forms a strong learner by combining weak learners. However, the decision tree used in the GBDT algorithm can only be a regression tree, because each tree of the algorithm learns the residual of the sum of all previous tree conclusions. This residual is a cumulative prediction value that can be obtained (By using the residual of each predicted result and the target value as the target of the next learning).

In order to make the models have better expressiveness for features, we use Scalable and Flexible Grandient Boosting Algorithm [4] for framework optimization. Considering the excellent learning

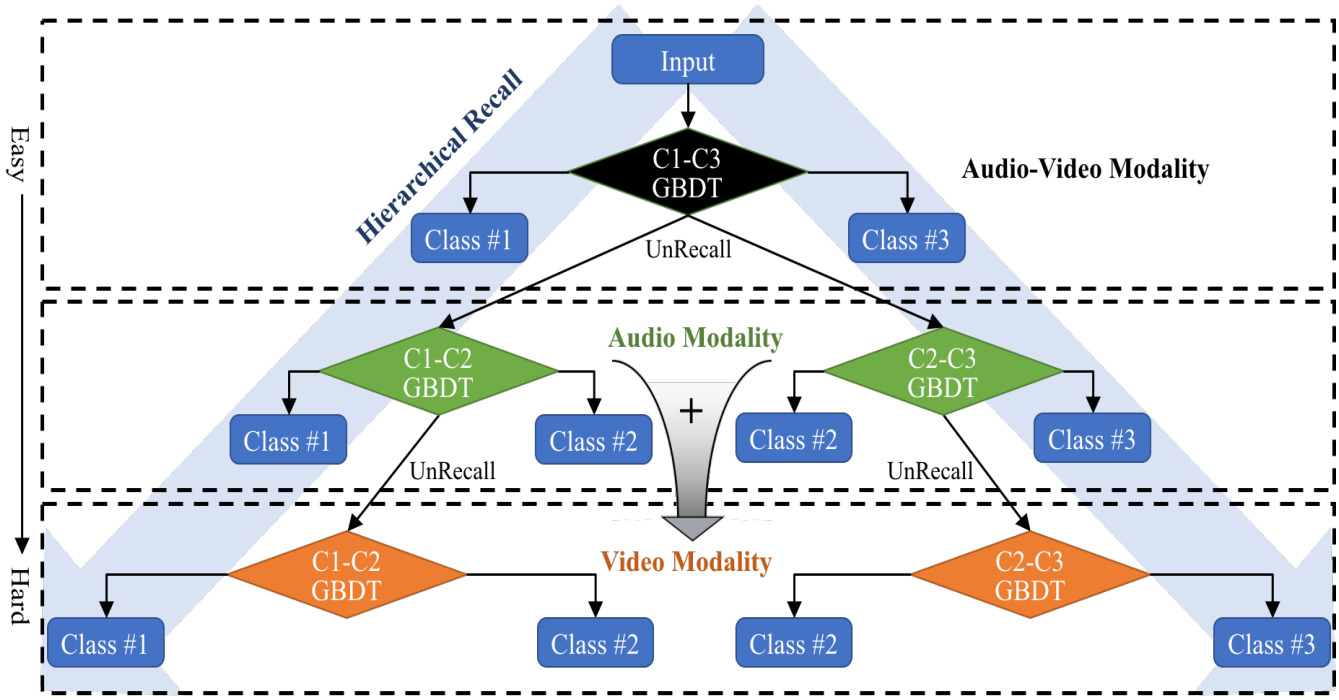


**Figure 1: Multi-modality hierarchical recall framework based on GBDTs. The final structure of hierarchical recall model uses five models, consisting of one audio-video model, two audio models and two video models. The Class #1, Class #2 and Class #3 represent labels of Remission, Hypo-mania, Mania respectively.**

performance and efficient training speed, the eXtreme Gradient Boosting (XGBoost), which is an efficient machine learning function library, is employed to our models.

### 3.2 Multi-modality Hierarchical Recall

The proposed multi-modality hierarchical recall framework is shown in Figure 1. It contains three layers from easy to hard, and hierarchically recall corresponding category. Subject with high confidence level is first recalled while the low one is delivered to the next layer for further judgment. In this way, we can filter out correct results layer-by-layer.

