# Cell-Coupled Long Short-Term Memory With $L$-Skip Fusion Mechanism for Mood Disorder Detection Through Elicited Audiovisual Features

Ming-Hsiang Su, *Member, IEEE*, Chung-Hsien Wu, *Senior Member, IEEE*,
Kun-Yi Huang, and Tsung-Hsien Yang

***Abstract*— In early stages, patients with bipolar disorder are often diagnosed as having unipolar depression in mood disorder diagnosis. Because the long-term monitoring is limited by the delayed detection of mood disorder, an accurate and one-time diagnosis is desirable to avoid delay in appropriate treatment due to misdiagnosis. In this paper, an elicitation-based approach is proposed for realizing a one-time diagnosis by using responses elicited from patients by having them watch six emotion-eliciting videos. After watching each video clip, the conversations, including patient facial expressions and speech responses, between the participant and the clinician conducting the interview were recorded. Next, the hierarchical spectral clustering algorithm was employed to adapt the facial expression and speech response features by using the extended Cohn–Kanade and eNTERFACE databases. A denoizing autoencoder was further applied to extract the bottleneck features of the adapted data. Then, the facial and speech bottleneck features were input into support vector machines to obtain speech emotion profiles (EPs) and the modulation spectrum (MS) of the facial action unit sequence for each elicited response. Finally, a cell-coupled long short-term memory (LSTM) network with an *L*-skip fusion mechanism was proposed to model the temporal information of all elicited responses and to loosely fuse the EPs and the MS for conducting mood disorder detection. The experimental results revealed that the cell-coupled LSTM with the *L*-skip fusion mechanism has promising advantages and efficacy for mood disorder detection.**

***Index Terms*— Cell-coupled LSTM, denoizing autoencoder (DAE), *L*-skip multimodal fusion, mood disorder detection.**

## I. Introduction

MOOD disorder is a mental illness, which has a high global prevalence. According to the classification in the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition (DSM-5) [1], bipolar disorder (BD) and unipolar depression (UD) are the two major mood disorders. The mood state of patients with UD only fluctuates within the euthymic and depressed states; on the other hand, the mood state of patients with BD can enter a manic state. Smith and Craddock [2] demonstrated that patients with BD who experience recurrent episodes of depression often have symptoms of mania, and such patients have been diagnosed as having UD in recent medical studies. This implies that a large percentage of patients with BD have been misdiagnosed as having UD but are not clinically diagnosed as such. The surveys reported in [3] suggest that patients with BD are often misdiagnosed on initial presentation and are mostly misdiagnosed as having UD. To avoid delay in appropriate treatment due to misdiagnosis, a mood disorder detection assistance system is highly desirable to help doctors while conducting one-time diagnoses. Numerous studies on mood disorder have focused on single mental illnesses such as UD or BD. In these studies, patients with UD or BD were distinguished from healthy controls in different fields [4]–[9]. However, only a few studies have considered the simultaneous detection of patients with UD and BD and healthy controls. In one study, depression identification was attempted using vocal utterances, facial expressions, and body movements [10]. In AVEC 2013 [11], audiovisual signals have been proven to be useful for emotion, depression, and mood disorder detection and, thus, are considered in this paper for mood disorder classification.

For various multimedia analysis tasks, multimodal fusion is used to leverage the complementary information among different data modalities to disclose the relations that cannot be found in unimodal data [12]. Many studies have obtained successful results on audiovisual fusion for numerous applications. However, considerable scope exists for exploring new techniques, and many challenges are presented in stream weighting and asynchrony problems [13]. Katsaggelos *et al.* [13] reviewed a recent study on audiovisual fusion and confirmed that deep learning can markedly improve audiovisual fusion performance, as it has been successfully applied in many other areas. With the capabilities of deep learning, data representation can be learned in an unsupervised manner without the requirement of hand-crafted labeling. In terms of multimodal fusion and deep learning, in this paper, we aimed to construct a diagnosis assistance system that applies deep learning technologies to combine speech and facial expression signals for conducting a one-time evaluation for patients and providing a reference criterion for doctors

Manuscript received May 7, 2018; revised November 15, 2018 and January 4, 2019; accepted February 12, 2019. Date of publication March 18, 2019; date of current version January 3, 2020. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract 107-2218-E-006-008. *(Corresponding author: Chung-Hsien Wu.)*

M.-H. Su, C.-H. Wu, and K.-Y. Huang are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan (e-mail: huntfox.su@gmail.com; chunghsienwu@gmail.com; iamkyh77@gmail.com).

T.-H. Yang is with Telecommunication Laboratories Chunghwa Telecom Co., Ltd., Taoyuan 326, Taiwan (e-mail: yasamyang@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2899884

to reduce misdiagnoses. Crucially, the proposed cell-coupled LSTM with an $L$-skip fusion strategy for the cells in the LSTM can precisely model the temporal and mutual relationship between the elicited audio and visual data, thus improving the accuracy of mood disorder detection.

In this paper, we presented two problems that are encountered during automatic mood disorder detection. First, Perlis [3] reported that patients with BD are often misdiagnosed on initial presentation and are mostly misdiagnosed with UD. To avoid delay in appropriate treatment due to misdiagnosis, a mood disorder detection assistance system to help doctors while conducting a one-time diagnosis is highly desirable. Second, few or no databases are available, which comprise speech signals and facial expressions for mood disorder detection. The main contributions of this paper are summarized as follows. First, we collected audiovisual data comprising elicited speech responses and facial expressions obtained after watching six emotion-eliciting videos. Few or no such databases are available for experimentation. Second, the hierarchical spectral clustering (HSC) and a denoizing autoencoder (DAE) method [14], [15] were adopted to deal with the problems of domain adaptation as well as bottleneck feature extraction. Third, we proposed a cell-coupled LSTM with an $L$-skip fusion mechanism for mood disorder detection by considering the changes in speech responses and facial expressions for different emotional stimuli. Although the proposed method is a heuristic modification of existing methods, it is still innovative in terms of the data characteristics. Fourth, the proposed elicitation-based mechanism enables a doctor to evaluate a patient with mood disorder at a reduced misdiagnosis rate during a one-time diagnosis. To the best of our knowledge, no similar system is currently available to assist doctors in evaluating a patient with a mood disorder.

The remainder of this paper is organized as follows. Section II presents the related research conducted on mood disorder recognition. Section III describes the preparation process of the mood disorder database. Section IV introduces the domain adaptation on the database. Section V describes the detailed system architecture, and Section VI presents the experiments and discussions. Finally, Section VII summarizes this paper.

## II. Related Work

In related studies, researchers have used speech and facial expressions' information to identify mood disorders, and their study results have provided considerable assistance in clinical diagnosis. In recent years, researchers have invested substantial effort in the field of UD detection. Patients with depression can be distinguished from the control group by using these approaches while conducting short-term detection [7], [16]–[19]. However, few studies have investigated the difference between BD and UD. Most of the studies have paid considerable attention to the long-term monitoring of the variations in the mood state of patients with mania, euthymia, and depression [5], [20], [21]. In short-term detection, several researchers have conducted their proposed procedures on different data types. Greco *et al.* [4] conducted a slideshow of pictures with high arousal and negative valence in the international affective picture system to provide emotional stimuli. They also analyzed the electrodermal response of patients with BD to identify their mental states. Lanata *et al.* [22] investigated whether features extracted from electrodermal responses can be beneficial for the development of a feasible and effective clinical decision support system for BD management. Bersani *et al.* [23] used videos to elicit the emotional state of the patients with BD and collected their facial expressions. These expressions were compared with those from patients with BD and schizophrenia. Bersani *et al.* [23] found that the patients with BD might elicit unpredictable emotion responses when they are presented with emotion stimuli.

