DepMSTAT: Multimodal Spatio-Temporal Attentional Transformer for Depression Detection

Yongfeng Tao, *Graduate Student Member, IEEE*, Minqiang Yang, *Member, IEEE*, Huiru Li, *Fellow, IEEE*, Yushan Wu, *Fellow, IEEE*, and Bin Hu, *Fellow, IEEE*

***Abstract*—Depression is one of the most common mental illnesses, but few of the currently proposed in-depth models based on social media data take into account both temporal and spatial information in the data for the detection of depression. In this paper, we present an efficient, low-covariance multimodal integrated spatio-temporal converter framework called DepMSTAT, which aims to detect depression using acoustic and visual features in social media data. The framework consists of four modules: a data preprocessing module, a token generation module, a Spatial-Temporal Attentional Transformer (STAT) module, and a depression classifier module. To efficiently capture spatial and temporal correlations in multimodal social media depression data, a plug-and-play STAT module is proposed. The module is capable of extracting unimodal spatio-temporal features and fusing unimodal information, playing a key role in the analysis of acoustic and visual features in social media data. Through extensive experiments on a depression database (D-Vlog), the method in this paper shows high accuracy (71.53%) in depression detection, achieving a performance that exceeds most models. This work provides a scaffold for studies based on multimodal data that assists in the detection of depression.**

***Index Terms*—Depression detection, spatio-temporal attention, transformer, vlog data.**

I. Introduction

DEPRESSION is one of the most common mental disorders. According to the World Health Organization (WHO), depression is expected to become the world's number one disease burden by 2030 [1]. The lifetime risk of depression is 15% to 18%, and the worst outcome is that many people eventually choose to commit suicide. About 50 per cent of suicides among suicide risk factors can be attributed to depression [2]. Despite efforts to recognise and treat depression, new data suggest that the prevalence of mental disorders may be increasing, particularly among younger people. To help clinicians work more efficiently, the first step in treating depression is depression detection, the process of assessing whether a person has depression or depressive symptoms. It is therefore crucial to develop an automated and intelligent depression detection system.

Depressed people have lower levels of social behaviour [3], such as fewer facial movements, fewer body and hand gestures, less eye contact and a lack of smiling. Current methods of diagnosing depression rely almost exclusively on doctor-patient interaction and scale analysis [4], with the obvious disadvantage that both the doctor's diagnosis and the patient's completion of the scale are subjective in nature. Therefore, an objective and valid method of predicting clinical outcomes in depression is important to improve the detection of depression. Most depression detection research consists mainly of hand-crafted features methods [5], [6], [7] and deep learning-based methods [8], [9]. Compared to gait [10] and physiological signals (electroencephalogram (EEG) [11], electrocardiography (ECG) [12], etc.), facial cues and voice signals are easier to capture. Therefore, the detection of depression through audiovisual [8] information has received considerable attention.

Most previous studies have extracted either semantic or temporal features of the data separately to detect depression, and few studies have considered both features of equal importance. We believe that video streaming data is a stack of semantic information in the temporal dimension, in which facial expression semantic features can be regarded as facial emotions formed by the coordinate position relationship of facial landmarks in a frame. As shown in Fig. 1, the temporal features of the facial expression data express more of the trend of emotional changes, while the semantic features carry more about the mutual information between facial landmarks to convey the facial emotion information at a certain moment.

Previous studies have used only face data [5], [13] or voice data [14] to extract features for depression detection. This line of studies ignored the interaction of information between the two types of data, face and speech. Other studies [15], [16] have

Manuscript received 30 March 2023; revised 9 November 2023; accepted 31 December 2023. Date of publication 5 January 2024; date of current version 10 June 2024. 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 62227807, in part by the Natural Science Foundation of Gansu Province, China, under Grant 22JR5RA488, in part by the Fundamental Research Funds for the Central Universities under Grant lzujbky-2023-16, and in part by Supercomputing Center of Lanzhou University. Recommended for acceptance by Lei Chen. *(Corresponding authors: Minqiang Yang; Bin Hu.)*

This work did not involve human subjects or animals in its research.

Yongfeng Tao, Minqiang Yang, Huiru Li, and Yushan Wu are with the Gansu Provincial Key Laboratory of Wearable Computing School of Information Science and Engineering, Lanzhou University, LanZhou 730000, China (e-mail: taoyf21@lzu.edu.cn; yangmq@lzu.edu.cn; hrli20@lzu.edu.cn; wuysh2021@lzu.edu.cn).

