Multimodal Deep Learning Framework for Mental Disorder Recognition

*Ziheng Zhang[1 5], *Weizhe Lin[2], Mingyu Liu[3] and Marwa Mahmoud[4]
[1 4] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom
[2] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[3] Department of Physics, University of Oxford, Oxford, United Kingdom
[5] Tencent Jarvis Lab, Shenzhen, China
[1]zihengzhang@tencent.com; {[2]wl356, [4]mmam3}@cam.ac.uk; [3]mingyu.liu@queens.ox.ac.uk

Abstract— **Current methods for mental disorder recognition mostly depend on clinical interviews and self-reported scores that can be highly subjective. Building an automatic recognition system can help in early detection of symptoms and providing insights into the biological markers for diagnosis. It is, however, a challenging task as it requires taking into account indicators from different modalities, such as facial expressions, gestures, acoustic features and verbal content. To address this issue, we propose a general-purpose multimodal deep learning framework, in which multiple modalities - including acoustic, visual and textual features - are processed individually with the cross-modality correlation considered. Specifically, a Multimodal Deep Denoising Autoencoder (multi-DDAE) is designed to obtain multimodal representations of audio-visual features followed by the Fisher Vector encoding which produces session-level descriptors. For textual modality, a Paragraph Vector (PV) is proposed to embed the transcripts of interview sessions into document representations capturing cues related to mental disorders. Following an early fusion strategy, both audio-visual and textual features are then fused prior to feeding them to a Multitask Deep Neural Network (DNN) as the final classifier. Our framework is evaluated on the automatic detection of two mental disorders: bipolar disorder (BD) and depression, using two datasets: Bipolar Disorder Corpus (BDC) and the Extended Distress Analysis Interview Corpus (E-DAIC), respectively. Our experimental evaluation results showed comparable performance to the state-of-the-art in BD and depression detection, thus demonstrating the effective multimodal representation learning and the capability to generalise across different mental disorders.**

I. INTRODUCTION

Mental disorders are highly prevalent worldwide. More than 300 million people were estimated by WHO to suffer from mental disorders in 2017 [1]. Not only these disorders can continuously impair an individual's well-being and ability to work but also some of them are associated with a significant mortality risk [1]. Although some psycho-therapeutic options are promising in relapse prevention, only a small proportion of individuals in need receive treatment because of limited medical resources and treatment refractoriness [2]. Automatically recognizing signs of mental disorder can help in early detection of relapses and reduce the treatment resistance [3], [4]. Moreover, such systems can help as a tool to assist psychologists during the face-to-face interview sessions and can even be deployed to mobile devices to facilitate public access to mental healthcare.

It has been shown in the literature [5] that depressed patients differ from normal and psychiatric comparison groups in terms of gross motor activity, body movements, speech, and motor reaction time. These psychomotor symptoms have high discriminative validity to distinguish depression subtypes, and they are also easier to obtain when compared to electroencephalogram (EEG) signals that require complicated systems. During clinical interviews, it is common practice for experts to capture these behavior signals including acoustic, visual, and textual modalities to reach the final diagnosis. While one single modality rarely provides complete information, *complementarity* [6] of the symptom, each modality brings some added value - known as *diversity* [6] - that cannot be obtained from any of other modalities. Therefore, it is significantly important to take advantage of different modalities and process them accordingly to obtain the final decision.

In this work, we propose a multimodal deep learning framework to automatically detect mental disorders symptoms or severity levels. Within the framework, different learning architectures are designed for different modalities. For audio-visual modalities, we present a Multimodal Deep Denoising Autoencoder (multi-DDAE) to learn the shared, frame-level representations of multiple audio-visual inputs, such as MFCC, eGeMAPS [7], facial landmarks, eye gaze, head pose, and facial action units. To generate session-level descriptors, we make use of Fisher Vector (FV) encoding to produce the Fisher Vectors for each interview session. The Paragraph Vector (PV) models [8] are utilized to encode transcripts into document embeddings for the textual modality. An early fusion strategy is applied to fuse features from both audio-visual and textual modalities before feeding the multimodal features to a Multitask Deep Neural Network (DNN), which addresses overfitting, on the final classification task.

