TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection

Li Zhou, Zhenyu Liu, *Member, IEEE*, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu, *Senior Member, IEEE*

***Abstract*— In recent years, with the widespread popularity of the Internet, social media has become an indispensable part of people's lives. People regard online social media as an essential tool for interaction and communication. Due to the convenience of data acquisition from social media, mental health research on social media has received a lot of attention. The early detection of psychological disorder based on social media can help prevent further deterioration in at-risk people. In this paper, depression detection is performed based on non-verbal (acoustics and visual) behaviors of vlog. We propose a time-aware attention-based multimodal fusion depression detection network (TAMFN) to mine and fuse the multimodal features fully. The TAMFN model is constructed by a temporal convolutional network with the global information (GTCN), an intermodal feature extraction (IFE) module, and a time-aware attention multimodal fusion (TAMF) module. The GTCN model captures more temporal behavior information by combining local and global temporal information. The IFE module extracts the early interaction information between modalities to enrich the feature representation. The TAMF module guides the multimodal feature fusion by mining the temporal importance between different modalities. Our experiments are carried out on D-Vlog dataset, and the comparative experimental results report that our proposed TAMFN outperforms all benchmark models, indicating the effectiveness of the proposed TAMFN model.**

***Index Terms*— Depression, vlog, non-verbal behaviors, automatic detection, time-aware attention-based multimodal fusion depression detection network (TAMFN).**

I. Introduction

DEPRESSION is a common and high incidence of mental disorders [1]. According to the World Health Organization (WHO) research on depression [2], there are about 350 million people with depression worldwide, and the incidence is increasing year by year. According to epidemiological studies, the lifetime prevalence of depression was 6.9% [3].

The main symptoms of depression are long-term low mood, loss of interest and pleasure, however, severe depression patients may self-harm and commit suicide [4]. Depression not only brings grave harm to the life of patients but also brings a burden to society [5]. Early detection of depression is beneficial in reducing the loss of individuals and society [6]. Therefore, early detection and intervention of depression are essential.

Currently, the typical clinical diagnosis methods of depression are mainly based on patients' self-evaluation and psychiatrist' clinical diagnoses. At present, the commonly used clinical diagnostic tools mainly include the Diagnostic and statistical manual of mental disorders (DSM-5) [7], Hamilton Depression Scale (HAMD) [8], Beck Depression inventory (BDI) [9], and Patient Health Questionaire - 9 items (PHQ-9) [10]. In the hospital, doctors mainly communicate with patients and make diagnosis according to the interviewer's performance, which is highly dependent on the doctor's communication style and clinical experience. Because patients with depression have a sense of stigma and are unwilling to communicate with others, this will interfere with psychiatrist' judgment [11]. The diagnosis method of depression based on consultation and questionnaire filling is subjective, and diagnosis result is primarily influenced by personal factors such as self-symptom description with patients, communication ability of psychiatrist, and clinical experience [12]. According to the survey report of Bishop et al. [13], the number of psychiatrists in the United States is constantly declining, and there is only one psychiatrist for every 100,000 residents in the hospital referral area. Due to the high misdiagnosis rate of depression and the shortage of psychiatrists, it is of great significance to explore and develop an additional diagnostic tool that can objectively evaluate depression accurately.

In recent years, the method of automatic detection of depression based on physiological signals has attracted many researchers' attention. Physiological signals are divided into internal and external ones. The internal physiological signals mainly include heart rate [14], [15], EEG [16], [17], galvanic skin [18], [19], nuclear magnetic [20], [21] and other signals. The external physiological signals mainly include eye

Manuscript received 19 August 2022; revised 23 September 2022 and 22 October 2022; accepted 15 November 2022. Date of publication 23 November 2022; date of current version 2 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200; in part by the National Natural Science Foundation of China under Grant 61632014, Grant 61627808, Grant 61802159, and Grant 61802158; and in part by the Fundamental Research Funds for Central Universities under Grant lzujbky-2019-26 and Grant lzujbky-2021-kb26. *(Corresponding authors: Zhenyu Liu; Bin Hu.)*

The authors are with the Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, Lanzhou, Gansu 730000, China (e-mail: zhoul2020@lzu.edu.cn; liuzhenyu@lzu.edu.cn; shanggzx20@lzu.edu.cn; yuanxy20@lzu.edu.cn; liyt20@lzu.edu.cn; bh@lzu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3224135

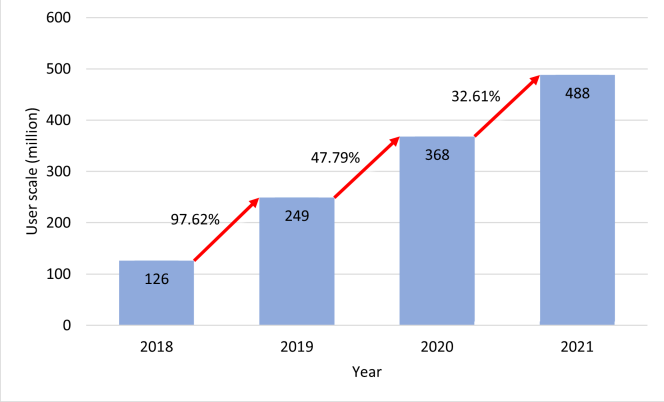


Fig. 1. The growth trend of vlog users in China.

movement [22], [23], gait [24], [25], speech [26], [27] and facial expression [28], [29], etc. The acquisition of these physiological signals above requires a lot of human and material resources and is not easy to promote. Nowadays, over 2 billion users worldwide use social media [30], among which vlog is popular. According to iiMedia Research Group's report (Fig. 1), the number of vlog users in China has increased from 2018 to 2021. As social media data are readily available, more and more researchers have begun to pay attention to the research on depression detection based on social media [31], [32].

In this paper, non-verbal (acoustic and visual) features in vlog data are used to detect depression. The following challenges are faced with depression detection in vlog data: 1) How to mine more helpful information from multi-modal data and obtain rich feature representations; 2) How to obtain efficient fusion feature representation. Therefore, this paper proposes a time-aware attention-based multimodal fusion network (TAMFN). The TAMFN model consists of the following modules: (i) Temporal convolutional network with global information (GTCN) module. The temporal convolutional network (TCN) can learn long-term information dependence by dilated convolution, but the receptive field is limited by the size of the convolution kernel using the convolution feature extraction method. To obtain more comprehensive temporal information, we propose the GTCN, which focuses not only on modeling local information but also on changes in global information. (ii) Intermodal feature extraction (IFE) module. To fully extract the interaction information between the early modalities, an additional GTCN model branch is adopted to fuse the acoustic and visual intermediate features extracted by the two GTCN models to improve the feature interaction between different modalities and increase the diversity of features. (iii) Time-aware attention multimodal fusion (TAMF) module. We thoroughly mine the temporal interaction information between different modalities to obtain the efficient fusion feature representation. Firstly, the temporal attention vector of each modality is constructed. Then the knowledge of the temporal attention vector between modalities is mixed by the sparse multi-layer perceptron. Finally, the hybrid temporal attention vector guides the feature fusion of multi-modalities.