Bin Hu is with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China, and with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China, and also with the TC Co-Chair of Computational psychophysiology in the IEEE Systems, Man, and Cybernetics Society (SMC), the TC Co-Chair of cognitive computing in IEEE SMC, and the Vice-Chair of the TC 9.1. Economic, Business, and Financial Systems on Social Media at the International Federation of Automatic Control (IFAC) 730000, China (e-mail: bh@lzu.edu.cn).

Digital Object Identifier 10.1109/TKDE.2024.3350071

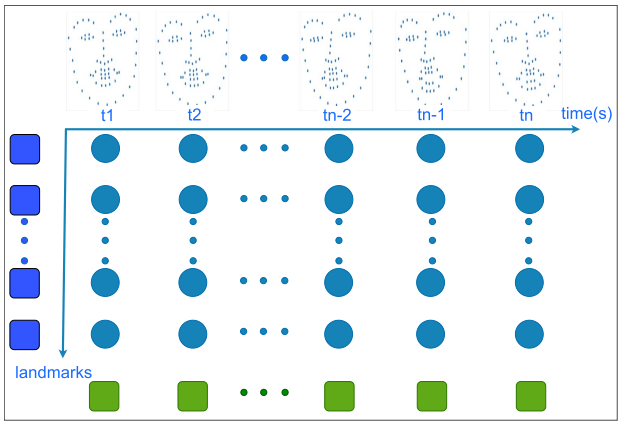


Fig. 1. Simulates the information contained in a sequence of faces. The coordinate relationship between all the facial landmarks in a frame can reflect the momentary semantic information, i.e. the facial emotions at that moment (indicated by blue squares). Changes in individual facial landmarks over time reflect information about changes in facial emotion (indicated by the green squares). The light blue circle in the middle of the diagram represents the facial landmarks data in the video stream.

addressed this issue by using text, speech and facial expressions as inputs to depression detection models. However, little attention has been paid to the semantic information in single modality data. Therefore, how to focus on both the semantic and temporal information of single-modality data to effectively explore these two types of information has become a problem that urgently needs to be solved in depression detection.

In this study, our research is motivated by the use of semantic information in the spatial dimension and temporal information of facial landmarks and speech data in a field environment to improve the accuracy of depression detection models. By combining features from different modalities, the emotional state of individuals can be better understood, facilitating early detection and intervention of depressive symptoms. We aim to further fuse spatio-temporal information from unimodal data in social network data to detect depression. In contrast to previous methods [9], [17], [18], [19], we propose an end-to-end lightweight Multimodal Spatio-Temporal Attention Transformer approach for Depression detection (named: DepMSTAT). The proposed Spatio-Temporal Attentional Transformer (STAT) module can be used to extract and fuse temporal and spatial information from facial and speech data. Specifically, the method first preprocesses the visual and acoustic data from the social media. Second, the Spatial Attention Block (SAB) and the Temporal Attention Block (TAB) in the STAT module are used to extract temporal and spatial features from the data, and these features are fused using the Multimodal Fusion Transformer Block (MTB). Finally, the features extracted from the social media data are fed into a classifier for voting classification. Fig. 2 shows a diagram of the main framework of our approach. The contributions of this paper are three-folds:

- We propose an end-to-end transformer framework for depression detection (DepMSTAT). The framework is able to fully integrate spatio-temporal features of streaming video data. The framework uses a simple voting mechanism for classification and is able to detect depression more effectively.
- We propose a plug-and-play multimodal spatio-temporal fusion attention module (STAT) consisting of a temporal attention block, a spatial attention block, and a multimodal fusion attention block. It captures the global dependencies of temporal and spatial information of visual and acoustic sequences in a video stream and fuses these features using an earlier fusion strategy.
- The DepMSTAT framework surpasses the accuracy of the baseline method on the Vlog dataset and drastically reduces the number of parameters compared to state-of-the-art algorithms.

The remaining sections of the paper proceed as follows. Section II discusses related work in current research. Section III describes the methodology of our work. The results and discussion are presented in Section IV. Section V presents the conclusions and suggestions for future research.

II. Related Work

In this section, we briefly describe the knowledge related to transformers and the evolution of previous research on depression detection based on deep learning.

A. Transformer and Self-Attention

The transformer was first proposed in [20] for machine translation by stacking multi-headed self-attentive and feed-forward MLP layers to capture the long-term relevance of words. The transformer has become increasingly popular in many natural language processing (NLP) tasks [21], [22], [23], due to its powerful performance. Recurrent neural networks have been replaced by transformer for sequential tasks (speech processing [23], [24], computer vision [25]). Transformer has been progressively extended to tasks that deal with non-sequential problems [26], [27]. This is made possible by the key components of Transformer, namely the self-attentive mechanism, the residual connections, and the feed-forward neural network. To facilitate the understanding of this paper, we present the detailed structure of the self-attentive mechanism with residual connections in Fig. 3. In the figure, the self-attentive mechanism looks at and decides the more important parts of the input sequence, thus facilitating the capture of global information from the input sequence. The advantages of residual connections are: alleviation of the gradient disappearance problem and alleviation of the weight matrix degradation problem.