Studies have indicated that patients with BD exhibit lower emotion perception accuracy compared with healthy people or patients with other mood disorders [24], [25]. Some studies have shown that the emotional responses of patients with UD and BD, elicited by emotional stimuli [26], [27], could differ from those of healthy people. These studies revealed that patients with both the BD and UD have emotion-processing defects, such as emotional sensitivity and emotional perception. Vederman *et al.* [25] suggested that BD might have a characteristic trait of disruption in the identification of negative emotions such as sadness and fear. As presented in these studies, patients with BD and UD were more sensitive to emotion, were less responsive to emotional stimuli, and had low accuracy in emotion perception. These investigations were helpful for constructing a diagnosis aid system for one-time evaluations based on elicited emotions and facial expressions. In general, the feelings and emotions of a patient can be outwardly revealed by their facial expressions and speech. These two modalities are helpful for determining a patient's mood. Thus, this paper focuses on analyzing variations in the speech responses and facial expressions of patients for conducting mood disorder detection.

In related studies, the data fusion strategy has been widely used to integrate facial and vocal cues to improve detection performance [28]. The three approaches were proposed for multiple-modality fusion—feature-level fusion (FF), decision-level fusion (DF), and model-level fusion (MF) [29]. FF uses the correlation between multiple features obtained from different modalities at an early stage by synchronizing multiple modalities. However, there are several drawbacks in FF as follows: 1) the components in the feature vectors that integrate different modalities are uncorrelated; 2) the components have different temporal structures and metric scales; and 3) the approach increases the dimensionality of the feature vector, which might cause dimensionality problems. Compared with FF, it is easier to combine the results from all unimodal models in DF [30]. DF is more flexible than FF because it enables the use of the most suitable models for analyzing each piece of unimodal data, such as LSTM for temporal structure data and convolutional neural network (CNN) for image data. The drawbacks of DF are that the correlation among modalities at an early stage cannot be utilized and that the learning process is tedious because each unimodal model must be trained separately. Moreover, to exploit the

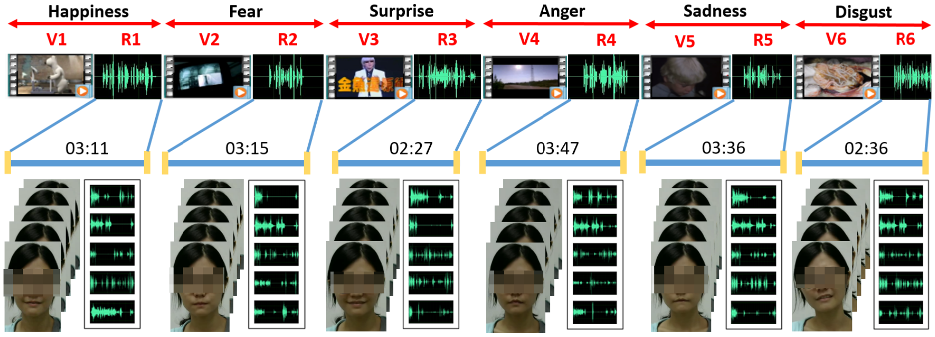


Fig. 1. Data structure of the facial expressions and speech responses of the participant after watching six eliciting video clips. Five facial image sequences and five speech responses corresponding to the five questions asked in each eliciting video are shown.

advantages of FF and DF, a system based on a hybrid strategy is more suitable and robust. MF utilizes the interaction and integration between modalities and incorporates FF and DF to alleviate their disadvantages. Moreover, temporal structures and correlations of the facial and vocal modalities serve a crucial role in the interpretation of human naturalistic audiovisual affective behavior [31]. To consider the temporal structures of the variation in facial action units (AUs) and speech emotions, the cell-coupled LSTM with the $L$-skip fusion mechanism, which integrates the audiovisual features and considers the temporal evolution and the response data structure of these two modalities, was proposed for improving the performance of mood disorder detection systems.

## III. Mood Database Collection

In this paper, a CHI-MEI mood disorder database was developed for system training and evaluation to distinguish between patients with BD, patients with UD, and healthy controls [14]. For data collection, the collection project was approved by the Institutional Review Board (IRB) of the CHI-MEI Medical Center, Taiwan, with IRB serial no: 10403-002. For the collection procedure, six emotional videos [14], [23] that were related to six basic emotions—happiness, fear, surprise, anger, sadness, and disgust—were used to elicit the emotional expressions of the participants. After watching each eliciting video, each participant was interviewed by a clinician, and the speech response and facial expressions of the participants were collected. To select the six eliciting emotion-eliciting candidates, eliciting emotional videos were first collected. Then, the videos were manually evaluated by 55 students to validate their effectiveness in terms of emotion elicitation. Finally, six videos that passed the chi-squared test were selected as the eliciting videos for this paper.

The CHI-MEI mood disorder database collection procedure is described as follows. First, each participant with BD or UD was assessed by a doctor to determine if their physical and mental states were stable before participating in the evaluation. The doctor asked the participants to complete a series of questionnaires to assess the physical and mental states of the participants. These questionnaires included the depression and somatic symptoms scale, mood disorder questionnaire, young mania rating scale, Simpson–Angus extrapyramidal side effects scale, Barnes Akathisia rating scale, and clinical global impression scale. Based on the questionnaire results, the doctor evaluated whether the participants were suitable for participation in the experiments [32]–[37]. Second, each participant was asked to watch six emotion-eliciting video clips one by one, and five prerecorded questions were asked to the participants sequentially after watching each video clip for response collection.

To construct a relaxed environment for data collection, the collection process was conducted in a closed room with one participant and one clinician conducting the process. The video clips were played on a portable monitor, and the elicited audio and video information of the participant was recorded using a webcam (including camera and microphone). Each participant answered five questions immediately after watching each of the six video clips. After displaying each video clip and asking questions related to the clip, a 20-s-long music track was played to relax the participant's emotional state before playing the next video clip.

The CHI-MEI mood disorder database comprises speech responses and facial expressions from 39 participants (12 males and 27 females), including 13 patients with UD, 13 patients with BD, and 13 healthy controls. The average age of the participants was 39.92 (std. 11.48). The videos were recorded in AVI format at a resolution of 640 × 480 and 30 frames/s. The recorded speech data were 16 bits at a frequency of 44.1 kHz and were monophonic. In total, 30 speech responses and 30 facial image sequences were collected from each participant. One facial image sequence and one speech response were collected for each question, and five questions were raised to the participant for each eliciting video, as presented in Fig. 1. Table I lists the average length of each response obtained after watching each video. The CHI-MEI mood disorder database contained 1170 speech responses and 1170 facial image sequences for the 39 participants. The overall speech duration of the six emotion stimulations was 12 h and 17 min.

## IV. Domain Adaptation on the Database

In this paper, an unlabeled domain-specific database (CHI-MEI) was constructed for mood disorder detection. As this paper focused on emotional expression in speech response and facial expressions, the unlabeled CHI-MEI mood disorder database had to be projected into an emotional

TABLE I

AVERAGE LENGTH (IN SECONDS) OF EACH RESPONSE FOR MOOD DISORDER AND HEALTHY CONTROLS

| | BD | UD | Healthy controls |
|---|---|---|---|
| Response 1 | 92.01 | 49.20 | 34.42 |
| Response 2 | 94.66 | 62.45 | 48.26 |
| Response 3 | 82.67 | 29.38 | 34.72 |
| Response 4 | 90.00 | 97.28 | 48.24 |
| Response 5 | 84.07 | 80.09 | 47.94 |
| Response 6 | 58.79 | 34.58 | 46.07 |

space in speech and facial emotional expressions for emotion labeling. The problem of domain adaptation is that the source and target domains are drawn from different distributions. To overcome this difference, several techniques have been investigated, such as covariate shift [38] and representation change [39]. Due to the data bias problem, the labeled big database could be adapted to fit the domain-specific CHI-MEI mood disorder database for training the speech and facial emotion models. In this paper, we used the HSC-based DAE method [14], [15] to solve the problem of domain adaptation as well as bottleneck feature extraction. The Cohn–Kanade (CK+) database and eNTERFACE database were first used to adapt the features of facial expression and speech response, respectively, using HSC algorithm. After domain adaptation, the bottleneck features, which have been proven useful to provide the more meaningful, abstract, and compressed latent representation of the domain-specific CHI-MEI mood disorder database, were extracted using the DAE for constructing the AU and speech emotion detectors.