The contributions of our work are summarized as follows:

1) We propose a general-purpose multimodal fusion framework to automatically estimate mental disorder scores, which fuses the audio-visual-textual features with different encoding methods and at different levels to retain the maximal amount of cues related to mental disorders.

*Both authors contributed equally to this work

2) We demonstrate that the proposed framework can be generalized across different prediction tasks related to mental disorder analysis. Our experimental evaluation shows that our framework generates comparable results with previous work in two different mental corpora without being specifically modified.

3) We also perform feature ranking in an ablation study to show that our framework has the potential to discover salient biological markers. These results could assist psychologists in mental disorder diagnosis.

II. Related Work

Previous research has demonstrated the effectiveness of deep architectures on multimodal data [9], [10], [11]. For audio-visual speech recognition, Ngiam *et al.* [9] presented bimodal deep autoencoders to capture the correlations across different modalities. For emotion recognition, Kim *et al* [10] and Ranganathan *et al.* [11] proposed multimodal Deep Belief Network (DBN) models, in which first-order representations on different modalities were fused to one shared hidden layer, and they reported the increased classification performance in comparison with unimodal baselines.

The mental disorder recognition differs from emotion recognition as it needs to capture the dynamic aspects of emotions and the temporal information from a variable-length period [12]. Many frameworks were proposed in AVEC2017 and AVEC2018 challenges [13], [14] to tackle this problem. Yang *et al.* [15] proposed several histogram-based features of arousal, speaking rate, and hands distance, and these features were fed into tree-based classifiers after dimension reduction, which might fail to capture the rapid changes in multimodal features. Syed *et al.* [16] proposed "turbulence features" to capture the sudden, erratic changes in feature trajectories. They also applied Fisher Vector (FV) encoding of one modality in feature aggregation step. Du *et al.* [17] presented IncepLSTM on a single modality to encode audio temporal representations on multiple scales [18], and with an improved triplet loss function, their framework was shown to capture dynamic information and to obtain a high-level descriptor for the whole audio clip. However, their framework only involved acoustic modality, and a performance gap was observed when compared with other multimodal frameworks. Dibeklioğlu *et al.* [19] presented Stacked Denoising Autoencoders (SDAE) to learn the non-linear mapping of facial landmarks and head pose respectively on frame-level, and the late fusion (or decision fusion) of all the modalities showed higher accuracy than unimodal models. Since strong correlations have been found between interview contents and depression symptoms [20], analyzing emotion-related textual modality has emerged as a new approach in mental disorder detection. In the work of Xing *et al.* [21] all available modalities were utilized in their framework, including audio, video, and transcribed text. They also introduced a hierarchical recall model for the classifications. Each layer within the model contained a Gradient Boosted Decision Tree (GDBTs) using different subsets of all features, and following a boosting strategy, the only un-recalled samples would be transmitted to the next layer.

III. Multimodal Learning Framework

To address the discrepancy and granularity between audio-visual-textual modalities, they are divided into two subgroups, audio-visual modality that is processed on frame-level, and textual modality that is processed on session-level. Fig. 1 illustrates our multimodal framework, in which audio-visual features are encoded via a Multimodal Deep Denoising Autoencoder (multi-DDAE) on frame-level and then transformed to fixed-length Fisher Vectors (FVs) on session-level. Textual features are obtained as transcripts and then embedded into fixed-length vectors with a Paragraph Vector (PV or doc2vec) model.

A. Audio-Visual modalities

For the audio-visual modalities, we propose a 3-layer Multimodal Deep Denoising Autoencoder (multi-DDAE) (i.e., DDAEs with 3 hidden layers) that learns shared and robust representations upon several modalities by reconstructing the denoised target from the noisy input. Different visual features, like facial landmarks and action units, are considered as different modalities because of the different ranges and expert-knowledge involved. The audio-visual features are typically Low-Level Descriptors (LLDs) that, for instance, can be MFCC, eGeMAPS [7], facial landmarks, facial landmarks, eye gaze, head pose, action units (AUs) or other representations learned from deep learning architectures. The number of input features can be extended to any number larger than one. Before feeding multimodal features into multi-DDAE, features must be aligned on frame-level first to ensure they are extracted from the same time interval. Some audio-visual features could be missing in some time frames, which are zeroed as the missing modalities in order to match the dimension.