II. Related Work

A. Depression Detection of Social Media Users Based on Single Modality

Hiraga et al. [33] explored the influence of different linguistic features on depression detection and found that depressed people like to use words such as 'become', 'get tired' and 'die', while healthy people like to use words such as 'go up', 'understand' and 'read'. Sadeque et al. [34] used the features based on depression vocabulary and a unified medical language system to achieve a better performance in the early detection of depression. Orabi et al. [35] proposed a word embedding optimization method, which achieved good performance in the depression detection task based on Twitter posts. Aragon et al. [36] proposed a new social media document representation method that automatically uses sub-word embedding to generate sentiment representations. Burdisso et al. [37] proposed a text classification supervised learning model named SS3. SS3 supports incremental classification and learning, which not only has good performance on the task of early depression detection but also has better interpretability. ALSAGRI et al. [38] extracted features such as self-centeredness, word usage, and emotion from users' posts to describe users' behaviors. The experimental results show that the more descriptive features used, the more accurate the detection of depressed users. Chiong et al. [39] proposed text preprocessing and text-based feature method for depression detection and proved the universality of their approach through cross-database experiments. Lara et al. [40] proposed the DeepBoSE model, which can extract the lexical sentiment information of user posts and utilize the attributes of deep learning models while retaining interpretability. Agirrezabal and Amann [41] integrated pre-trained BERT, RoBERTa, and XLNET models to extract text features and vote for depression detection. Kayalvizhi et al. [42] proposed a standard dataset for depression detection in social media, detecting depression levels from user posts. To solve the problem of sample imbalance, they introduced Word2Vec vectorized data augmentation technology.

B. Depression Detection of Social Media Users Based on Multimodality

De Choudhury et al. [43] constructed a probabilistic model through the user's social activities, emotions, and language signals and used the model to introduce the social media depression index to evaluate the user's depression level. Shen et al. [44] constructed a multimodal learning dictionary for depression detection by extracting features related to depression. They found that the first-person pronouns in the posts of depressed users were nearly 200% higher than those of healthy people. Vedula et al. [45] collected clues such as user language style, emotional signals, and user engagement on Twitter. They found that depression users published posts with stronger negative emotions, frequent self-focused pronouns, and less communication engagement. Lin et al. [46] proposed the SenseMood system, which uses the pre-trained Bert and CNN models to extract the features of pictures and texts posted by users, respectively, and then uses a multimodal

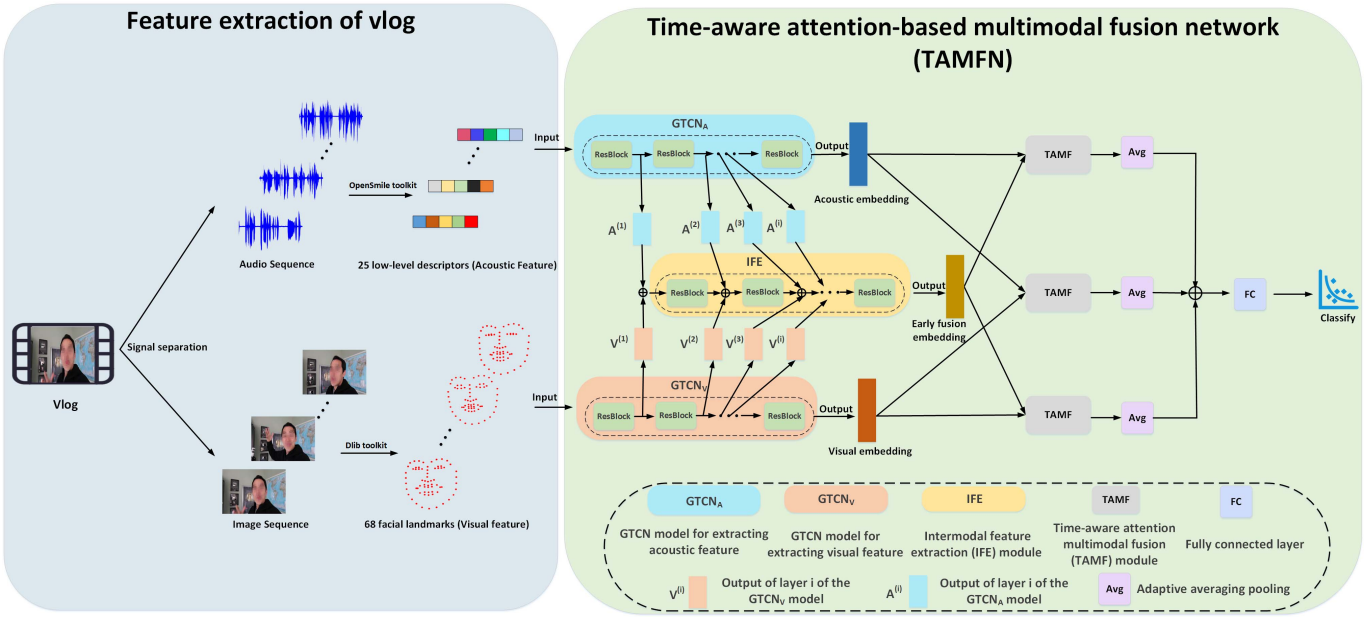


Fig. 2. The detailed structure of TAMFN.

feature learning method to integrate the features of images and texts. Hussain et al. [47] analyzed whether the user suffered from depression by mining the user update status of the user account, the content of the pages the user liked, and the content of the groups the user joined. Mann et al. [48] extracted image and text features through ResNet and ELMo, respectively, fused the two modalities' information through a fusion module and found that the model combination had the best performance. Shatte et al. [49] used the behavior, emotion, language style, and discussion topics of posts on Reddit to describe the risk of developing postpartum depression. Zogan et al. [50] proposed the depression-net model to detect Depression by combining user behavior, release history, and activity. Safa et al. [51] proposed a method for automatic collection of self-reported statements and evaluation of postings, and proposed a multimodal framework to process textual and visual features to describe the user's depression level in a lexicographical way. Cheng et al. [52] used the T-LSTM model to analyze the importance of each user's post for depression detection by taking the text, image, and time posted by the user as input features.

III. Proposed Method

To improve the depression detection ability of vlog data on social media, we proposed the TAMFN model. The structure of TAMFN is shown in Fig. 2. TAMFN model is mainly composed of three core modules: GTCN, IFE, and TAMF. The GTCN model is used to extract acoustic and visual features, the IFE module integrates early acoustic and visual interaction features, and the TAMF module fuses multiple modal features through time-aware attention. The detailed structure of each module is described below.

A. GTCN Model

The temporal convolutional network (TCN) [53] is a structure based on the convolutional neural network, which can effectively extract features for time series and avoid gradient disappearance or gradient explosion. Dilated convolution in TCN allows the convolution kernel to sample at intervals and adjusts the receptive field of the convolution kernel by changing the expansion coefficient so the TCN model can flexibly receive historical information. Causal convolution can make the information of the previous layer of the model closely related to the knowledge of the next layer. The transmission of information is strictly one-way, and the output of each moment is only affected by the previous historical information. However, TCN is based on the structure of the convolutional network. The receptive field of the convolution kernel limits the range of feature extraction at each step of the TCN model, so the TCN model fails to capture the time series' global information. Furthermore, we propose a temporal convolutional network with the global information (GTCN), which introduces a branch of global information extraction into the residual block of TCN to supplement the global information to TCN.