### A. Target and Source Databases

For domain adaptation, the extended CK+ [40] and eNTERFACE [41] databases were adopted as the databases of the source domain for facial AU and speech emotion adaptation, respectively. In the eNTERFACE database, each subject provided five different sentences for all six emotions. Therefore, 42 participants (18 of which were women) were included, and 30 sentences were recorded for each participant. Similarly, the CK+ database [40] was regarded as the source domain. In total, 593 frame sequences of facial expressions from 123 subjects were included in the CK+ database, and the length of these frame sequences ranged from 10 to 60 frames.

### B. HSC-DAE for Data Adaptation and Bottleneck Feature Extraction

In the domain adaptation process, we transferred the eNTERFACE database to fit the CHI-MEI speech response data and the CK+ database to fit the CHI-MEI facial expression data by adopting the HSC-DAE method [14]. The concept of the HSC algorithm for database adaptation is presented in Fig. 2. Here, **U** represents the source database (eNTERFACE and CK+), and **V** represents the target database (speech and facial data in the CHI-MEI mood disorder database). The shift vector $\Delta_i$ between these two centroids can be estimated as follows:

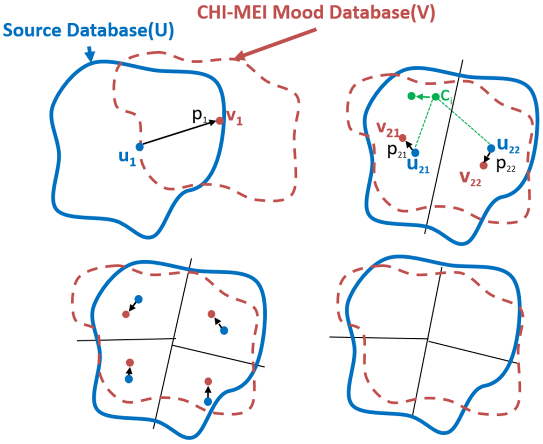


Fig. 2. Concept of the HSC algorithm for database adaptation.

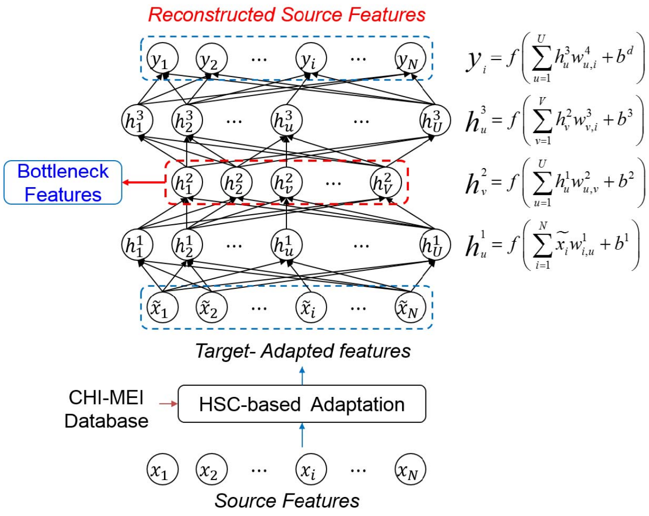

Reconstructed Source Features
$y_1$ $y_2$ ... $y_i$ ... $y_N$
$h_1^3$ $h_2^3$ ... $h_u^3$ ... $h_U^3$
Bottleneck Features
$h_1^2$ $h_2^2$ ... $h_v^2$ ... $h_V^2$
$h_1^1$ $h_2^1$ ... $h_u^1$ ... $h_U^1$
$\tilde{x}_1$ $\tilde{x}_2$ ... $\tilde{x}_i$ ... $\tilde{x}_N$
Target- Adapted features
CHI-MEI Database
HSC-based Adaptation
$x_1$ $x_2$ ... $x_i$ ... $x_N$
Source Features
$y_i = f\left(\sum_{u=1}^{U} h_u^3 w_{u,i}^4 + b^d\right)$
$h_u^3 = f\left(\sum_{v=1}^{V} h_v^2 w_{v,i}^3 + b^3\right)$
$h_v^2 = f\left(\sum_{u=1}^{U} h_u^1 w_{u,v}^2 + b^2\right)$
$h_u^1 = f\left(\sum_{i=1}^{N} \tilde{x}_i w_{i,u}^1 + b^1\right)$

Fig. 3. Adapted data reconstruction process of HSC-DAE for bottleneck feature extraction.

$$\Delta_i = \frac{\sum_{m=1}^{M} w_{im}(\boldsymbol{v}_m - \boldsymbol{u}_m)}{\sum_{m=1}^{M} w_{im}} \tag{1}$$

$$w_{im} = 1/[d(\boldsymbol{x}_i, \boldsymbol{u}_m)], \quad m = 1, 2, \ldots, M \tag{2}$$

where $u_{\boldsymbol{m}}$ is the centroid of the $m$th cluster in the source database, $v_{\boldsymbol{m}}$ is the corresponding target data cluster, $M$ is the total number of clusters, and $d(\cdot)$ is the Euclidean distance.

After data adaptation, the adjusted vector $\widetilde{\boldsymbol{x}_i}$ is fed to the DAE to obtain the output $y_i$. Moreover, we expected the output be the same as the source data $x_i$, as presented in Fig. 3. The adapted data reconstruction process using DAE is described as follows. As shown in Fig. 3, in the DAE map, the input $\widetilde{x_i}$ to a hidden representation $\boldsymbol{h}$ using $f_\theta(\tilde{x}) = S(\boldsymbol{W}\tilde{x} + \boldsymbol{b})$ is presented. Then, $y$ is reconstructed using $\boldsymbol{y} = g_{\theta'}(\boldsymbol{h}) = S(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{b}')$, where $\boldsymbol{W}$ is a weight matrix, $\boldsymbol{b}$ is a bias vector, and $S$ is a logistic sigmoid function. To minimize the average reconstruction error, the backpropagation algorithm was used to update the weights of the DAE by using the following cost

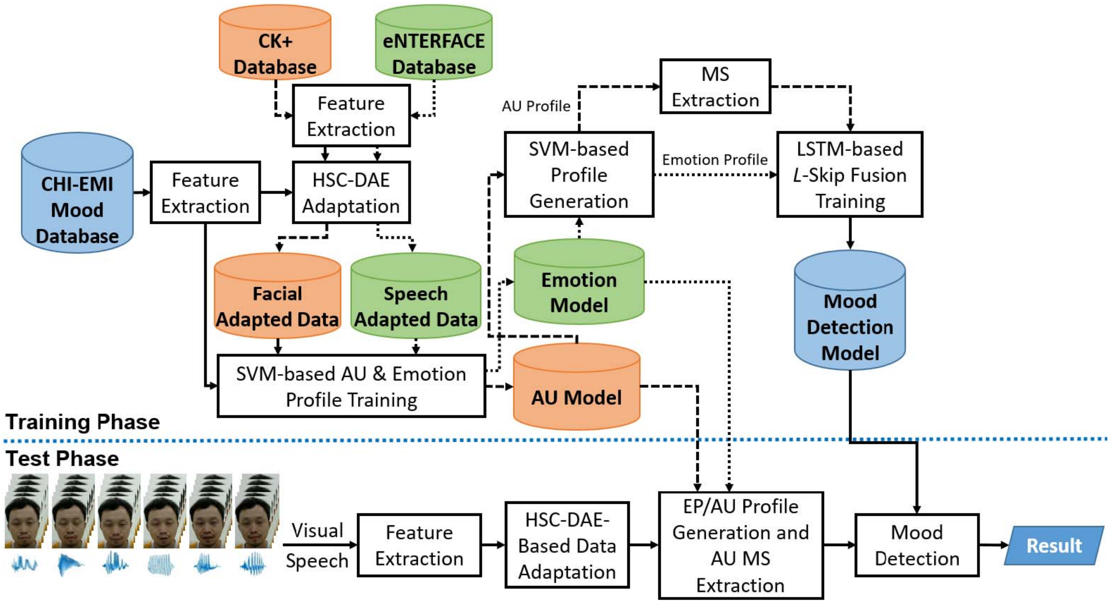


Fig. 4. System framework.

function:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\|x - y\|^2 + \frac{\lambda}{2}\sum_{l=1}^{L}\sum_{j=1}^{J}\left(W_j^l\right)^2. \tag{3}$$

Therefore, by using HSC-DAE, the features of the adapted source database, which has been adapted to fit the CHI-MEI mood disorder database, could be reconstructed back to the features of the original source database. Then, the bottleneck features obtained from the middle bottleneck layer of the DAE can be used to train the support vector machine (SVM)-based AU and speech emotion detectors.