As Fig. 1 (1) shows, multiple encoders are merged into one shared layer after one hidden layer, which ensures multimodal features are fused at "middle-level" as the first-order representations [9] as it would be difficult to correlate raw data of different modalities. From the shared layer, multimodal features are reconstructed via their decoders with the weighted sum of mean squared errors on all reconstructed inputs as the loss function. Since the denoising criterion in autoencoders helps to learn robust features, masking noise that corrupts a specific portion of the inputs to zero is implemented prior to the per-modality normalization. The multi-DDAE is compiled with the Adam optimizer with 0.001 learning rate and the weight setting within the loss function varies across different corpora. Two hyperparameters are investigated in multi-DDAE as listed in Table I: noise level (a) and hidden ratio (h), which is defined as the dimension ratio between two consecutive hidden layers. For instance, with hidden ratio 0.5, the dimensions of hidden layers would be $\{0.5d, 0.25d, 0.5d\}$ where d represents the input dimension. As the unsupervised feature learning, the

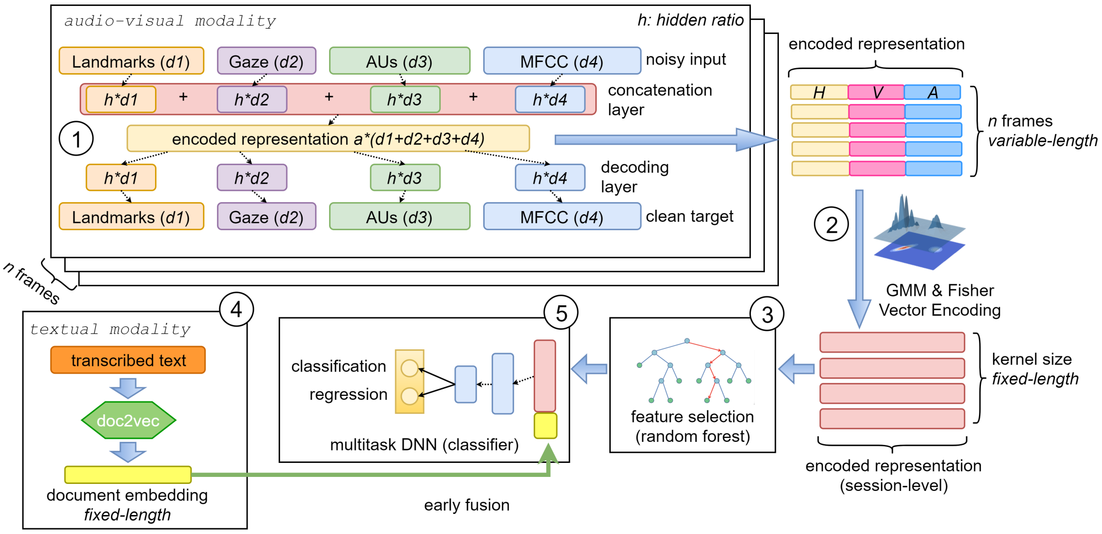

Fig. 1. The pipeline of the proposed framework: (1) audio-visual features are encoded by the multi-DDAE to generate compact per-frame representations; (2) these representations along with computed dynamics are encoded into fixed-length per-session descriptors using GMM and Fisher Vector encoding; (3) most discriminative features are selected via random forests; (4) session-level document embeddings are produced by the PV/doc2vec model; (5) early-fused multimodal features are fed into the multitask DNN for the classification task.

multi-DDAE is trained on all available unlabelled audio-visual data.