Fig. 3 shows the structure of the GTCN model and the residual block of the GTCN model. The residual block consists of three branches. This module is composed of three branches. GTCN retains two branches of TCN model – branch 1 and branch 3, and adds branch 2 on the basis of TCN. Since the dilated causal convolution in branch 1 extracts features in a sliding window, the range of feature extraction in each step is limited by the size of the convolution kernel, leading to a lack of global information in TCN. Therefore, we introduce branch 2 to extract global information. In the first layer of branch 2, AdaptiveAvgPool is used to map the dimension (N, C, L) of time series to (N, C, 1), and the global feature representation is obtained by aggregating the time axis information of the sequence, where N represents Batch, C represents the number of channels, and L is the time series length. In the case of underfitting due to high data complexity in the D-Vlog dataset, we therefore do not use the

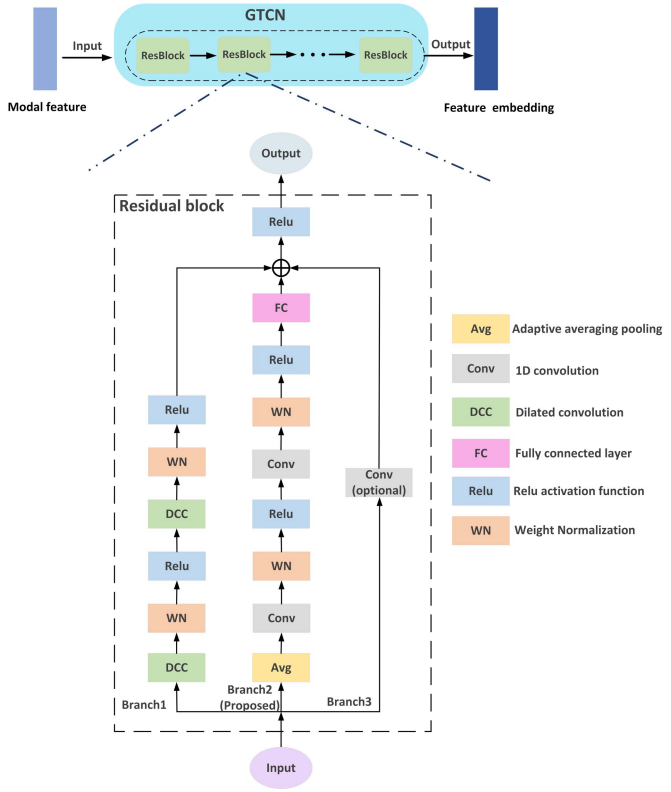


Fig. 3. The detailed illustration of the GTCN model (top) and the residual block of the GTCN model (bottom).

dropout regularization technique here. The feature extraction steps for the GTCN residual blocks are specified as follows:

1) For the input sequence F_{Input}, branch 1 adopts two dilated convolution layers to extract features. The first dilated convolution layer DCC_1 extracts the first-level feature M_1, and the second dilated convolution layer DCC_2 extracts the second-level feature M_2.

$$M_1 = Relu(WeightNorm(DCC_1(F_{Input}))) \quad (1)$$
$$M_2 = Relu(WeightNorm(DCC_2(M_1))) \quad (2)$$

2) Branch 2 uses a convolutional layer with an expansion coefficient of 1. First, the information on the time series is aggregated through adaptive averaging pooling $AdaptiveAvgPool$, and then the global information is extracted through two convolutional layers with the size of convolution kernel 1 to obtain M_5. Finally, the time dimension of feature M_5 is extended to be consistent with the input F_{Input} time dimension through the fully connected layer FC.

$$M_3 = AdaptiveAvgPool(F_{Input}) \quad (3)$$
$$M_4 = Relu(WeightNorm(Conv_1(M_3))) \quad (4)$$
$$M_5 = Relu(WeightNorm(Conv_2(M_4))) \quad (5)$$
$$M_6 = FC(M_5) \quad (6)$$

3) With greater network depth, the GTCN network can capture longer temporal dependencies. To maintain the stability of the network, identity mapping, namely branch 3, is added. If the channel dimension $Channel_{In}$ of the input sequence is inconsistent with the channel dimension $Channel_{Out}$ of the output sequence, one-dimensional convolution $Conv_3$ with convolution kernel size 1 is used for dimensional transformation. If they are consistent, the convolution operation is not selected.

$$M_7 = \begin{cases} F_{Input}, & if\ Channel_{In} == Channel_{Out} \\ Cov_3(F_{Input}), & esle \end{cases} \quad (7)$$

4) By adding the outputs of branch 1, branch 2 and branch 3, and then by Relu activation function, the final residual block output F_{Output} can be obtained.

$$F_{Output} = Relu(M_2 + M_6 + M_7) \quad (8)$$

Facing the problem of depression detection based on vlog, we need to fully use local context-dependent and global information to evaluate the depression state more reasonably and accurately. The GTCN model can aggregate the historical knowledge of a long time span through multiple layers of dilated causal convolution and model the information of different time segments. At the same time, the GTCN model also supplements the global information and improves the depression detection ability of the model.

B. IFE Module

Most of the current multimodal fusion studies usually use multiple encoders to extract multiple modal features separately, and then fuse them based on the extracted high-level semantic features [54], [55], [56]. These studies lack the interactive modeling of the features of each phase in the encoder of each modality and cannot fully exploit the complementary relationship between the modal information. Therefore, we propose an inter-modal feature extraction (IFE) module that fuses the semantic features of each phase in the encoders of multiple modalities to further mine the inter-modal interaction information.

The intermediate module in Fig. 4 is the IFE module. The IFE model is also a GTCN model in essence, but the input of this module consists of two GTCN models extracting acoustic and visual multi-stage features, respectively. The input features of the first layer of the IFE module are acoustic and visual features of the first layer extracted by GTCN. The input features of the other stages of the IFE module are features extracted from the previous layer of IFE, acoustic characteristics of the same stage, and visual characteristics of the same stage. The specific expression of the process is as follows:

$$IFE^{(i)}_{Input} = \begin{cases} A^{(1)} + V^{(1)}, & if \quad i == 1 \\ A^{(i)} + V^{(i)} + IFE^{(i-1)}_{Output}, & esle \end{cases} \quad (9)$$

Among them, $IFE^{(i)}_{Input}$ represents the input of the i-th layer of the IFE module; $A^{(i)}$ and $V^{(i)}$ and describe the acoustic and visual features of the i-th layer extracted by GTCN, respectively; $IFE^{(i-1)}_{Output}$ is denoted as the output feature of the i-1th layer of the IFE module.