## V. Proposed Methods

Fig. 4 illustrates the proposed system architecture in the training and testing phases. For feature extraction, 16 low-level descriptors and their deltas with 12 functional types were extracted as speech features. For facial signals, 49 facial feature points were extracted using the Gauss–Newton Deformable Part Model (GN-DPM), which includes 10 landmarks for the eyebrow, 12 landmarks for the eye, 9 landmarks for the nose, and 18 landmarks for the mouth. Next, we used the HSC-DAE to obtain facial and speech features from the eNTERFACE and CK+ databases for the speech response part and for the facial expression part of the CHI-MEI mood disorder database, respectively.

Both the audio and facial features extracted were low-level features and were further used to obtain the emotion and AU profiles by using the SVM-based detectors, respectively. In this paper, the adapted speech and facial data were used to train the SVM-based emotion and AU detectors. Based on the trained SVMs, the emotion profiles (EPs) and the AU profiles were generated for feature representation.

Then, the modulation spectrum (MS) of an AU profile sequence was further extracted as the facial feature representation corresponding to one question response. After feature extraction, the features of the entire facial response for each subject were composed of six elicited facial AU MS sequences with respect to six emotion-eliciting videos, each containing five AU modulation spectra corresponding to five questions in each elicitation. Similarly, the features of the entire speech response for each subject comprised six EP sequences, each consisting of five EPs. Finally, the cell-coupled LSTM with the *L*-skip fusion mechanism was adopted for mood disorder detection. The proposed method combined the EP sequence and the AU MS sequence to characterize the response-based temporal context of speech emotion and facial expression.

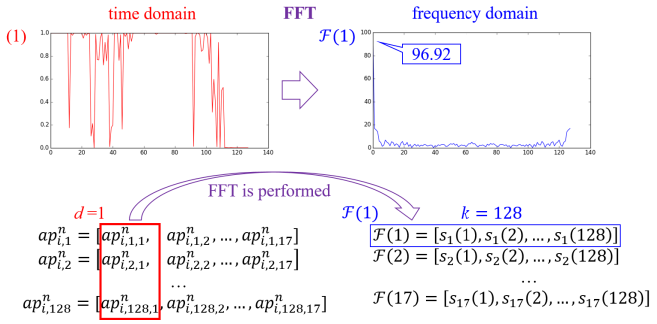

time domain
FFT
frequency domain
(1)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
$\mathcal{F}(1)$
100
80
60
40
20
0
96.92
FFT is performed
$d=1$
$\mathcal{F}(1)$
$k = 128$
$ap_{i,1}^n = [ap_{i,1,1}^n, ap_{i,1,2}^n, \dots, ap_{i,1,17}^n]$
$ap_{i,2}^n = [ap_{i,2,1}^n, ap_{i,2,2}^n, \dots, ap_{i,2,17}^n]$
$\dots$
$ap_{i,128}^n = [ap_{i,128,1}^n, ap_{i,128,2}^n, \dots, ap_{i,128,17}^n]$
$\mathcal{F}(1) = [s_1(1), s_1(2), \dots, s_1(128)]$
$\mathcal{F}(2) = [s_2(1), s_2(2), \dots, s_2(128)]$
$\dots$
$\mathcal{F}(17) = [s_{17}(1), s_{17}(2), \dots, s_{17}(128)]$

Fig. 5. MS of temporal trajectory of the AU profile sequence.

### *A. Facial Feature Extraction*

For each frame of the video clips in the CHI-MEI mood disorder database, we applied OpenCV Haar classifiers [42] to detect facial region, and then, we adopted the GN-DPM [43] for facial feature extraction. We extracted 49 inner facial points from each frame of the facial image sequence. We applied the HSC-DAE adapted database to train the AU detectors and then to estimate the occurrence probability of each AU, known as the AU profile. Then, we obtained the AU profiles that were generated from all the frames in the video clip for forming the AU profile sequence by using the MS method. The concept of the MS method is illustrated in Fig. 5.

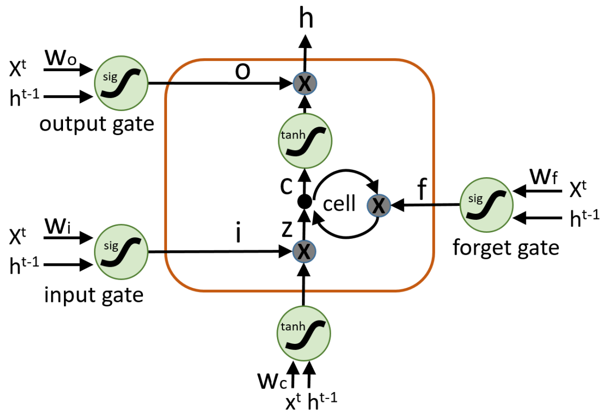


Fig. 6. Structure of the LSTM block.

Here, the time-domain function is the AU profile sequence in which $n$ denotes the corresponding experimental phases and $i$ denotes the video sequence number (128 frames for each sequence). $F(d)$ represents the frequency-domain results from the fast Fourier transform (FFT) performed on the $d$th dimension of the AU profiles. The length of $F(d)$ is $k$, which represents the number of transformed components from the FFT. Energy distributions in the MS represent the temporal fluctuations of the facial expressions. Finally, the variation in the AU profile in the sequence was used to distinguish between the AU profile sequences for mood disorder detection.

### *B. Speech Feature Extraction*

For the speech segments, we adopted a method proposed in [44] to remove the speech segments produced by the clinician and retain the speech segments that only contain speech responses from the patient. After speech segmentation, we applied openSMILE [45] to extract 384-D acoustic features. Then, we constructed an SVM-based EP detector to obtain EPs [46] to express the degree of a set of basic emotions to represent the emotional expression of the input speech.

### *C. Long Short-Term Memory*

Recently, recurrent neural networks (RNNs) have been proven to be useful for yielding the state-of-the-art performance on a variety of real-world sequence processing tasks [47]–[49]. However, conventional RNNs still encounter the vanishing gradient problem [50]. LSTM [51]–[53], as presented in Fig. 6, is an effective technique for overcoming the vanishing gradient problem. LSTM stores the long-term information in memory cells and learns the contextual information that is helpful for the classification task. In contrast to an RNN, a special type of memory block, instead of the nonlinear hidden units in an RNN, is used in an LSTM network. Each memory block is composed of one or more recurrently connected memory cells and three multiplicative units (input, output, and forget gates). The multiplicative gates allow LSTM memory cells to store and access information from long sequences. The output of the block is recurrently connected back to the block input and all the gates. The equations for the forward pass of the LSTM layer are as follows:

$$z^t = g\left(W_z^x x^t + W_z^h h^{t-1} + b_z\right) \tag{4}$$

$$i^t = \sigma\left(W_i^x x^t + W_i^h h^{t-1} + b_i\right) \tag{5}$$

$$f^t = \sigma\left(W_f^x x^t + W_f^h h^{t-1} + b_f\right) \tag{6}$$

$$o^t = \sigma\left(W_o^x x^t + W_o^h h^{t-1} + b_o\right) \tag{7}$$

$$c^t = i^t \otimes z^t + f^t \otimes c^{t-1} \tag{8}$$

$$h^t = o^t \otimes g(c^t) \tag{9}$$

where $z^t$ is the block input, $i^t$ is the input gate, $f^t$ is the forget gate, $o^t$ is the output gate, $g$ is the hyperbolic tangent function, and $\sigma$ is the logistic sigmoid function. $z^t$ is activated using $g$ of the weighted sum of the current input vector $x^t$ and the previous cell state vector $h^{t-1}$. Equation (8) shows that the current state $c^t$ of the cell is summed with the block input $z^t$ that is multiplied by the activation of the input gate $i^t$, and the previous cell value $c^{t-1}$ is multiplied by the forget gate $f^t$, where the pointwise multiplication of two vectors is denoted as $\otimes$. Equation (9) demonstrates that the block output $h^t$ is equal to the current state $c^t$ that is multiplied pointwise by the activation of the output gate $o^t$.