In this framework, we combine three modalities: audio, video and audio-video. The features used in this framework are extracted via topic modeling with the help of text transcriptions. The models of audio-video, audio and video modality are trained by audio-video, audio and video features respectively. All the models above output the probability of two categories, and decide whether they directly output the classification result or deliver the subject into the next layer.

Taking an example to illustrate, we firstly send a subject to C1-C3 GBDT model in the first layer, where the probabilities of Class #1 and Class #3 of this subject are obtained. If the output probability of this model is greater than a predetermined threshold, we specify it directly as the corresponding category. Otherwise, this sample called "unrecall ones", will be sent into the next layer for further judgement. The models in the later layer are similar to the operation of the previous ones. In response to the previous layer output unrecall result for different categories of propensity, we sent it into the corresponding model. For example, in the second layer, those "unrecall ones" from first layer with higher probability of Class #1 than Class #3 are delivered to C1-C2 GBDT model and the other are delivered to C2-C3 GBDT model. The output of the second layer also directly classifies the category or sent it into the last layer. In order to make the framework have better generalization ability, we only use three layers. So the sample that is "unrecalled ones" in the first two layers will get the specific results in the later layers.

Each layer of the above performs a two-category recall of the sample. After a multi-hierarchical recall, we can get a satisfactory classification result. Our framework is a layered and symmetrical multi-modality framework. The experimental results show that our framework has good accuracy and robustness.

## 4 EXPERIMENTS

### 4.1 The AVEC2018 Bipolar Disorder Sub-Challenge Dataset

The AVEC2018 Bipolar Disorder Sub-Challenge Dataset is part of the Turkish Audio-Visual Bipolar Disorder Corpus [5]. It includes audio and visual recordings of structured interviews performed by 47 Turkish speaking subjects aged 18-53. All those Turkish subjects are bipolar disorder patients and were recorded into video camera in every follow up day (0th- 3rd- 7th- 14th- 28th day) during hospitalization and after discharge on the 3rd month. In each interview, they were asked to answer questions in Table 1. In the AVEC2018 Bipolar Disorder Sub-Challenge Dataset, all recordings are split

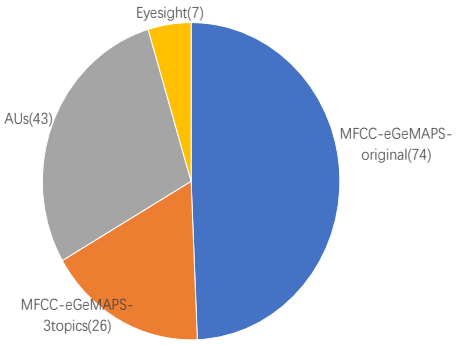


**Figure 2: Distribution of feature categories corresponding to the selected features**

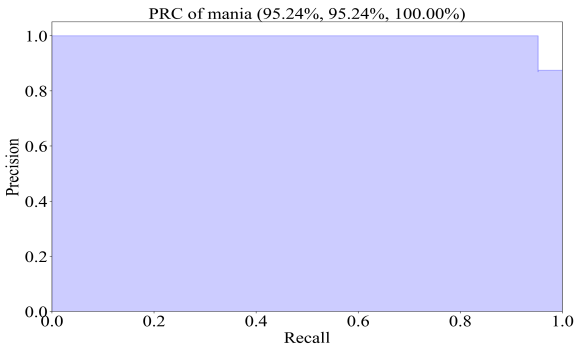


**Figure 3: The Precision-Recall Curve of mania**

into three parts: 104 recordings in training set, 60 recordings in development set and 54 recordings in test set. Besides, AVEC2018 provides the label of Young Mania Rating Scale scores and the ternary value of remission/hypo-mania/mania according to YMRS scores as follow:

1.$Remission : Y_t \leq 7$
2.$Hypo - mania : 7 < Y_t < 20$
3.$Mania : Y_t \geq 20$

## 4.2 Selected Feature Analysis

Although we have tried to extract various features, only some of them will actually work in the end. The selected features are visualized in Figure 2, which contain 100 audio features and 50 video features. Audio statistic features in eGeMAPS and MFCC features are chosen in majority, due to the importance of the spectral and prosodic information. Furthermore, these statistics features both on original and three topics account for a large proportion of selected features, which means that the topic modeling successfully extended the effective features. As for the video features, AUs features play a strong role in this task, since the facial expressions of the patients are different from normal persons. Eyesight features are also useful because the eyes of mania patients tend to be erratic. Moreover, it is worth mentioning that even if the features are not selected, they may still be useful in other models.