Dosovitskiy et al. proposed the Visual Transformer (VIT) [26], revealing the great potential of transformer-based models for image classification. As a result, transformers soon had a profound impact on many computer vision tasks [28], [29]. The researchers worked on applying transformer networks to spatio-temporal data modelling [30] and multimodal learning tasks [31], [32] by introducing spatio-temporal attention mechanisms and multimodal encoder-decoder designs and achieving more comprehensive and accurate modelling. Ding et al. have proposed a simple yet effective Dual Attention Visual Transformers (DaViT) [33], which uses self-attention mechanisms

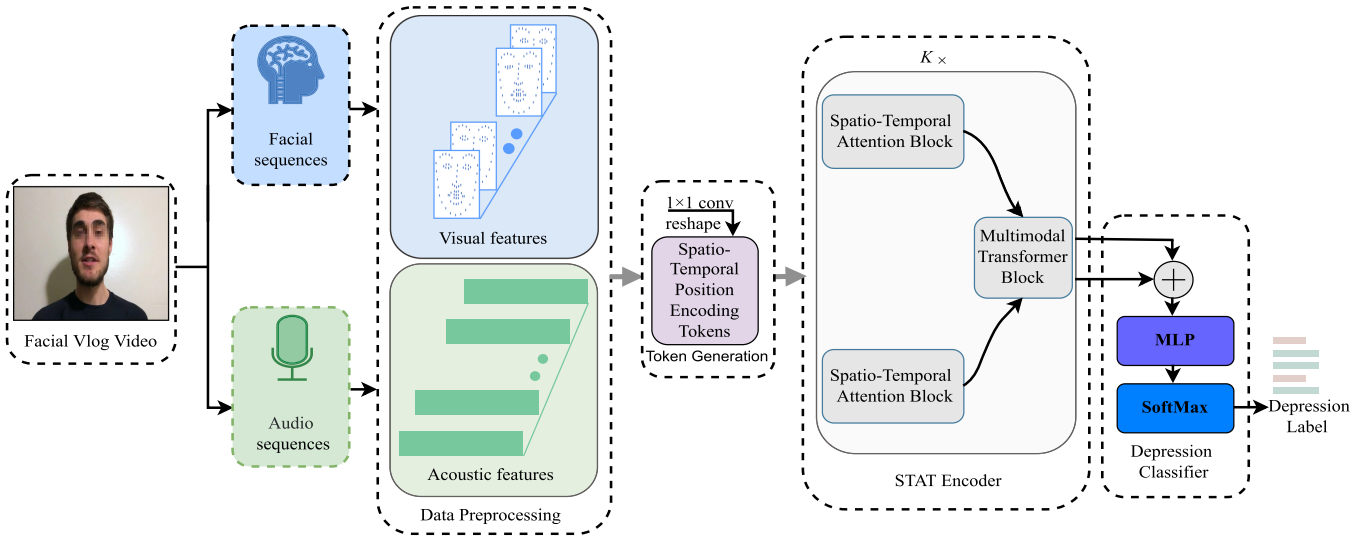


Fig. 2. DepMSTAT model depression detection processing flow. The visual and acoustic features obtained from the video stream are fed into K tandem STAT encoder modules after data preprocessing and data embedding to extract temporal features, spatial features, and fused features. Depression classifier receives the features sequence representation from the STAT encoder and predicts the depression label.

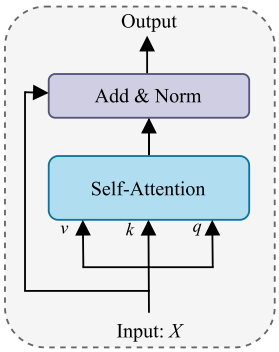


Fig. 3. Detailed structure of self-attention in a residual form.

with spatial and channel tokens to capture global context while maintaining computational efficiency. It has shown excellent performance on image classification tasks. Deep fusion of multimodal data is an important topic that improves performance by using multiple input sources. However, it is a challenge to extract data features efficiently using a single transformer, while ensuring that features between different modalities can be fully fused. Xu et al. systematically reviewed the design and training of transformers in multimodal environments and summarised the main challenges and solutions in the multimodal domain [34]. In recent years, multimodal research based on the Transformer and Attention mechanisms in various fields has contributed significantly to the development of their respective fields, e.g. humour detection [35], product detection [36], data alignment in time series modelling of human behaviour [37], dance choreography [38], and so on. The transformer encoder was first applied in [17] to extract long-term temporal context information from long sequences of audio and visual data on depression. Guo et al. [18] proposed a multimodal depression detection method based on the TOpic Attentive Transformer (TOAT) by introducing a transformer pre-training model. The transformer encoder was introduced directly in [19] to extract features for depression detection. Unlike the aforementioned methods, we aim to build an end-to-end fusion pipeline capable of effectively detecting depression, extracting temporal and spatial features from each type of multimodal data, and employing an early feature fusion strategy.