### *D. Cell-Coupled LSTM With the L-Skip Fusion Mechanism*

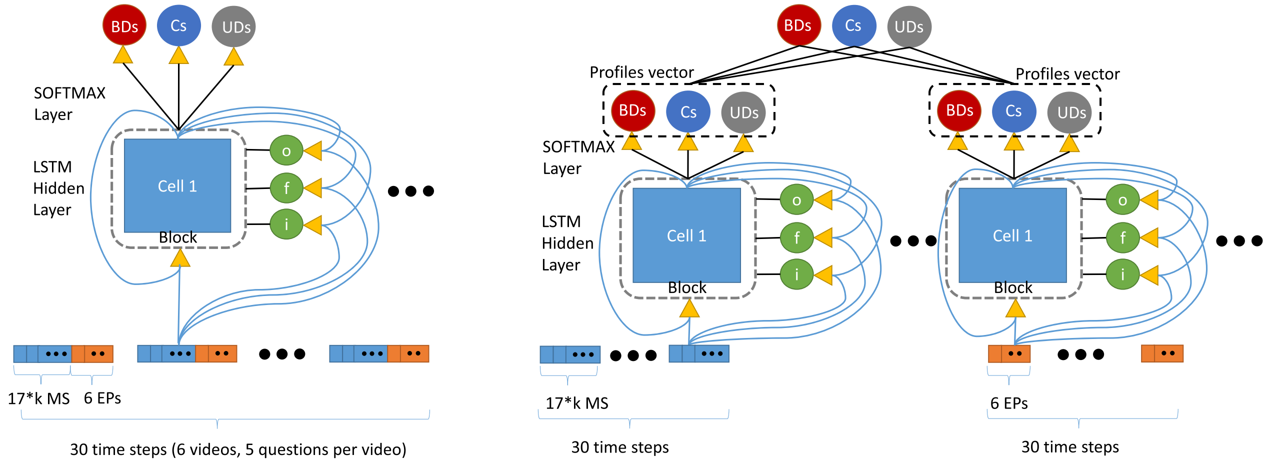
In general, in a multimodal system, different channels and cues are fused to make a decision. Moreover, the system is expected to provide a more precise recognition result compared with that gained using unimodal systems. Fig. 7 illustrates FF and DF by using the LSTM memory block with a single-cell architecture. The structure of Cell 1 is displayed in Fig. 6, and the recurrence is limited to the hidden layer and controlled using three gates. In the FF, we concatenated EPs and MSs to form a supervector as an input to develop a classification model. In the DF, we used EPs and MSs to develop its own unimodal model first. Finally, the unimodal decision results were combined with the product of all the results from the same category. In MF, a cell-coupled LSTM was proposed, and its structure is displayed in Fig. 8. In this paper, the gates for the DF were separated to obtain individual results from the audio and facial signals. In the MF, the results were obtained from the interaction between cells, and thus, the same gates were used.

In a cell-coupled LSTM, a memory block contains two interacting cells. For these cells, (8) and (9) in the vector equations for the forward pass of an LSTM layer were modified to obtain the following equations, respectively,

$$c_i^t = i^t \otimes z^t + f^t \otimes c_i^{t-1} + f^t \otimes c_{\neg i}^{t-1}, \quad i \in \{\text{cell1}, \text{cell2}\} \tag{10}$$

$$h^t = o^t \otimes g\left(c_1^t\right) + o^t \otimes g(c_2^t). \tag{11}$$

Equation (10) presents the memory sharing between two cells inside the memory block and (11) illustrates the summed output of the two cells which delivers recurrent states to the next time step. The process of the proposed model-based fusion is as follows: the MSs and EPs were concatenated in the FF and are used as the input to the memory block.



(a) (b)

Fig. 7. (a) Feature-level and (b) DL multimodal fusions.

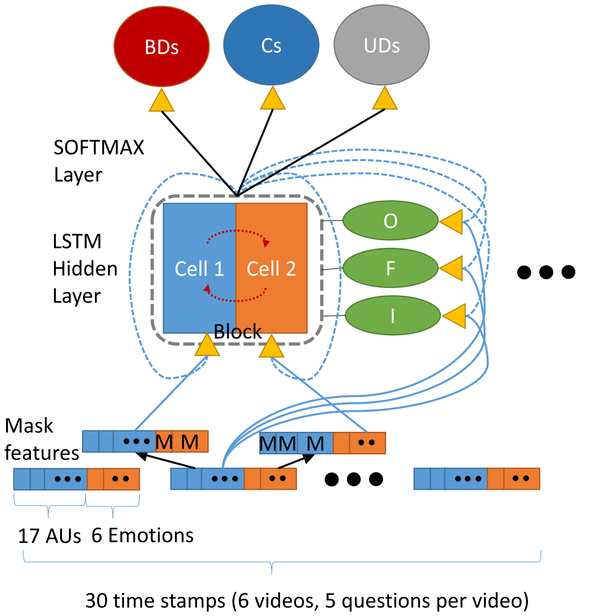


Fig. 8. Model-level fusion using cell-coupled LSTM.

Then, we duplicated the input sequence, masked the EP feature part of one sequence to Cell 1, and masked the MS features of the other sequence to Cell 2. This means that we only sent the MS and EP information to different cells. Each cell maintains its memory history by itself but can interact with each other. The red dotted lines present the interaction between two cells that share the same input inside the memory block. The forget and output gates indicated by the blue dashed lines are the previous hidden output, which is the sum of the cell outputs. According to the data structure of the CHI-MEI mood disorder database, a participant provided 30 speech responses and facial expression sequences after watching six videos, that is, five responses were obtained for each emotion-eliciting video from each participant. As the participant requires some response time to accumulate his or her emotions after watching each eliciting video, a new fusion mechanism considering the time for emotion accumulation is desirable. Skip grams are widely used in the field of speech and language processing [54]. They allow tokens to be “skipped” to overcome the data sparsity problem. The same technique can be applied to the elicited responses in this paper. Accordingly, this paper proposed an $L$-skip loosely coupled strategy to fuse two cells every $L$ (in this paper, $L = 5$) time steps to model the emotion accumulation effect. Thus, the main focus was on loosely coupled time series, in which only the final elicited responses of each elicitation are coupled in time.

As described in the beginning of Section V, the facial features of the entire facial response for each participant comprised six elicited facial AU MS sequences with respect to the six emotion-eliciting videos. Each sequence contains five AU modulation spectra corresponding to five questions in each elicitation. The emotional features of the entire speech response for each subject comprise six EP sequences, each consisting of five EPs. In the cell-coupled LSTM with the five-skip model, the five time-step EPs were fed to Cell 1, on the other hand, the five time-step MSs were fed to Cell 2. Then, the outputs of Cell 1 and Cell 2 interacted (coupled) with each other after five time steps for each facial and audio response, that is, the cell-coupled LSTM with the five-skip loosely coupled fusion mechanism was applied to fuse the two modalities after every five time steps. The alignment between facial and speech responses was based on response instead of time (or frame). This implies that each speech response was aligned to one facial response. For the entire facial and speech elicited responses, six interactions were noted between Cells 1 and 2 in the cell-coupled LSTM with the five-skip fusion model. The final mood disorder detection result was obtained through a softmax layer.

## VI. Experimental Results and Discussion

The proposed multimodal fusion method based on cell-coupled LSTM was evaluated and compared with different classifiers for comparing their performance. First, the effectiveness of using the HSC-DAE adaptation method and MS was tested. Subsequently, the performance of unimodal

TABLE II
ABBREVIATIONS USED IN EXPERIMENTS

| Abbreviations | Descriptions |
|---|---|
| MS | Modulation spectrum of AU profiles (unimodal) |
| EP | Emotion profiles (unimodal) |
| FF | Feature-level fusion (multimodal) |
| DF | Decision-level fusion (multimodal) |
| MF | Cell-coupled model-level fusion (multimodal) |
| MF5 | Cell-coupled model-level fusion with 5-skip fusion mechanism (multimodal) |

approaches for detecting mood disorder was evaluated. Finally, the performance of the proposed method was compared with that of other multimodal fusion methods and classifiers.