The per-frame representations learned in multi-DDAE can be regarded as static because they solely encode the static element of the input, such as locations of facial landmarks or pitch of audio signal without any temporal information. Similar as the work in [19], we extend these representations with the per-frame dynamics. Considering the latent representations as a matrix $H \in \mathcal{R}^{n \times d}$, in which n denotes the number of frames and d the final dimension of representations. Each column in $H_i (i \in \{1, 2...d\})$ corresponds to one node in the representation layer. We then compute the first-order dynamics, velocity V, of H by the 1^{st} derivative $V_i = \frac{dH_i}{dt}$, measuring the velocity of the change between per-frame representations. We continue to calculate the second-order dynamics, acceleration A, of H by the 2^{nd} derivative $A_i = \frac{d^2 H_i}{d^2 t}$, measuring the acceleration of the change. To align H, V, and A, the first two frames are discarded in video sessions, and three features H, V, and A are concatenated as frame-level representations, as seen in Fig. 1 (2).

Because the video sessions vary in length, we encode frame-level representations with Fisher encoding to produce Fisher Vectors (FVs), fixed-length descriptors on session-level, by fitting them into a Gaussian Mixture Model (GMM) [22]. Specifically, a GMM is firstly built with a specified number of kernels to estimate the probability distribution of multiple multivariate Gaussian distributions on the time-series frame-level representations. The investigated values for the number of GMM kernels (g) are shown in Table I. FVs are calculated afterwards with the estimated distributions along with their mean and variance. We also implement the power normalization and l_2 normalization to generate the Improved Fisher Vectors (IFVs) [22] as the session-level descriptors in the framework.

To reduce redundancy and select the most informative feature set, we apply a tree-based model (Random Forest) to select salient features in our framework ((3) in Fig. 1). The selection process is based on the feature importance by computing the information gain $Gain(S, A)$:

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

where $v(A)$ is the set of all possible values for feature A relative to a dataset S, and S_v is the subset of S for which feature A has value v. As Random Forest is robust to redundant features and insensitive to irrelevant information, the feature importance usually leads to a reliable and discriminative subset of features.

TABLE I
HYPERPARAMETER SETTINGS IN THE FRAMEWORK

Model	Hyperparameter	Investigated values
multi-DDAE	Hidden ratio (h)	{0.4, 0.5}
	Noise level (a)	{0.1, 0.2, 0.4}
FV	GMM Kernel (g)	{16, 32}
PV	Architecture	{PV-DM, PV-DBOW}
	Vector size (v)	{25, 50, 100}
	Window size (w)	{5, 10}
	Negative words (n)	{5, 10}

B. Textual modality

To incorporate textual modality into our framework, the transcripts of recordings (psychotherapy interviews) are firstly obtained via open-source API, such as Google Cloud Platform (GCP). As shown in Fig. 1 (4), Paragraph Vector (PV) models, or doc2vec models [8], are utilised to embed the variable-length transcripts into fixed-length, session-level representations. Generally, there are two architectures within doc2vec, PV with Distributed Memory (PV-DM) and PV with Distributed Bag-of-Words (PV-DBOW) [8]. Two architectures are experimented with different hyperparameter settings as Table I. To boost the performance of doc2vec models, the model is pre-trained on an additional corpus, which helps to learn mental implication.

C. Multimodal fusion

We follow the early fusion strategy to concatenate the compact-size vectors from audio-visual modalities and textual modality to obtain a joint multimodal representation. As the limited size is a typical issue for most mental disorder corpora, a Multitask Deep Neural Network (DNN) with hard parameter setting is built for the classification task to overcome overfitting, as shown in Fig. 1 (5). The proposed Multitask DNN has the shared hidden layers between two tasks with unchanged task-specific output layers where one task is the classification of disorder levels while the other task could be the regression of disorder scores [23] or other auxiliary tasks. The joint loss function is adjusted as the weighted sum of cross-entropy loss for classification and the Euclidean loss for regression, as shown in Eq. 2.

$$\begin{aligned}\mathcal{L} &= \mathcal{W}_c\mathcal{L}_c + \mathcal{W}_r\mathcal{L}_r \\ &= \mathcal{W}_c(-\sum_{c=1}^{N} y_c \log(p_c)) + \mathcal{W}_r(\frac{1}{M}\sum_{i=1}^{M} \parallel y_r - p_r \parallel^2)\end{aligned} \qquad (2)$$

where the weights for two losses are denoted as $\mathcal{W}_c$ and $\mathcal{W}_r$, ground-truth values as y_c and y_r, predicted values as p_c and p_r, the number of classes as N and the number of samples as M.