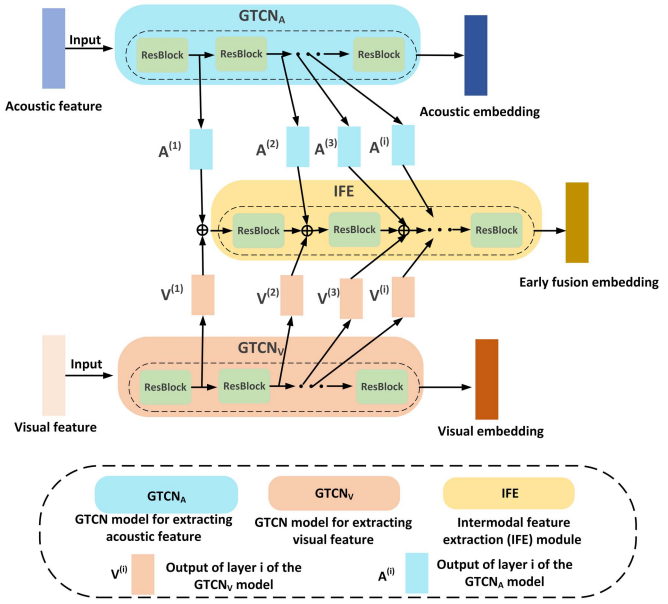


Fig. 4. The structure of the IFE module.

C. TAMF Module

In the vlog-based depression detection task, the contribution of different modalities at different times to depression detection is often different. To obtain more efficient multimodal fusion characteristics, we propose a time-aware attention multimodal fusion (TAMF) module, which can model the interaction between different modalities through the temporal importance between modalities and then guide the fusion of different modalities. The TAMFN model extracts a total of three modalities using two modules, GTCN and IFE, which are acoustic feature, visual feature and early fusion feature. Based on this, the modality pairs with different combinations of these three modalities are fused by the TAMF module respectively. The structure of the TAMF module is shown in Fig. 5. The TAMF module consists of three main steps: temporal feature extraction for each modality, mixed attention vector extraction, and modal pair feature fusion. The module first extracts the long-term context dependencies of each modality using LSTM, and then further extracts the temporal vectors of each modality. After that, the sparse MLP [57] is used to mix the information of the temporal importance of the two modalities to obtain the attention vector with interaction information. Finally, the attention vector with interaction information is used to guide multimodal feature fusion. For the fusion of a pair of modalities, the processing flow of the TAMF module is as follows:

1) For any two modalities F_1 and M_1, use two LSTM models with a layer number of 1 to learn the timing information of the two modalities respectively and concatenate the outputs of different modalities at each moment to obtain the features F_2 and M_2. By using the adaptive averaging pooling $AdaptiveAvgPool$, the information of hidden layer representation dimension of F_2 and M_2 features is aggregated; that is, the time dimension is retained, and the temporal feature representation F_3 and M_3 are obtained.

$$F_2 = LSTM_1(F_1) \tag{10}$$
$$M_2 = LSTM_2(M_1) \tag{11}$$
$$F_3 = AdaptiveAvgPool(F_2) \tag{12}$$
$$M_3 = AdaptiveAvgPool(M_2) \tag{13}$$

2) The temporal features of the two modalities are concatenated to obtain the concatenated vector $Concat\_vector$. To interact with the timing information of the two modalities, we mix the information in the vertical and horizontal directions of the concatenated vector $Concat\_vector$ through weight sharing and sparse connection in the sparse MLP [57], respectively, to obtain the mixed attention vector x_mix.

$$Concat\_vector = Concat(F_3, M_3) \tag{14}$$
$$x\_v = proj\_v(Concat\_vector) \tag{15}$$
$$x\_mix = proj\_h(x\_v) \tag{16}$$

3) The mixed attention vector x_mix are separated, and the temporal attention vectors F_4 and M_4 with interactive information of two modalities are obtained, respectively. Then the temporal attention vectors F_4 and M_4 are applied to modal features F_1 and M_1 by element-wise multiplication. Then the two weighted modal features are added to obtain the final fusion feature $Fusion$.

$$F_4, M_4 = Split(x\_mix) \tag{17}$$
$$Fusion = F_4 \times F_1 + M_4 \times M_1 \tag{18}$$

The TAMF module differs from the method that only focuses on the time series importance modeling on a single modality. TAMF models the importance of different modalities at different times from a global perspective, fully considering the interaction of different modalities. The TAMF module rationally and effectively fuses features from different modalities through interactive attention vectors.

IV. Experimental Setup

A. Dataset

The vlog data used was derived from the D-Vlog dataset [58] collected from YouTube, which contains 961 videos from 816 different people, including 322 males and 639 females, and 555 depression and 406 non-depression individuals. The source data for the D-vlog was obtained from YouTube, and Yoon et al. [58] analyzed posted videos between January 1, 2020 and January 31, 2021 to collect vlogs based on keywords. the keywords for the depression vlog were ‘depression daily vlog’, ‘depression journey’, ‘depression vlog’, ‘depression episode vlog’, ‘depression video diary’, ‘my depression diary’, and ‘my depression story’, while the keywords for non-depression vlogs are ‘daily vlog’, ‘grwm (get ready with me) vlog’, ‘haul vlog’, ‘how to vlog’, ‘day of vlog’, ‘talking vlog’ and so on. To effectively label downloaded videos, Yoon performed two tasks. First, check to see if the downloaded video is in “vlog” format, that is, if a person is speaking directly to the camera. Videos that are