As mentioned in Section III, 39 participants were included, of which 13 had UD, 13 had BD, and 13 were healthy controls, in the CHI-MEI mood disorder database. Therefore, in this paper, the experimental results were evaluated using a 13-fold cross validation. We randomly sampled one participant without replacement from each category (those with UD, BD, and healthy controls) to form 13 subsets. Each subset contained three subjects (one from each category) for testing, and the remaining 12 subsets (36 subjects) were used to train the model. Before experimentation, we linearly scaled each attribute to the range of [0, 1] for both the training and testing data. The process was repeated 13 times, and the results were averaged to obtain the evaluation results.

As the three popularly used tools—SVM [55], LSTM, and DAE—were used in the following experiments, we first fine-tuned and validated these models to ensure appropriate implementation and use. First, we conducted experiments to evaluate the performance of the SVM-based AU and emotion detection conducted on the training data sets: the CK+ and eNTERFACE databases. The LibSVM tool was adopted to obtain the probability outputs of each AU and emotion. The output of the LibSVM tool was represented as a probability. The accuracy of the AU on the training data set was 90.95%, and the accuracy of the EP on the training data set was 74.40% when a fivefold cross validation was used. Second, we used an LSTM implemented by Keras and Theano [56], [57] to classify emotions in the Berlin Emotion Database [58], and the accuracy achieved was 85.78%. Finally, a DAE was trained to reconstruct a clean "reconstructed" input from a corrupted version of the input data [59]. We conducted experiments to evaluate the performance of the LSTM emotion detection based on an HSC-DAE on the eNTERFACE database. The inputs of LSTM were the acoustic features with or without using an HSC-DAE. The accuracy of the LSTM without an HSC-DAE was 75.08% and that with an HSC-DAE was 97.42% when fivefold cross validation was used. From the validation results, we confirmed the effectiveness of these models and the tools in the experiments. For further details on the experiments, see Table II, which lists and describes the abbreviations used in the following experiments.

### *A. Performance of HSC-DAE Adaptation and MS*

As mentioned in Section III, the CHI-MEI mood disorder database contained both the facial and audio modalities, and all facial expressions and speech responses were spontaneous. We considered both modalities that reflect different properties together for mood disorder detection. For each speech response datum, we extracted 384 acoustic features for the response-based speech data and the 98-D AU profile features of 128 frames for the response-based facial data. To optimize the parameters used in the HSC-DAE, the number of hidden nodes was determined, as presented in Fig. 9.

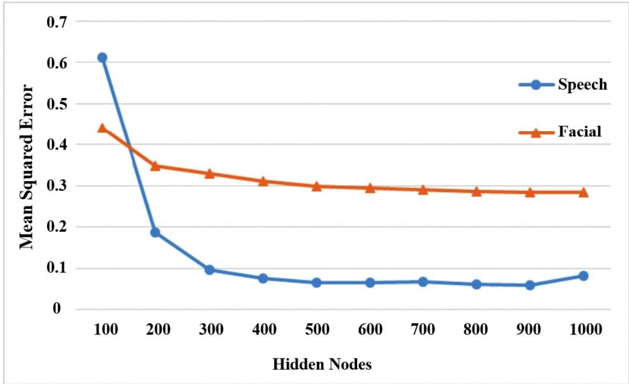


Fig. 9. Performance of the HSC-DAE between different numbers of hidden nodes in the bottleneck layer of the DAE.

TABLE III
COMPARISON OF THE METHODS WITH AND WITHOUT HSC-DAE FOR MOOD DISORDER DETECTION

| | Without HSC-DAE | With HSC-DAE |
|---|---|---|
| AU | 41.60% | **55.60%** |
| EP | **50.09%** | 49.83% |

For both the speech and facial data, we selected 500 hidden nodes of the HSC-DAE to obtain the minimum mean-squared error. To verify if the HSC-DAE was helpful in distinguishing between mood disorders, we used the EPs predicted by two SVMs. One of the SVMs was trained by the data adapted using the HSC-DAE, and the other was trained by the original source data. For comparison, the accuracy of the methods with and without using the HSC-DAE is presented in Table III. Although the method that used the HSC-DAE presented an improved mood disorder detection performance in terms of the AU profiles, comparable results were obtained for the method with and without HSC-DAE in terms of EP.

As indicated by Yen [60], the MS has an extraordinary ability for analyzing temporal fluctuations in AU for mood disorder. We applied the MS to obtain the temporal fluctuation information of AU profiles. The value of $k$ denotes the number of transformed components from the 128-point FFT (128 is the length of the data sequence). After applying FFT to each dimension of the AU profiles, we only retained the first $k$ low-pass components of the FFT-transformed results and concatenated them to form a feature vector. Fig. 10 presents the results obtained after comparing different values of $k$ for MS extraction in our experiments. The results revealed that when $k = 24$, the proposed approach achieved the highest accuracy. However, as the value of $k$ increased, the performance

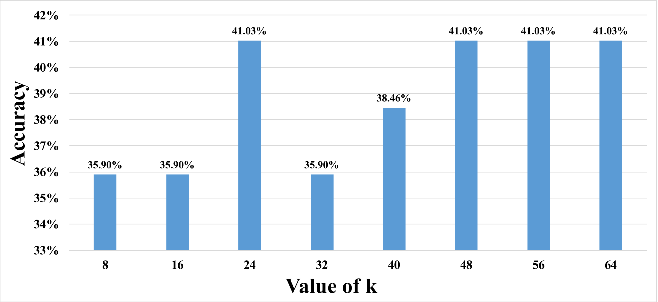


Fig. 10. Comparison for applying different values of $k$ to MS.

TABLE IV

LSTM-Based Detection Performances for Mood Disorder Detection Using Unimodal Features for Different Hidden Node Numbers

| Nodes | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|
| MS | 28.2% | 35.9% | **41.0%** | 33.3% | 33.3% | 35.9% |
| EP | 66.7% | 66.7% | **69.2%** | 66.7% | **69.2%** | 66.7% |

decreased and finally recovered until $k$ was 48. Moreover, equivalent performance was obtained, when $k$ was 24 and 48. This implies that the components between $k = 24$ and 48 were unreliable, and thus, the performance deteriorated. In this paper, further experiments were conducted using a $k$ value of 24.

### *B. Unimodal Approaches for Mood Disorder Detection*

In this section, the performances of approaches using unimodality (MS or EP) were compared. We first used MS and EP separately to train the LSTM-based systems with 16, 32, 64, 128, 256, and 512 hidden nodes. In the MS comparison experiment, the input of LSTM was a 408 (17 × 24)-D MS of AU profiles, and the output of LSTM was the detection results of mood disorders. In the EP comparison experiment, the input of LSTM was a 6-D EP, and the output of LSTM was the detection results of mood disorders. As mentioned in Section V-D, we obtained 30 feature sequences for each participant to train the LSTM-based systems [61]. Accordingly, the backward signals were truncated into a 30-time-step feature sequence using the backpropagation algorithm through time with the stochastic gradient descent (SGD) optimization. The detection performance results for the test set are presented in Table IV. The highest value of each row (modality) is highlighted in bold. As can be seen, when the number of nodes was equal to 64, both the MS and EP achieved the highest accuracy, each in its own modality. Furthermore, the accuracies of the LSTM using EP were consistently superior to the accuracies of the LSTM using MS for all hidden node numbers. The results of the LSTM using unimodal features confirmed that EP tends to be more crucial for discriminating between mood disorders. For the MS of the AU profiles, a lower performance was obtained for LSTM modeling. A major reason for the low performance was the difficulty in detecting meaningful AUs in real-world spontaneous facial expression data without well-labeled data.