IV. BIPOLAR DISORDER RECOGNITION

We first evaluate our framework on the Bipolar Disorder Corpus (BDC) [4]. We also compare our framework with the models presented at AVEC2018 [14].

A. Dataset

The BDC was introduced by Çiftçi *et al.* [4] to provide an insight for the personalized treatment of BD patients. The BDC is annotated by psychiatrists with BD states as well as the Young Mania Rating Scale (YMRS) scores which were obtained at session level such that each score corresponds to one patient on one of the pre-determined test days. The data format in the corpus is a set of audio-visual recordings of structured interviews performed by 46 Turkish speaking patients. For the classification experiments, the BDC contains 104 recordings as the training partition, 60 recordings as the development partition, and 54 recordings as the test partition. The provided ground-truth labels are clinician-annotated YMRS scores of corresponding sessions, and the recordings are also grouped into three disjoint subgroups as follows, thus leading to a ternary classification task: 1) Depression: YMRS ≤ 7; 2) Hypo-mania: $7 <$ YMRS < 20; 3) Mania: YMRS ≥ 20.

The availability of the BDC is upon request but labels for the test set cannot be obtained outside the challenge. Therefore, in the following experiments, we evaluated our framework on the pre-determined training and development sets and compared our framework with the competing frameworks [15], [17], [21], [16] in AVEC2018.

B. Multimodal Features and Preprocessing

The baseline features (LLDs) used in AVEC2018 were extracted with open-source toolkits, such as acoustic features extracted via OpenSMILE[1], visual features extracted via OpenFace [24]. The baseline features include MFCC and eGeMAPS [7] for acoustic modality, and facial landmarks, eye gaze, head pose, and action units for visual modality. MFCC features are computed at the frame level while eGeMAPS features are computed at speaker turn level that however can be aligned with other modalities unless given the frame time. Since MFCC and eGeMAPS have overlapping elements, the shared representation layer in our multi-DDAE was trained from only one acoustic feature, either MFCC features or eGeMAPS features.

To align multimodal features, 3 contiguous acoustic features were concatenated as each input that has approximately the same duration as 1 visual feature. Thus, a total of five modalities were used in one multi-DDAE for the BD recognition, which included facial landmarks, eye gaze, head pose, action units and acoustic features (MFCC or eGeMAPS).

The PV models for the textual modality were pre-trained on an additional Turkish corpus, triwiki[2], which contained various kinds of texts in Turkish, such as articles and primary meta-pages.

C. Experimental Results

The multi-DDAE and the PV were evaluated on the classification of BD symptoms with the same multitask DNN classifier in the framework.

For evaluation, we used the unweighted F1-score which ignores the imbalance between the three classes. As seen in Table II, in audio-visual modalities, multi-DDAE using MFCC features achieves a performance of 0.656 UAR and 0.667 F1, slightly better than multi-DDAE using eGeMAPS features. Furthermore, the multi-DDAE using MFCC tends to perform better with a more compact size of the latent representation while the multi-DDAE using eGeMAPS favors a less compact size. In textual modality, the experimental results show better performance in PV-DBOW in comparison with PV-DM (0.013 higher in UAR and 0.038 higher in F1).

With the best-performing multi-DDAEs and doc2vec models, we evaluate 4 multimodal fusion models as shown in Table II (5-8). The best-performing multimodal framework is based on multi-DDAE using MFCC and PV-DM and it obtains a UAR of 0.709 and F1 of 0.721, significantly better than the unimodal architectures.

Additionally, to validate the generalization of the proposed framework, 10-fold cross-validation (CV) was implemented on the entire dataset (training set + dev set) and the best-performing framework ((5) in Table II) also reaches the highest averaged UAR at approximately 0.60. This shows

[1] https://www.audeering.com/opensmile/

[2] https://dumps.wikimedia.org/trwiki/

TABLE II

Performance in audio-visual modalities, textual modality and multimodal fusion on BD classification. Best performance in **bold**. Metrics include Unweighted F1, Unweighted Average Recall (UAR), and Unweighted Average Precision (UAP).