B. Deep Learning-Based Depression Detection

Current research focuses on applying deep learning [8] to depression detection, which is more effective than hand-crafted [6], [39] feature extraction. Most of the research data on depression detection is not publicly available, and only a few small datasets in laboratory scenarios are publicly available (AVEC2019 [40], etc.). Most of these are based on laboratory scenarios, with a lack of data studies in real-world scenarios. Social media users share their emotions on social media, including users with mental health problems. As a result, some studies [41], [42] are beginning to use social media data for the early detection of depression. The public availability of the large social media dataset D-Vlog [19] provides favourable data to study depression detection in a real-world scenario.

Some researchers have used single-modality data to detect depression. He et al. [43] used a single face image as input to the model and extracted local and global features in the face. To increase the range of receptive field of the convolutional neural network feature map, the face was divided into four regions of different sizes as input to the model in [44]. However, similar studies have ignored the dynamic changes between face sequences. A 3D network model for depression detection with residual connectivity that effectively avoids overfitting problems is developed in [45]. The model consists of a multiscale spatiotemporal network (MSN) to effectively represent facial information related to depressive behavior in videos. However, this

study first requires pre-training on the VGGFace2 [46] dataset. Al Jazaery et al. [47] proposed an RNN-C3D depth model for depression detection that incorporates facial changes and head motion information. The model is able to model local and global spatio-temporal information from successive facial expressions. Due to the limited training data for deep models, some studies have used pre-training for depression detection [48], [49]. There is also research working towards an end-to-end model. He et al. [13] propose an end-to-end deep system capable of extracting high-level features from video frames. The model is a 3D convolutional neural network equipped with the spatiotemporal feature aggregation module (STFAM). An audio-based bipolar disorder detection method has been proposed by Du et al. [50]. The method integrates the inception module with long-term memory (LSTM) in feature sequences and is able to acquire multi-scale temporal information for depression detection.

Some researchers believe that multimodal data contain rich information and that multimodal fusion is better than single-modality data for depression detection. Dongle et al. [51] used pre-trained models to extract deep speaker recognition (SR) and speech emotion recognition (SER) features and combined the complementary information of these two features to capture the difference between voice and emotion. For multimodal data fusion studies, gated recurrent unit (GRU) is used to extract features from text, speech and video data to identify depression [16]. Recently, some studies have introduced transformer encoders to depression recognition tasks [17], [18]. In our previous work [9], by embedding the spatio-temporal features in the matrix V, the attention mechanism was guided to learn the global information efficiently. However, the model has a large number of parameters, which is not very friendly for training on small datasets, so it is necessary to reduce the number of parameters while extracting the spatio-temporal fusion features to improve the stability of the model.

III. Method

In this section, we formulaically define the data sequences and propose a novel framework, i.e. DepMSTAT, to model the multimodal data information of each frame and capture changes in the temporal dimension.

A. Problem Formulation

Let us denote a single-modality data sequence as X_{Sm}. Suppose each sequence X_{Sm} consists of T frames in length with N feature points in each frame, and $X_{Sm} \in \mathbb{R}^{T\times N\times C}$ can be denoted as:

$$X_{Sm} = \{X_1, X_2, \ldots, X_t, \cdots X_T\} \tag{1}$$

where m indicates a single-modality, in this paper, $m \in \{a, v\}$ is used to mark acoustic and visual data sequences, respectively. $X_t = \{x_t^1, x_t^2, \cdots x_t^N\} \in \mathbb{R}^{N\times C}$ indicates the data points of the tth frame in a specific order. Each x denotes a data vector of dimension C for each single-modality. From another point of view, X_{Sm} could be expressed as:

$$X_{Sm} = \{X_1, X_2, \ldots, X_n, \cdots X_N\} \tag{2}$$

where $X_n = \{x_1^n, x_2^n, \cdots x_T^n\} \in \mathbb{R}^{T\times C}$ represents the sequence composed of the nth data point in all frames.