TABLE V

Accuracy Comparison of Unimodal Approaches

| | HMM | SVM | MLP | LSTM | CNN | RNN | GRU |
|---|---|---|---|---|---|---|---|
| MS | 35.9% | **41.0%** | 35.0% | **41.0%** | 30.76% | 35.90% | 30.77% |
| EP | 53.9% | 49.8% | 42.0% | **69.2%** | 38.46% | 41.02% | 33.33% |

We also compared the performance of the LSTM-based method with those of the commonly used classifiers such as the hidden Markov model (HMM), SVM, multilayer perceptron (MLP), CNN, RNN, and gated recurrent unit (GRU), all of which use HSC-DAE. Multiclass SVMs with radial basis function kernels and sequential minimal optimization were adopted to train each modality. The three-layer MLPs with a hyperbolic tangent (tanh) activation function were trained through SGD backpropagation. A participant could be categorized as UD, BD, or C by detecting the mood of the participant using a 30-time-step feature sequence in the LSTM-based method. For comparison between the SVM and MLP methods, the product of the 30 profiles, each obtained from one time step predicted by an SVM or MLP, was estimated to detect mood disorder for each participant. For comparison with the method using HMM [62], three HMM-based recognition models, each comprising six states, were trained for UD, BD, and C. The input of the HMMs was 30 MS feature sequences of 30 time steps, and the output was the classification results. Optimal performance was attained when six states were used in the HMM. This is because the six states coincided with the six elicited emotions; thus, optimal performance was attained. For comparisons with CNN, RNN, and GRU models, MS and EP features of 30 time steps were used as inputs of these models, and the output was the classification results. The use of MS features, 32 filters, 10 kernel sizes, 36 batches, and 6 strides in CNN achieved optimal performance. The use of EP features, 32 filters, 6 kernel sizes, 36 batches, and 6 strides in CNN achieved optimal performance. The use of MS and EP features, 32 hidden nodes, and 12 batches in RNN achieved optimal performance. The use of MS and EP features and 12 batches in GRU achieved optimal performance. Table V presents the unimodal comparison results for seven different methods: HMM, SVM, MLP, CNN, RNN, GRU, and LSTM. The highest accuracies of the methods obtained after parameter tuning are listed in Table V. The comparison results revealed that the LSTM-based approach outperformed the HMM-, SVM-, MLP-, CNN-, RNN-, and GRU-based unimodal approaches.

### *C. Multimodal Approaches Used for Mood Disorder Detection*

The performance of the multimodal fusion approaches, including FF, DF, cell-coupled MF, and cell-coupled MF with a five-skip mechanism (MF5), was compared in this section. We established the parameters of the three detection methods that were specified in Sections VI-A and B. Briefly, in FF, EPs and MSs were concatenated and used as input features for the FF. In DF, the results were obtained from the product of the unimodal outputs from EPs and MSs. In MF, the interaction of two modalities was used as the communication between

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT LSTM-BASED MULTIMODAL FUSION APPROACHES

| Nodes | 16 | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|---|
| FF | 38.5% | 43.6% | **48.7%** | 35.9% | 35.9% | 38.5% |
| DF | 38.5% | 46.2% | **74.5%** | 61.5% | 69.2% | 64.1% |
| MF | 59.0% | **64.1%** | **64.1%** | 53.8% | 61.5% | 59.0% |
| MF5 | 69.2% | 71.8% | **76.9%** | 74.5% | **76.9%** | 74.5% |

two cells in an LSTM block. Table VI presents the detection performance when LSTM-based methods with different multimodal fusion approaches were used. In the multimodal comparison experiment, the input of the proposed cell-coupled LSTM with the $L$-skip method was a 414-D vector (408-D AU profile combined with 6-D EP). Moreover, the output was the detection results of mood disorders. We found that the hidden layer with 64 nodes achieved the highest accuracy for all fusion methods. This finding is consistent when all unimodal methods are used, thus implying that the topology with 64 hidden nodes is suitable for our data set. In general, based on the experimental results, the multimodal fusion methods outperformed the unimodal methods. An optimal accuracy of 76.9% was achieved when MF5 with 64 and 256 hidden nodes was used. Moreover, FF and MF outperformed the unimodal approaches that use MS. Early fusion methods, such as FF, which interacts among different modalities at the lower level, could be undermined by one poor modality. The proposed MF fusion also interacted among different modalities from the beginning stage. However, the modalities were only connected to each other in the recurrent state; thus, the performance degradation was not severe. The other reason for the poorer performance of the MF is that it generally requires a large amount of training data to obtain a well-trained model. In this paper, the collected data were not sufficiently large due to the difficulty of data collection; thus, the detection performance was degraded. In contrast to the FF and MF methods, a late fusion method, such as DF, outputs the results to the classifiers of other modalities at the final stage to form a high-level representation. This representation might prevent an inappropriate interaction effect; thus, more favorable performance (74.5%) is attained while using DF than unimodal methods. Therefore, the proposed cell-coupled LSTM with five-skip MF delays the interaction between two cells after every five time steps to prevent an inappropriate interaction effect.

## VII. CONCLUSION

Based on the temporal variations in speech response and facial expressions, this paper proposed a cell-coupled LSTM with an $L$-skip fusion mechanism to model long-range contextual information and conduct mood disorder detection. This system can be used to avoid delay in providing suitable treatment due to misdiagnosis.

To overcome database bias, the CK+ and eNTERFACE databases were adapted to the CHI-MEI mood disorder database in this paper by employing the HSC-DAE method. Then, for AU profile and EP generation, data were used to construct an SVM-based detector. The AU profiles were further used to extract MSs to characterize long-range contextual information of facial expressions. Finally, a cell-coupled LSTM with an $L$-skip fusion mechanism was applied to fuse the MSs and the EPs for mood disorder detection. Experimental results revealed that the proposed cell-coupled LSTM with the $L$-skip method outperformed other classifiers. The contributions of this paper are as follows: 1) the promising fusion approach is based on facial expressions and speech emotions for diagnosis assistance and 2) the proposed approach is noninvasive, cost-effective, and requires less time compared with other methods with biosignal-based modalities, such as EEG and functional magnetic resonance imaging (fMRI). Although this is a preliminary study that used a small database, the encouraging results provide doctors with additional information for mood disorder diagnosis and further study.

In the future, we can improve the system performance by combining other forms of information, such as head pose, gesture, and lexical information. We can also consider the personality traits of individual patients as a crucial factor for the detection of mood disorders.

## REFERENCES

[1] P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5 (R))*. Arlington, VA, USA: American Psychiatric, 2013.

[2] D. J. Smith and N. Craddock, "Unipolar and bipolar depression: Different or the same?" *Brit. J. Psychiatry*, vol. 199, no. 4, pp. 272–274, 2011.

[3] R. H. Perlis, "Misdiagnosis of bipolar disorder," *Amer. J. Managed Care*, vol. 11, no. 9, pp. S271–S274, 2005.

[4] A. Greco, G. Valenza, A. Lanata, G. Rota, and E. P. Scilingo, "Electrodermal activity in bipolar patients during affective elicitation," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 6, pp. 1865–1873, Nov. 2014.

[5] A. Grünerbl *et al.*, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 1, pp. 140–148, Jan. 2015.

[6] Y. Katyal, S. V. Alur, S. Dwivedi, and R. Menaka, "EEG signal and video analysis based depression indication," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol. (ICACCCT)*, May 2014, pp. 1353–1360.

[7] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel weighted speech classification system for prediction of major depression in adolescents," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 2, pp. 497–506, Feb. 2013.

[8] M. N. Stolar, M. Lech, and N. B. Allen, "Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 987–991.

[9] S. Mantri, P. Agrawal, D. Patil, and V. Wadhai, "Cumulative video analysis based smart framework for detection of depression disorders," in *Proc. Int. Conf. Pervasive Comput. (ICPC)*, Jan. 2015, pp. 1–5.

[10] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 525–536, Mar. 2018. doi: 10.1109/JBHI.2017.2676878.

[11] M. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[12] S. K. D'mello, and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 43:1–43:36, 2015.

[13] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[14] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[15] S. Furui, "Unsupervised speaker adaptation based on hierarchical spectral clustering," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 12, pp. 1923–1930, Dec. 1989.

[16] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–7.

[17] L. S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 5154–5157.

[18] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011.

[19] M. H. Sanchez *et al.*, "Using prosodic and spectral features in detecting depression in elderly males," in *Proc. INTERSPEECH*, 2011, pp. 3001–3004.