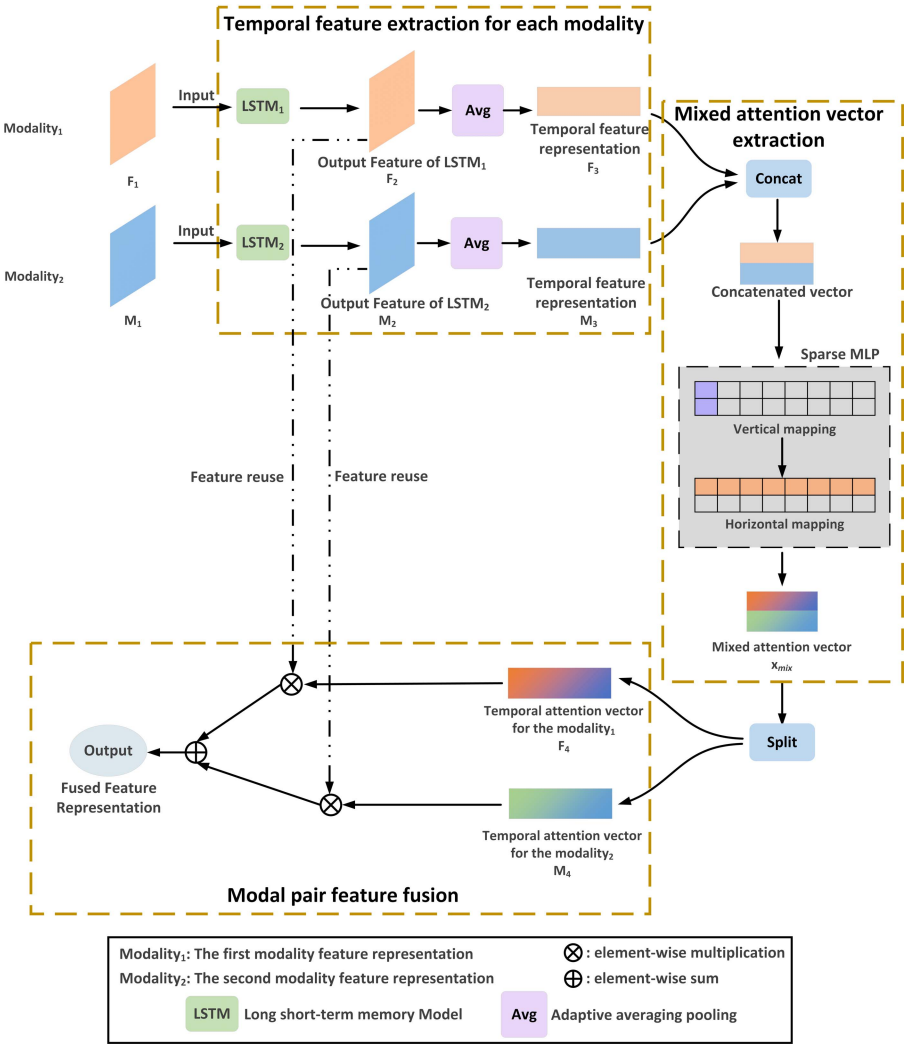


Fig. 5. The structure of the TAMF module.

TABLE I
SAMPLE ALLOCATION OF TRAIN, VALIDATION AND TEST SET OF D-VLOG DATASET

Gender	Train	Val	Test
Male	216	40	66
Female	431	62	146

not in "vlog" format (for example, a group of people or no faces on the video) will be deleted. Second, videos with automatically generated text were watched and analyzed to determine if the speakers in the vlogs were depressed. After the above two tasks, the labeled vlogs is obtained, and Yoon et al. [58] splited the dataset into the train, validation, and test sets with a 7:1:2 ratio, and the specific allocation is shown in Table I. To avoid leakage of privacy of vlog photographers, D-Vlog provides two non-intuitive features, low-level descriptor features extracted by OpenSmile [59] and facial landmark features extracted by Dlib [60].

B. Settings

This paper uses the pytorch framework [61] to implement our model, and models in this paper run on on NVIDIA PCIE A100 graphics card with 40G memory. In this paper, the Adam optimizer [62] is used to optimize the weight update of the model. The learning rate, weight decay and eps of the Adam optimizer are set to 1e-4, 5e-4 and 1e-8, respectively. We optimize the model parameters based on the validation set and use the weighted average f1 score as the evaluation index of the validation set. The weighted averaged f1 score is calculated by taking the mean of all per-class f1 scores while considering each class's support, where support refers to the number of actual occurrences of the class in the dataset. In addition, the weighted average precision and recall are calculated in a similar way to the weighted average F1 score. At the same time, the early stop mechanism is used to avoid over-fitting, and the patience parameter is set to 4. The acoustic and visual feature representation length in vlog is 596, the batch size is set to 32, and the epoch is set to 30. The model's performance is comprehensively evaluated for the testing phase by the weighted average precision, recall, and f1 score. In addition, the number of GTCN layers in TAMFN model is 5, and the number of channels in each layer is 128. The number of layers of LSTM model used to learn the temporal characteristics of each modality is 1.

TABLE II
PERFORMANCE COMPARISONS BETWEEN BENCHMARK MODELS AND THE PROPOSED MODEL

Model	Precision ($\times 10^{-2}$)	Recall ($\times 10^{-2}$)	F1-Score ($\times 10^{-2}$)
LR	54.86	54.72	54.78
SVM	53.10	55.19	52.97
RF	57.69	58.49	57.84
KNN-Fusion	57.86	59.43	54.25
BLSTM	60.81	61.79	59.70
TFN	61.39	62.26	61.00
Fusion_Concat	62.51	63.21	61.10
Fusion_Add	59.11	60.38	58.11
Fusion_Multiply	63.48	64.15	63.09
Depression Detector	65.40	65.57	63.50
TAMFN (proposed)	**66.02**	**66.50**	**65.82**

V. EXPERIMENTAL RESULTS

A. Compared With the State-of-the-Art Models

To verify the effectiveness of the TAMFN model, we compared all the benchmark models proposed by Yoon et al. [58]. Table II shows the performance of all the benchmark models and the TAMFN model on the D-Vlog dataset. The weighted average precision, recall, and f1 of the TAMFN model reached 66.02 ($\times 10^{-2}$), 66.50 ($\times 10^{-2}$), and 65.82 ($\times 10^{-2}$), respectively. It can be seen that the performance of the TAMFN model is better than that of all the benchmark models. The benchmark model is divided into traditional machine learning and deep learning. The machine learning models include Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors based Fusion (KNN-Fusion). The traditional machine learning inputs are features flattened and concatenated by acoustic and visual features. These four machine learning models perform poorly, indicating that the acoustic and visual features in vlog data are highly complex and difficult to distinguish. Deep learning models are Bi-directional LSTM (BLSTM), Tensor Fusion Network (TFN), and Depression Detector. Yoon et al. [58] used Concat, Add and Multiply feature fusion methods to replace the multimodal Transformer encoder in Depression Detector and added three more benchmark fusion models. Depression Detector stands out among benchmark models among deep learning models due to its powerful ability to capture multimodal data. However, our model performed better than Depression Detector. Compared with the Depression Detector model, the TAMFN model improved the weighted average precision, recall, and f1 score by 0.94%, 1.4%, and 3.6%, respectively. From the performance of the comparison experiment, it can be observed that the proposed TAMFN model can effectively evaluate the depression state in the vlog-based data, which indicates that the TAMFN model has strong ability to extract and fuse the temporal information.

B. Ablation Study

In the previous subsection, our proposed TAMFN model achieved the best performance on the D-Vlog dataset compared to the current models. In this subsection, we will explore the impact of each sub-module in TAMFN on the model's depression detection ability. We designed five model structures to explore the importance of each module in TAMFN: (i) Temporal Convolutional Network (TCN). We used two TCN model structures with five layers and 128 channels to extract features, then concatenate all the features by channel, and finally Ordinary Fusion Module (OFM) is used for depression detection. The OFM model is composed of a convolution kernel with size 1 and number of output channels 1 combined with a fully connected layer; (ii) TCN+TAMF, the model uses TCN The model extracts multimodal features, and combines the multimodal features with the TAMF module to detect depression; (iii) GTCN, this model uses the GTCN model to extract features, then concatenates the features, and finally uses the OFM structure for depression detection; (iv) GTCN+IFE, this model uses GTCN as the benchmark model to build a three-stream network structure extraction model, concatenate all features by channel, and then uses OFM module for depression detection; (v) TAMFN, a model composed of three modules GTCN, IFE, and TAMF.