	ID	Hyperparameters				Metrics		
Audio-Visual		Acoustic	h	a	g	UAR ↑	UAP ↑	F1* ↑
	(1)	MFCC	0.4	0.1	32	0.656	0.678	0.667
	(2)	eGeMAPS	0.5	0.1	32	0.622	0.665	0.642
Textual		Model	v	w	n	UAR ↑	UAP ↑	F1* ↑
	(3)	PV-DM	50	10	5	0.492	0.481	0.486
	(4)	PV-DBOW	50	-	5	0.505	0.544	0.524
Multimodal Fusion		Audio-Visual		Textual		UAR ↑	UAR (CV) ↑	F1* ↑
	(5)	(1) MFCC		(3) PV-DM		**0.709**	**0.598**	**0.721**
	(6)	(1) MFCC		(4) PV-DBOW		0.667	0.572	0.673
	(7)	(2) eGeMAPS		(3) PV-DM		0.675	0.543	0.691
	(8)	(2) eGeMAPS		(4) PV-DBOW		0.659	0.581	0.665

that our framework is not overfitted to the pre-determined development set and generalizes well on unseen data.

TABLE III

Comparison of our framework with AVEC2018 frameworks

Framework	UAR (dev.) ↑	Acc. (dev.) ↑
Yang *et al.* 2018 [15]	0.714	**0.717**
Du *et al.* 2018 [17]	0.651	0.650
Xing *et al.* 2018 [21]	**0.868**	NA
Syed *et al.* 2018 [16]	0.635	NA
Ours	0.709	**0.717**

Table III lists the experimental results obtained by four frameworks in AVEC2018 on the development set. It is clear that our framework, (5) in Table II, outperforms the frameworks proposed by [17] and [16] in both UAR and accuracy. Furthermore, our framework achieves the same accuracy as [15] and a close UAR (only 0.005 lower) even though [15] benefited from extra data as they extracted the ”arousal features” from pre-trained LSTM-RNN model on AVEC2015 affective dataset. [21] seems to have a better performance than ours, but [21] suffers overfitting issues due to the performance drop, 0.868 UAR on the development but only 0.574 UAR on the test set. Therefore, our proposed multimodal framework shows comparable performance to the state-of-the-art on BD recognition.

V. Depression Detection

To further demonstrate the generalisability, we experiment our framework with the depression detection task on the Extended Distress Analysis Interview Corpus (E-DAIC) [25], which is used in AVEC2019 challenge [26]. Note that we only compare our experimental results with the AVEC2019 baseline system.

A. Dataset

E-DAIC is the extended version of WOZ-DAIC which contains semi-clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety and depression [25]. The dataset excluding the un-released test set is partitioned into a training set of 163 samples and a development set of 56 samples while the overall diversity of the speakers - in terms of age, gender distribution, and the eight-item Patient Health Questionnaire (PHQ-8) scores - is preserved. Our analysis was based on the training and development sets following the restrictions of AVEC2019 [26]. Because PHQ-8 values and depression binary labels are reported for each participant, the challenge task could be thus considered as a combination of a regression and a binary classification.

B. Multimodal Features and Preprocessing

Since the multi-DDAE with MFCC performed the best in BD classification, we defined the acoustic features as MFCC and therefore, MFCC features, FAUs (including pose, gaze, and action units), and ResNet deep spectrum features were chosen as the fundamental frame-level features in the multi-DDAE. To avoid biasing the multi-DDAE by feeding modalities of unbalanced dimensions or incompatible magnitudes, PCA was performed on ResNet features to reduce the dimension from 2048 to 200 and z-norm was applied for each feature. We investigated different settings for the number of features selected in Random Forests as a considerable impact was found on the performance.

Finally, MFCC, ResNet features, Pose, Gaze, and Action Units were used for training and 3 contiguous acoustic features were concatenated as each input to match the duration time for 1 visual feature. We also ran the session-level analysis with textual information but the PV models were pre-trained on the transcripts only without additional English corpora.