## 4.3 Parameters of Multi-modality Hierarchical Recall Model

The final structure of hierarchical recall model uses five models, consisting of one audio-video model, two audio models and two video models. For designing all these models, we have tested several parameters and selected those which obtained the highest UAR on the development set. The parameters of all models used in the hierarchical recall model are shown in table 3. In addition, the threshold is set according to the balance of the precision and recall rates.

## 4.4 Results

As shown in Figure 1, the hierarchical multi-modality fusion model recalls the samples layer-by-layer. Table 4 shows the sample numbers of models recalled per layer. We can see that Layer1 recalls none sample due to its highest threshold, since the function it wants to achieve is grouping. After the first two layers, there are still 16 and 9 samples that have not been recalled in the Dev and Test set respectively. Fortunately, the third-layer model fills the gap well and perfectly realizes multi-modality fusion.

We evaluated each of the five models separately. The results in the development set are shown in Table 6 with the metrics using recall and accuracy rate. Note that the evaluation samples used here removes samples that do not belong to the category of the corresponding model. Obviously, the first C1-C3 model has the best performance, which combines the advantages of audio and video.

In order to better demonstrate the performance of our sub-model, we plotted the Precision Recall Curve (PRC) of the C2-C3 audio model as Figure 3 , where the recall rate is all 100% when the corresponding precision rate is 95%, 90%, 85%. We find that high recall rate is maintained even at high precision rate. This means that there are very few samples that are misclassified, and this model has excellent performance.

As Table 5 shows, our model achieves competitive performance by gradually recalling samples. Note that the baseline system has a lot of performance degradation on the test set. This implies that the overfitting of the baseline system is very serious, which should be avoided. Our result in the dev set is better than any others, and the result in the test set is still clearly better than all baseline, which demonstrates that our model has particularly good generalization performance. At the same time, we also conducted a direct tenary-classification task. The prediction is performed using a C1-C2-C3 model trained by similar features. We can see that although the performance of the tenary-classification model on the Dev set is better than the proposed model, there is still a large performance loss on the test set. This shows that that direct multimodal model fusion is not feasible, and the Multi-modality Hierarchical Recall Model solves the problem that the modal is difficult to fuse through Hierarchical Recall and achieves competitive results.

# 5 CONCLUSIONS

In this paper, we targeted to AVEC 2018 Bipolar Disorder Challenge and conducted multi modalities analysis for BD corpus. In order to perform domain adaptation for each patient and hard sample mining for special patients, we proposed a new hierarchical recall model, where patients of different mania level are recalled at multi

**Table 3: Parameters of Multi-modality Hierarchical Recall Model**

| Layer | Modality | Model | Learning rate | L2 regularization parameter | Threshold |
|---|---|---|---|---|---|
| Layer1 | Audio+Video | C1-C3 | 0.03 | 0.2 | 0.95 |
| Layer2 | Audio | C1-C2 | 0.13 | 0.1 | 0.55 |
| | | C2-C3 | 0.09 | 0.5 | 0.55 |
| Layer3 | Video | C1-C2 | 0.19 | 0.1 | \ |
| | | C2-C3 | 0.01 | 0.1 | \ |

**Table 4: The number of models recalled per layer**

| | Layer | Class1 | Class2 | Class3 | Uncertain |
|---|---|---|---|---|---|
| Dev | Layer1 | 0 | 0 | 0 | 60 |
| | Layer2 | 12 | 16 | 16 | 16 |
| | Layer3 | 17 | 23 | 20 | 0 |
| Test | Layer1 | 0 | 0 | 0 | 54 |
| | Layer2 | 5 | 23 | 17 | 9 |
| | Layer3 | 7 | 27 | 20 | 0 |

**Table 5: Bipolar disorder recognition results**

| | Proposed(%) | Tenary classification model(%) | Baseline(Audio)(%) | Baseline(Visual)(%) | Baseline(Audiovisual)(%) |
|---|---|---|---|---|---|
| Dev | 86.77 | 72.00 | 69.84 | 57.14 | 79.37 |
| Test | 57.41 | 53.70 | 50.00 | 33.33 | 44.44 |