Table IV describes the depression detection performance of five different structural models. TCN has poor depression detection performance, possibly due to the sparsity of dilated causal convolution. Although the receptive field is expanded, it leads to the loss of continuous neighborhood information. Next, we investigate the role of each module in the TAMFN network by comparing each two model pairs, i.e., TCN with TCN+TAMF, TCN with GTCN, and GTCN with GTCN+IFE. As can be seen in Table IV, the TCN+TAMF, GTCN, and GTCN+IFE models all show some degree of improvement in performance over their respective control groups, indicating the effectiveness of the TAMF, GTCN, and IFE modules. For the TAMFN model, GTCN and IFE modules, as feature encoders, can effectively extract the features of a single modality and the interaction between modalities, respectively. On the basis of obtaining effective feature representation and combining with useful TAMF fusion module, the depression detection performance of TAMFN model is greatly improved.

C. Feature Analysis in TAMFN Model

The previous section's ablation experiment shows that the TAMF module's introduction has dramatically improved the depression detection performance of the TAMFN model. In this section, we explore the changes in practical information contained in the features before and after the fusion of different modalities. We use the Principal Component Analysis (PCA)

TABLE III
CLUSTERING EFFECT EVALUATION OF DIFFERENT FEATURES

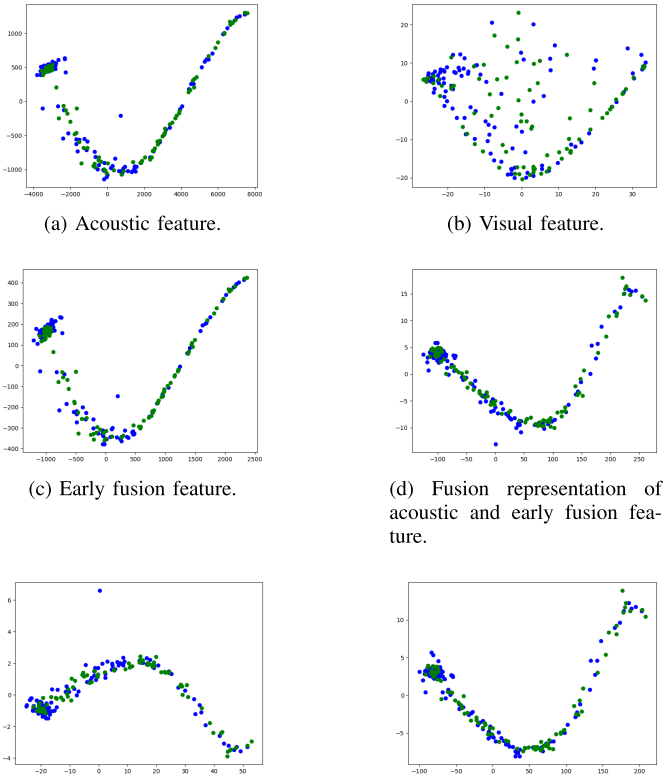
Feature	Silhouette coefficient ($\times 10^{-2}$)	Davies bouldin score
Visual	2.49	6.65
Acoustic	7.68	3.59
Early fusion	7.67	3.56
Acoustic + Early fusion	7.85	3.47
Visual + Early fusion	7.92	3.41
Visual + Acoustic	7.81	3.49

TABLE IV
PERFORMANCE OF DIFFERENT STRUCTURAL MODELS ON D-VLOG DATASET

Model	Precision ($\times 10^{-2}$)	Recall ($\times 10^{-2}$)	F1-Score ($\times 10^{-2}$)
TCN	54.30	54.71	54.46
TCN+TAMF	57.84	59.43	56.12
GTCN	57.24	58.96	55.75
GTCN+IFE	58.49	58.49	58.49
TAMFN	**66.02**	**66.50**	**65.82**

method to reduce high-dimensional feature representation in TAMF to two-dimensional representation and visualize them. Fig. 6 shows the visual results of visual feature, acoustic feature, early fusion feature, fusion representation of visual and acoustic feature, fusion representation of acoustic and early fusion feature, and fusion representation of visual and early fusion feature. We found that the features of depression and healthy people are challenging to distinguish after dimensionality reduction. The reason is that depression detection based on vlog data is difficult, which can be seen from the unsatisfactory performance of the ten benchmark models provided by Yoon et al. [58].

Further, we try to evaluate the quality of the features extracted by each submodule in the TAMFN model using two clustering evaluation metrics, Silhouette coefficient [63] and Davies bouldin score [64]. Silhouette Coefficient combines intra-cluster distance and nearest-cluster distance to evaluate the clustering results of extracted features, and the higher the value, the higher the quality of the features. The davies bouldin score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances, and the lower value indicates the better quality of the extracted features. In addition, we leverage the trained TAMFN model to extract features from the test set of D-Vlog. The silhouette coefficient and davies bouldin score of the features extracted by each submodule of the TAMFN model are shown in Table III. It can be seen that the value of silhouette coefficient of the extracted acoustic features is higher than that of visual features, and the davies bouldin score is lower than that of visual features, which indicates that the quality of acoustic features is better than that of visual features. Moreover, the silhouette coefficient and davies bouldin score of the early fusion features extracted by the IFE module are close to those of the acoustic features, indicating that the IFE module is effective. Then, the value of the silhouette coefficient of the fusion features (Acoustic + Early fusion, Visual + Early fusion and Acoustic + Visual) obtained by the corresponding



(a) Acoustic feature. (b) Visual feature. (c) Early fusion feature. (d) Fusion representation of acoustic and early fusion feature. (e) Fusion representation of visual and early fusion feature. (f) Fusion representation of visual and acoustic feature.

Fig. 6. Feature dimensionality reduction visualization of different modules in TAMFN. In the figures, the blue points denote the dimensionally reduced characteristics of depressed people, and the green points denote the dimensionally reduced characteristics of healthy people.

TAMF module increases to some extent and the davies bouldin score decreases to some extent compared to the original features (Visual, Acoustic and Early fusion). For example, the value of the silhouette coefficient for Visual + Early fusion feature increases and the davies bouldin score decreases compared to the Visual feature. These results demonstrate that the TAMF module effectively extracts the interaction information between the modalities and improves the quality of the fused features.

D. Performance of TAMFN Model for Depression Detection Task in Other Scene

To further evaluate the effectiveness of the TAMFN model, we conduct an extended experiment based on the EATD-CORPUS dataset, which is not the same scene as vlog. the EATD-CORPUS dataset is constructed by Shen et al. [56],

TABLE V
RESULTS OF EXPERIMENTS ON EATD-CORPUS

Features	Models	F1 Score	Recall	Precision
Audio	Multi-modal LSTM [65]	0.49	0.56	0.44
	SVM	0.46	0.41	0.54
	RF	0.50	0.53	0.48
	Decision Tree	0.45	0.44	0.47
	GRU model [56]	0.66	0.78	0.57
Text	Multi-modal LSTM [65]	0.57	0.63	0.53
	SVM	0.64	**1.00**	0.48
	RF	0.57	0.53	0.61
	Decision Tree	0.49	0.43	0.59
	BiLSTM model [56]	0.65	0.66	0.65
Fusion	Multi-modal LSTM [65]	0.57	0.67	0.49
	GRU/BILSTM-BASED Model [56]	0.71	0.84	0.62
	TAMFN (proposed)	**0.75**	0.85	**0.69**

who develop an app that uses a virtual interviewer to ask three questions to interviewees at random while collecting their audio responses. The interviewees can submit their corresponding data online, and in addition, the interviewees have to complete an SDS questionnaire [66], whose scores indicate depression severity. Currently, they have collected information from 162 interviewees. By assessing the scores of the SDS questionnaire, 30 interviewees are considered depressed and 132 interviewees are non-depressed. Shen et al. [56] obtain the audio data of the interviewees and performed data preprocessing, first, removing the muted audio, the audio less than 1 second, and the muted segments at the beginning and end of each recording. Then background noise is removed using RNNoise [67] with default parameters. After that, Kaldi [68] is used to extract texts from the audio. Finally, all texts are manually checked and corrected.