C. Experimental Results

Following the AVEC2019 baseline [26], Concordance Correlation Coefficient (CCC) [27] was used to evaluate the regression task, and we additionally reported mean-squared-error (MSE) for regression and unweighted F1-score and accuracy for classification.

The multi-DDAE and the PV are evaluated using our proposed multitask DNN classifier in the framework. The multi-DDAE is validated to perform the best with hyperparameters of 0.5 hidden ratio, 0.1 corruption noise level, and the FV encoding with 32 GMM kernels.

As seen in Table IV, the baseline system that fuses MFCC, ResNet, and FAUs features shows a CCC value of 0.336. For

TABLE IV

Proposed framework performance in audio-visual modalities, textual modality and multimodal fusion on depression detection. The best performance for every column is given in **bold** and unweighted F1-score is used.

		Regression		PLSR	Classification	
	fused dimension	CCC ↑	MSE ↓	CCC ↑	F1* ↑	Acc. ↑
Audio-Visual	500	0.464	26.75		0.774	0.857
	700	0.452	25.73		0.838	0.857
	1000	0.386	29.38		0.884	**0.893**
Textual	50 (PV-DM)	**0.560**	21.72		0.907	0.839
Multimodal Fusion	500	0.506	20.85		0.894	0.821
	700	0.528	**20.06**		**0.917**	0.857
	1000	0.504	20.67		0.896	0.821
10-fold CV	500	0.419	26.47	0.352	0.845	0.785
	700	0.423	25.85	0.382	0.884	0.817
	1000	0.385	28.85	0.360	0.868	0.785
Baseline (AVEC2019)	Late fusion	0.336	-	-	-	-

audio-visual modalities, the multi-DDAE reports a maximum CCC value of 0.464 on the development set with 500 features selected, better than the fusion of all baseline features. The framework with 1000 features selected, however, shows a better performance than the one with 500 features in terms of classification. In textual modality, the PV-DM model achieves a performance of 0.560 CCC and 0.907 F1, showing a significant gain from the baseline.

Other than for the BD classification, we noticed the performance gap between different numbers of selected features shown in Table IV. Specifically, with more features selected in the audio-visual modalities, the proposed framework achieved better classification but unstable performance in regression. As shown in Table IV, an increasing number of selected audio-visual features (from 500 to 700) benefits the final result, but however, with 1000 features, the performance is suppressed. As we adopted an early-fusion strategy where selected audio-visual features are merged with textual features, we believe that when selected audio-visual features are above 700, the redundant information biased the classifier and deteriorated results. Therefore, we conclude that the audio-visual features have less discriminative features than textual features in depression detection.

The performance improvement of multimodal fusion over the multi-DDAE is observed (0.076 higher in CCC and 0.079 higher in F1) when 700 features are selected. However, when compared with PV-DM on textual modality, the multimodal fusion shows a better classification result (0.01 higher in F1) and a worse regression result (0.032 lower in CCC).

The 10-fold Cross Validation (CV) demonstrates that our framework is not overfitted to the development set with an averaged CCC of 0.423 and F1 of 0.884. Furthermore, we implemented Partial Least Square Regression (PLSR) in 10-fold CV to validate the effectiveness of the proposed multitask DNN in the regression. As seen in Table IV, while the multitask DNN obtains a CCC value of 0.423, the PLSR obtains a CCC of 0.263 with the same multimodal features, proving the superiority of our proposed multitask configuration.

D. Ablation Study

To gain some insights into salient biological markers, we conducted an ablation experiment to investigate the importance of individual audio-visual features in our framework. It can also be considered as an audio-visual feature ranking. We investigated every individual feature by excluding it from our multi-DDAE and then proceeding with the rest of 10-fold CV experiments. In other words, different feature combinations were fed into the multi-DDAE, and the investigated feature was the one missing from the combination.