[20] O. Schleusing, P. Renevey, M. Bertschi, S. Dasen, J. M. Koller, and R. Paradiso, "Monitoring physiological and behavioral signals to detect mood changes of bipolar patients," in *Proc. Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, Mar. 2011, pp. 130–134.

[21] Z. N. Karam *et al.*, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4858–4862.

[22] A. Lanata, A. Greco, G. Valenza, and E. P. Scilingo, "A pattern recognition approach based on electrodermal response for pathological mood identification in bipolar disorders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3601–3605.

[23] G. Bersani *et al.*, "Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders," *J. Neuropsychiatric Disease Treat.*, vol. 9, p. 1137, Aug. 2013.

[24] D. P. David, M. G. Soeiro-de-Souza, R. A. Moreno, and D. S. Bio, "Facial emotion recognition and its correlation with executive functions in bipolar I patients and healthy controls," *J. Affect. Disorders*, vol. 152, pp. 288–294, Jan. 2014.

[25] A. C. Vederman *et al.*, "Modality-specific alterations in the perception of emotional stimuli in bipolar disorder compared to healthy controls and major depressive disorder," *Cortex*, vol. 48, no. 8, pp. 1027–1034, 2012.

[26] M. Summers, K. Papadopoulou, S. Bruno, L. Cipolotti, and M. A. Ron, "Bipolar I and bipolar II disorder: Cognition and emotion processing," *Psychol. Med.*, vol. 36, no. 12, pp. 1799–1809, 2006.

[27] G. Bersani *et al.*, "Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders," *Neuropsychiatric Disease Treat.*, vol. 9, p. 1137, Aug. 2013.

[28] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, pp. e12-1–e12-18, Nov. 2014.

[29] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[30] P. K. Atrey, M. S. Kankanhalli, and J. B. Oommen, "Goal-oriented optimal subset selection of correlated multimedia streams," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 3, no. 1, pp. 1–24, 2007.

[31] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su, "Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals," *J. Ambient Intell. Hum. Comput.*, vol. 8, no. 6, Nov. 2017, pp. 895–906. doi: 10.1007/s12652-016-0395-y.

[32] M. Hamilton, "A rating scale for depression," *J. Neurol., Neurosurg., Psychiatry*, vol. 23, no. 1, p. 56, 1960.

[33] R. M. Hirschfeld *et al.*, "Development and validation of a screening instrument for bipolar spectrum disorder: The mood disorder questionnaire," *Amer. J. Psychiatry*, vol. 157, no. 11, pp. 1873–1875, 2000.

[34] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, "A rating scale for mania: Reliability, validity and sensitivity," *Brit. J. Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[35] S. Leucht, G. Pitschel-Walz, D. Abraham, and W. Kissling, "Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials," *Schizophrenia Res.*, vol. 35, no. 1, pp. 51–68, 1999.

[36] T. R. Barnes, "A rating scale for drug-induced Akathisia," *Brit. J. Psychiatry*, vol. 154, no. 5, pp. 672–676, 1989.

[37] W. Guy, "ECDEU assessment manual for psychopharmacology-revised," US Dept. Health, Educ., Welfare, Public Health Service, Alcohol, Drug Abuse, Mental Health Admin., Nat. Inst. Mental Health, Psychopharmacology Res. Branch, Division Extramural Res. Programs, Bengaluru, India, 1976, pp. 534–537.

[38] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learn. Res.*, vol. 8, pp. 985–1005, May 2007.

[39] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, pp. 137–144.

[40] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 94–101.

[41] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. IEEE Workshop Multimedia Database Manage.*, Apr. 2006, pp. 1–8.

[42] G. Bradski, "The OpenCV library," *Doctor Doobs J.*, vol. 25, no. 11, pp. 120–126, 2000.

[43] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1851–1858.

[44] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in MATLAB," Dept. Inform. Telecommun., Comput. Intell. Lab. (CIL), Inst. Inform. Telecommun., Univ. Athens, Athens, Greece, 2009.

[45] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[46] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[47] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3128–3137.

[48] A. Graves, D. Eck, N. Beringer, and J. Schmidhuber, "Biologically plausible speech recognition with LSTM neural nets," in *Proc. 1st Int. Workshop Biol. Inspired Approaches Adv. Inf. Technol.*, 2004, pp. 127–136.

[49] N. Beringer, "Human language acquisition methods in a machine learning task," in *Proc. 8th Int. Conf. Spoken Lang. Process.*, 2004, pp. 2233–2236.

[50] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[52] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 115–143, 2003.

[53] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[54] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *Proc. 5th Int. Conf. Lang. Resour. Eval. (LREC)*, 2006, pp. 1–4.

[55] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[56] F. Bastien *et al.*, "Theano: New features and speed improvements," in *Proc. Deep Learn. Unsupervised Feature Learning NIPS Workshop*, 2012, pp. 1–10.

[57] J. Bergstra *et al.*, "Theano: A CPU and GPU math expression compiler," in *Proc. Python Sci. Comput. Conf. (SciPy)*, 2010, pp. 1–7.

[58] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. 1st Richmedia Conf.*, 2003, pp. 109–119.

[59] A. Budiman, M. I. Fanany, and C. Basaruddin, "Stacked denoising autoencoder for feature representation learning in pose-based action recognition," in *Proc. IEEE 3rd Global Conf. Consum. Electron. (GCCE)*, Oct. 2014, pp. 684–688.

[60] H.-H. Yen, "Detection of mood disorder using modulation spectrum of facial action unit profiles," M.S. thesis, Dept. Comput. Sci. Inf. Eng., Nat. Cheng Kung Univ., Tainan, Taiwan, 2015.

[61] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su, "Detection of mood disorder using speech emotion profiles and LSTM," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Tianjin, China, Oct. 2016, pp. 1–5.

[62] K. Murphy. (1998). *Hidden Markov Model (HMM) Toolbox for MATLAB*. [Online]. Available: https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

**Ming-Hsiang Su** (M'18) received the B.S. degree in computer science and information engineering from Tunghai University, Taichung, Taiwan, in 2001, the M.S. degree in management information systems from the National Pingtung University of Science and Technology, Pingtung City, Taiwan, in 2003, and the Ph.D. degree in computer science and information engineering from National Chung Cheng University, Tainan, Taiwan, in 2013.

He is currently a Post-Doctoral Fellow with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan. His current research interests include e-learning, artificial intelligence, machine learning, multimedia signal processing, and personality detection.

**Chung-Hsien Wu** (SM'03) received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively.

Since 1991, he has been with the Department of Computer Science and Information Engineering, NCKU, where he became the Chair Professor in 2017. In 2003, he joined the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, as a Visiting Scientist. From 2009 to 2015, he was the Deputy Dean of the College of Electrical Engineering and Computer Science, NCKU. His current research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing.

Dr. Wu is currently a member of APSIPA BoG. He was a recipient of the 2018 APSIPA Sadaoki Furui Prize Paper Award in 2018 and the Outstanding Research Award of Ministry of Science and Technology, Taiwan, in 2010 and 2016. He was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2010 to 2014 and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING from 2010 to 2014. He is currently an Associate Editor of the *ACM Transactions on Asian and Low-Resource Language Information Processing*.

**Kun-Yi Huang** received the B.S. and M.S. degrees in computer science and information engineering from the Southern Taiwan University of Science and Technology, Tainan, Taiwan, in 2010 and 2012, respectively, and the Ph.D. degree with the Institute of Computer Science and Information Engineering, National Cheng Kung University, Tainan, in 2019, respectively.

His current research interests include biomedical signal processing, emotion recognition, spoken language processing, and speech recognition.

**Tsung-Hsien Yang** received the M.S. degree in management information systems from the National Pingtung University of Science and Technology, Pingtung City, Taiwan, and the Ph.D. degree in management information systems from National Sun Yat-sen University, Kaohsiung, Taiwan.

He was a Post-Doctoral Fellow with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He is currently a Researcher with the Telecommunication Laboratories Chunghwa Telecom Co., Ltd., Taoyuan, Taiwan. His current research interests include artificial intelligence, machine learning, multimedia signal processing, dialogue systems, and natural language understanding.