Shen et al. [56] evaluate the model performance on EATD-CORPUS dataset by three-fold cross-validation, taking F1 Score, Recall and Precision as evaluation indicators. For the data imbalance problem, they expanded the depressed dataset by resampling method, i.e., reordering the three responses and resampling these reordered responses to create new training samples. Because there are 6 ways of response rearrangement for each individual, the size of the depressed class can be enlarged 6 times. We use NetVLAD [69] and ELMo [70] to extract the features of audio and text respectively, and then use the features of audio and text as the input of TAMFN model. The parameters of TAMFN model are consistent with the parameter settings shown in Subsection IV, Part B. In addition, the average result of three-fold cross-validation of TAMFN model for ten times on EATD-CORPUS dataset are reported.

Table V shows the performance of different models on the EATD-CORPUS dataset. It can be seen that among the unimodal model performances, the GRU/BILSTM-BASED model proposed by Shen et al. [56] achieves the best performance in comparison to other models on audio and text features, respectively. Although the SVM achieves a recall value of 1.00 for text-based feature prediction, the precision value is low, i.e., depression detection performs poorly. However, in the depression detection based on Fusion feature, our proposed TAMFN model outperforms the GRU/BILSTM-BASED model [56] and the Multi-modal LSTM model [65]. In particular, the precision of the TAMFN model reaches 0.69, which is 11.2% higher than the GRU/BILSTM-BASED model, indicating that our proposed TAMFN model has better depression detection ability.

VI. CONCLUSION

In this paper, we propose a time-aware attention multimodal fusion network (TAMFN). The TAMFN model consists of three modules, GTCN, IFE and TAMF. The GTCN module is used for temporal feature extraction of each modality, the IFE module is used to extract acoustic and visual interaction features, and the TAMF module guides the fusion of multiple modal features through the time-aware attention mechanism. The ablation experiments show that the three submodules GTCN, IFE and TAMF in the TAMFN model have a positive impact on the generalization ability of the TAMFN model, verifying the effectiveness of our proposed submodules. Meanwhile, we conducted experiments on the D-Vlog dataset, and the TAMFN model achieved the best performance compared to all benchmark models. In addition, we conducted an extended experiment to evaluate the performance of the TAMFN model on the EATD-CORPUS dataset, which is not the same scene as vlog. The results show that the TAMFN model also has great performance of depression detection.

Due to the high complexity and noise of non-verbal features in vlog data, the data distribution of the training set and test set may be quite different, making it difficult for current methods to achieve satisfactory depression detection results. In future work, we will introduce a test-time adaptation (TTA) technique, which aims to utilize the unlabeled test set for inverse computation, which allows the model's weights to be updated during the test phase. The variation of the potential distribution between the training set and the test set can be overcome by the TTA technique, and the TTA technique has certain application prospects for the vlog-based depression detection task.

REFERENCES

[1] A. A. Moustafa, R. Tindle, D. Frydecka, and B. Misiak, "Impulsivity and its relationship with anxiety, depression and stress," *Comprehensive Psychiatry*, vol. 74, pp. 173–179, Apr. 2017.
[2] *Depression and Other Common Mental Disorders: Global Health Estimates*, World Health Org., Geneva, Switzerland, Tech. Rep., 2017.
[3] Y. Huang et al., "Prevalence of mental disorders in China: A cross-sectional epidemiological study," *Lancet Psychiatry*, vol. 6, no. 3, pp. 211–224, 2019.

[4] D. M. Maurer, "Screening for depression," *Amer. Family Physician*, vol. 85, no. 2, pp. 139–144, 2012.

[5] H. A. Pincus et al., "The societal costs of chronic major depression," *J. Clin. Psychiatry*, vol. 62, pp. 5–9, Jan. 2001.

[6] C. F. Reynolds et al., "Early intervention to reduce the global health and economic burden of major depression in older adults," *Annu. Rev. Public Health*, vol. 33, no. 1, pp. 123–135, Apr. 2012.

[7] F. Edition et al., "Diagnostic and statistical manual of mental disorders," *American Psychiatric Association*, vol. 21, no. 21, pp. 591–643, 2013.

[8] M. Hamilton, "The Hamilton rating scale for depression," in *Assessment of Depression*. Cham, Switzerland: Springer, 1986, pp. 143–152.

[9] A. T. Beck, R. A. Steer, and G. Brown, "Beck depression inventory—II," *Psychol. Assessment*, Jan. 1996.

[10] K. Kroenke and R. L. Spitzer, "The PHQ-9: A new depression diagnostic and severity measure," *Psychiatric Ann.*, vol. 32, no. 5, pp. 509–515, 2002.

[11] C.-F. Yen, C.-C. Chen, Y. Lee, T.-C. Tang, C.-H. Ko, and J.-Y. Yen, "Association between quality of life and self-stigma, insight, and adverse effects of medication in patients with depressive disorders," *Depression Anxiety*, vol. 26, no. 11, pp. 1033–1039, Nov. 2009.

[12] P. S. Wang et al., "Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys," *Lancet*, vol. 370, no. 9590, pp. 841–850, Sep. 2007.

[13] T. F. Bishop, J. K. Seirup, H. A. Pincus, and J. S. Ross, "Population of U.S. Practicing psychiatrists declined, 2003–13, which may help explain poor access to mental health care," *Health Affairs*, vol. 35, no. 7, pp. 1271–1277, Jul. 2016.

[14] C. Koch, M. Wilhelm, S. Salzmann, W. Rief, and F. Euteneuer, "A meta-analysis of heart rate variability in major depression," *Psychol. Med.*, vol. 49, no. 12, pp. 1948–1957, Sep. 2019.

[15] D. Kuang et al., "Depression recognition according to heart rate variability using Bayesian networks," *J. Psychiatric Res.*, vol. 95, pp. 282–287, Dec. 2017.

[16] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level fusion approaches based on multimodal EEG data for depression recognition," *Inf. Fusion*, vol. 59, pp. 127–138, Jul. 2020.

[17] I. Kakkos et al., "EEG fingerprints of task-independent mental workload discrimination," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3824–3833, Oct. 2021.

[18] M. N. Naszariahi, K. N. A. Khaleeda, and N. A. M. Mortar, "The development of galvanic skin response for depressed people," in *Proc. Adv. Mater., Eng. Technol.*, 2020, Art. no. 020096.