TABLE V

Performance comparison of different feature combinations. (*) represents the original setting as the baseline

Feature Combination	CCC change (multitask DNN)	CCC change (PLSR)
MFCC + ResNet + FAUs (*)	0.419	0.352
MFCC + ResNet	**0.167** ↓	**0.147** ↓
ResNet + FAUs	0.137 ↓	0.106 ↓
MFCC + FAUs	0.040 ↓	0.054 ↓

Table V shows performance changes of three feature combinations compared with the original setting, which indicates the audio-visual feature ranking as: FAUs > MFCC > ResNet. Based on the findings that textual modality itself gives high performance (0.560 CCC as shown in Table IV), we can reach a conclusion that the depression cues are more likely to be the speech content and patients' facial expression, which could assist psychologists in the depression diagnosis.

VI. Conclusion

This work proposed a multimodal deep learning framework to automatically analyze signs of mental disorders, especially from video, audio and textual data. A Multimodal Deep Noising Autoencoder (multi-DDAE) was used to encode the per-frame representations across multiple audio-visual features and Fisher Vector (FV) encoding was used to encode the compact per-session descriptors. The document embeddings of interview transcripts, inferred by Paragraph Vector (PV) models were incorporated to helped to improve the performance. To handle overfitting, a multitask learning model was proposed. Experimental evaluation showed that our proposed framework achieved comparable performance to previous work in bipolar disorder recognition and baselines in depression detection, showing effective multimodal representation learning. Moreover, without being specially optimized for the learning task, our framework generalizes well across different mental disorder corpora and shows the potential to discover the biological markers with the feature ranking, which helps the diagnosis of mental disorders.

As future work, we evaluated an improved version of our framework in another audio-visual dataset [28]. In the future, We will also investigate the semantic interface between all audio-visual-textual modalities to address the discrepancy and granularity and compare it with the early fusion strategy in the current framework. Furthermore, it would be worthwhile to incorporate the spatial information in our framework with additional layers in the multi-DDAE, such as convolutional and pooling.

References

[1] W. H. Organization *et al.*, "Depression and other common mental disorders: global health estimates," World Health Organization, Tech. Rep., 2017.

[2] A. E. Kazdin and S. L. Blase, "Rebooting psychotherapy research and practice to reduce the burden of mental illness," *Perspectives on psychological science*, vol. 6, no. 1, pp. 21–37, 2011.

[3] I. E. Bauer, J. C. Soares, S. Selek, and T. D. Meyer, "The link between refractoriness and neuroprogression in treatment-resistant bipolar disorder," in *Neuroprogression in Psychiatric Disorders*. Karger Publishers, 2017, vol. 31, pp. 10–26.

[4] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, "The turkish audio-visual bipolar disorder corpus," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 5 2018, pp. 1–6.

[5] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, and E. Haffen, "Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications," *BioMed research international*, vol. 2013, 2013.

[6] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

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

[8] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1188–II–1196.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. USA: Omnipress, 2011, pp. 689–696.

[10] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5 2013, pp. 3687–3691.

[11] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3 2016, pp. 1–9.

[12] R. W. Picard, *Affective computing*, 2000.

[13] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '17. New York, NY, USA: ACM, 2017, pp. 3–9.

[14] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. A. Salah, and M. Pantic, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18. New York, NY, USA: ACM, 2018, pp. 3–13.

[15] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, "Bipolar disorder recognition with histogram features of arousal and body gestures," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18. New York, NY, USA: ACM, 2018, pp. 15–21.

[16] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18. New York, NY, USA: ACM, 2018, pp. 39–45.

[17] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18. New York, NY, USA: ACM, 2018, pp. 23–30.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[19] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 525–536, 2017.

[20] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 12 2016, pp. 136–143.

[21] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality hierarchical recall based on gbdts for bipolar disorder classification," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC'18. New York, NY, USA: ACM, 2018, pp. 31–37.

[22] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 143–156.

[23] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017.

[24] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[25] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[26] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th Annual Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '19. New York, NY, USA: ACM, 2019.

[27] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[28] W. Lin, I. Orton, M. S. Liu, and M. Mahmoud, "Automatic detection of self-adaptors for psychological distress," in *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020)*. IEEE, 2020.