**Table 6: The experimental results of five single-modal models**

| Model | Recall(%) | Accuracy(%) |
|---|---|---|
| Audio+Video C1-C3 | 100.0 | 100.0 |
| Audio C1-C2 | 87.30 | 87.17 |
| Audio C2-C3 | 97.61 | 97.61 |
| Video C1-C2 | 100.0 | 100.0 |
| Video C2-C3 | 76.19 | 76.19 |

layers instead of single layer. The UAR of our proposed model significantly outperforms the baseline method. As far as we know, this is the first time that hierarchical recall model is proposed and applied to MD analysis.

In our future work, deep semantic features will be extracted and fused with hierarchical recall model to improve UAR further.

## REFERENCES


[1] American Psychiatric Association et al. 2013. Diagnostic and Statistical Manual of Mental Disorders, 5th edn (Washington, DC: APA). (2013).

[2] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report. Citeseer.

[3] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining.. In *AAAI fall symposium: commonsense knowledge*, Vol. 10.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[5] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. 2018. The Turkish Audio-Visual Bipolar Disorder Corpus. *Asian Conference on Affective Computing and Intelligent Interaction* (2018).

[6] Scott A Crossley, Laura K Allen, Kristopher Kyle, and Danielle S McNamara. 2014. Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes* 51, 5-6 (2014), 511–534.

[7] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods* 49, 3 (2017), 803–821.

[8] Ting Dang, Brian Stasak, Zhaocheng Huang, Sadari Jayawardena, Mia Atcheson, Munawar Hayat, Phu Le, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. 2017. Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 27–35.

[9] 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. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[10] Fabien Ringeval and Björn Schuller and Michel Valstar and Roddy Cowie and Heysem Kaya and Maximilian Schmitt and Shahin Amiriparian and Nicholas Cummins and Denis Lalanne and Adrien Michaud and Elvan Çiftçi and Hüseyin Güleç and Albert Ali Salah and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*, Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic (Eds.). ACM, Seoul, Korea.

[11] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[12] Yuan Gong and Christian Poellabauer. 2017. Topic Modeling Based Multi-modal Depression Detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 69–76.

[13] K. Kyle. 2017. the suite of Linguistic Analysis Tools (SALAT). http://www.kristopherkyle.com/

[14] Harold D Lasswell and J Zvi Namenwirth. 1969. The Lasswell value dictionary. *New Haven* (1969).

[15] Scikit learn Developers. 2017. sklearn feature selection. http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html#sklearn.feature_selection.f_classif

[16] Ltd. Megvii Technology Co. 2018. Face++. https://www.faceplusplus.com.cn/

[17] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *ACM International Workshop on Audio/visual Emotion Challenge*. 21–30.

[18] Kathleen R Merikangas, Minnie Ames, Lihong Cui, Paul E Stang, T Bedirhan Ustun, Michael Von Korff, and Ronald C Kessler. 2007. The impact of comorbidity

of mental and physical conditions on role disability in the US adult household population. *Archives of general psychiatry* 64, 10 (2007), 1180–1188.

[19] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[20] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).

[21] Georgios Paltoglou and Michael Thelwall. 2013. Seeing Stars of Valence and Arousal in Blog Posts. *IEEE Transactions on Affective Computing* 4, 1 (2013), 116–123.

[22] Fisher Ra. 1968. XV.–The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 20, 4 (1968), 402.

[23] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 3–9.

[24] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[25] Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A Random Forest Regression Method With Selected-Text Feature For Depression Assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 61–68.

[26] Zafi Sherhan Syed, Kirill Sidorov, and David Marshall. 2017. Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 37–43.

[27] Annex Table. 2004. 3: burden of disease in DALYs by cause, sex and mortality stratum in WHO regions, estimates for 2004. *The world health report* (2004).

[28] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

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

[30] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal Measurement of Depression Using Deep Learning Models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 53–59.

[31] Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. 2017. Hybrid Depression Classification and Estimation from Audio Video and Text Information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 45–51.

[32] RC Young, JT Biggs, VE Ziegler, and DA Meyer. 1978. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry* 133, 5 (1978), 429–435.