[19] S. Pal, N. Mishra, P. Singh, S. Sood, and H. Kumar, "Study of galvanic skin response in patients of moderate depression," *Indian J. Health Wellbeing*, vol. 10, nos. 1–3, pp. 29–31, 2019.

[20] J. Xiao et al., "Discriminating poststroke depression from stroke by nuclear magnetic resonance spectroscopy-based metabonomic analysis," *Neuropsychiatric Disease Treatment*, vol. 12, p. 1919, Jun. 2016.

[21] J. Yan, F. Zhang, M. Wang, H. Tang, Q. Hu, and X. Zhang, "Classification of rock-electro parameters of low-permeability sandstone based on nuclear magnetic resonance log and its appication: An example of E_{s4} in south slope of the Dongying depression," *Chin. J. Geophys.*, vol. 62, no. 7, pp. 2748–2758, 2019.

[22] F. Rahimi et al., "The effect of the eye movement desensitization and reprocessing intervention on anxiety and depression among patients undergoing hemodialysis: A randomized controlled trial," *Perspect. Psychiatric Care*, vol. 55, no. 4, pp. 652–660, Oct. 2019.

[23] D. Zhang et al., "Effective differentiation between depressed patients and controls using discriminative eye movement features," *J. Affect. Disorders*, vol. 307, pp. 237–243, Jun. 2022.

[24] T. Wang et al., "A gait assessment framework for depression detection using Kinect sensors," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3260–3270, Feb. 2020.

[25] B. Miao, X. Liu, and T. Zhu, "Automatic mental health identification method based on natural gait pattern," *PsyCh J.*, vol. 10, no. 3, pp. 453–464, Jun. 2021.

[26] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7159–7163.

[27] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103107.

[28] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep neural networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks," *Frontiers Neurosci.*, vol. 15, Apr. 2021, Art. no. 609760.

[29] Y. Xu et al., "Inconsistency-based multi-task cooperative learning for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Aug. 9, 2022, doi: 10.1109/TAFFC.2022.3197414.

[30] S. Ziyadin, R. Doszhan, A. Borodin, A. Omarova, and A. Ilyas, "The role of social media marketing in consumer behaviour," in *Proc. E3S Web Conf.*, vol. 135, 2019, Art. no. 04022.

[31] S. Saha, S. K. Thakur, and R. Ponmagal, "Machine learning approach to detect depression using social media posts," in *Ambient Communications and Computer Systems*. Cham, Switzerland: Springer, 2022, pp. 291–301.

[32] R. M. Ortega-Mendoza, D. I. Hernández-Farías, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "Revealing traces of depression through personal statements analysis in social media," *Artif. Intell. Med.*, vol. 123, Jan. 2022, Art. no. 102202.

[33] M. Hiraga, "Predicting depression for Japanese blog text," in *Proc. ACL, Student Res. Workshop*, 2017, pp. 107–113.

[34] F. Sadeque, D. Xu, and S. Bethard, "UArizona at the CLEF eRisk 2017 pilot task: Linear and recurrent models for early depression detection," in *Proc. CEUR Workshop*, vol. 1866, 2017, pp. 1–15.

[35] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., Keyboard Clinic*, 2018, pp. 88–97.

[36] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y-Gómez, "Detecting depression in social media using fine-grained emotions," in *Proc. Conf. North*, 2019, pp. 1481–1486.

[37] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Syst. Appl.*, vol. 133, pp. 182–197, Nov. 2019.

[38] H. S. Alsagri and M. Ykhlef, "Machine learning-based approach for depression detection in Twitter using content and activity features," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 8, pp. 1825–1832, 2020.

[39] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104499.

[40] J. S. Lara, M. E. Aragón, F. A. González, and M. Montes-y Gómez, "Deep bag-of-sub-emotions for depression detection in social media," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, 2021, pp. 60–72.

[41] M. Agirrezabal and J. Amann, "KUCSTLT-EDI-ACL2022: Detecting signs of depression from social media text," 2022, *arXiv:2204.04481*.

[42] S. Kayalvizhi and D. Thenmozhi, "Data set creation and empirical analysis for detecting signs of depression from social media postings," 2022, *arXiv:2202.03047*.

[43] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. 5th Annu. ACM Web Sci. Conf. (WebSci)*, 2013, pp. 47–56.

[44] G. Shen et al., "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proc. IJCAI*, Aug. 2017, pp. 3838–3844.

[45] N. Vedula and S. Parthasarathy, "Emotional and linguistic cues of depression from social media," in *Proc. Int. Conf. Digit. Health*, Jul. 2017, pp. 127–136.

[46] C. Lin et al., "SenseMood: Depression detection on social media," in *Proc. Int. Conf. multimedia Retr.*, 2020, pp. 407–411.

[47] J. Hussain et al., "Exploring the dominant features of social media for depression detection," *J. Inf. Sci.*, vol. 46, no. 6, pp. 739–759, 2020.

[48] P. Mann, A. Paes, and E. H. Matsushima, "See and read: detecting depression symptoms in higher education students using multimodal social media data," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 440–451.

[49] A. B. R. Shatte, D. M. Hutchinson, M. Fuller-Tyszkiewicz, and S. J. Teague, "Social media markers to identify fathers at risk of postpartum depression: A machine learning approach," *Cyberpsychol., Behav., Social Netw.*, vol. 23, no. 9, pp. 611–618, Sep. 2020.

[50] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: A novel summarization boosted deep framework for depression detection on social media," 2021, *arXiv:2105.10878*.
[51] R. Safa, P. Bayat, and L. Moghtader, "Automatic detection of depression symptoms in Twitter using multimodal analysis," *J. Supercomput.*, vol. 78, no. 4, pp. 4709–4744, Mar. 2022.
[52] J. C. Cheng and A. L. Chen, "Multimodal time-aware attention networks for depression detection," *J. Intell. Inf. Syst.*, pp. 1–21, Apr. 2022.
[53] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 156–165.
[54] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2982–2990.
[55] D. Cao, L. Miao, H. Rong, Z. Qin, and L. Nie, "Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 106114.
[56] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6247–6251.
[57] C. Tang, Y. Zhao, G. Wang, C. Luo, W. Xie, and W. Zeng, "Sparse MLP for image recognition: Is self-attention really necessary?" in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2344–2351.
[58] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022.
[59] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Jun. 2016.
[60] D. E. King, "Dlib-ML: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.
[61] E. Stevens, L. Antiga, and T. Viehmann, *Deep Learning With PyTorch*. Shelter Island, NY, USA: Manning Publications, 2020.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[63] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[64] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[65] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, Sep. 2018, pp. 1716–1720.
[66] W. W. Zung, "A self-rating depression scale," *Arch. Gen. Psychiatry*, vol. 12, no. 12, pp. 63–70, 1965.
[67] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–5.
[68] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," 2014, *arXiv:1410.7455*.
[69] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.
[70] W. Che, Y. Liu, Y. Wang, B. Zheng, and T. Liu, "Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation," 2018, *arXiv:1807.03121*.