B. Method Framework

As shown in Fig. 2, our model consists of four main functional modules: Data Preprocessing Module, Token Generation Module, Spatio-Temporal Attentional Transformer Encoder (STAT-encoder) and Depression Classifier. The STAT-encoder extracts spatio-temporal information from multimodal data and performs feature fusion. The Depression Classifier receives a sample representation from the STAT-encoder and then predicts the label of depression.

1) Data Preprocessing: Length normalization of social media video data of varying lengths is necessary and serves to increase the amount of data in order to prevent over-fitting of the model. Specifically, the data is cropped to a given length, L. For data of insufficient length L, a polynomial interpolation fitting method is used to interpolate the operation. Following this approach, the number of samples in the training set, validation set, and test sets is increased. Looking at the data durations in the D-Vlog dataset, the longest duration is 3968.59 s and the shortest duration is 23.62 s. Therefore, using too large a value for L will result in a serious loss of information during upsampling for data with short durations. Using too small a value for L will result in segmented data with less time information. Therefore, the sample length L is set to 300 for all data sequences. In our previous work we also used this type of pre-processing [9].

To improve the convergence speed of the model, we perform a max-min normalization on the data, which can be formulated as:

$$X'_{Sm} = \frac{X_{Sm} - \min(X_{Sm})}{\max(X_{Sm}) - \min(X_{Sm})} \tag{3}$$

2) Token Generation: To ensure uniformity of dimensionality in feature fusion, we manipulated the dimensions of visual and speech information as follows:

$$X_{Sm} = \text{Reshape}\,(X'_{Sm}) \tag{4}$$

where Reshape $(\cdot)$ is a reshape operator that makes X_{Sa} and X_{Sv} have the same shape. And we use a two-layer Conv1d to implement Reshape $(\cdot)$.

To enable the transformer encoder to exploit the sequential relationship of the social media data, we append the learnable location-based information of the sequential data, which can be formulated as:

$$X_t = X_t + PE_N, X_n = X_n + PE_T \tag{5}$$

where PE_N is the matrix representing N semantic tokens shared in all frames. X_n means the sequence composed of the same type of data points in all frames. All kinds of data points share the same PE_T, which means all data points in the same frame share the same temporal position encoding. The PE_N and PE_T are trained jointly with the whole model. Following the data representation in Fig. 1, we add the semantic tokens and the temporal position encoding tokens to the input data to form a complete representation of each data point.

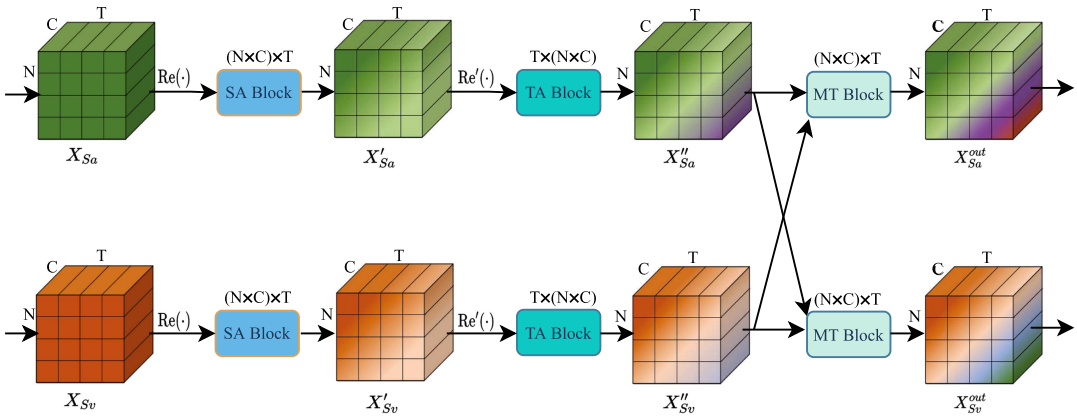


Fig. 4. Main components of the STAT module proposed in this paper are the SAB (SA Block) for acquiring spatial information, the TAB (TA Block) for acquiring temporal dimensional information, and the MTB (MT Block) fusion module for extracting information before the two modes. The illustration shows the structure of STAT and the flow representation of the data. It helps to understand the flow of information in single-modality during spatial-temporal feature extraction and multimodal information fusion.

3) STAT-Encoder: The key plug-and-play STAT module for feature extraction is presented, as shown in Fig. 4. The STAT consists of three blocks: the Spatial Attention Block (SAB), the Temporal Attention Block (TAB), and the Multimodal Fusion Transformer Block (MTB). The SAB is designed to model the facial expression or voice state represented by each data point in each frame. The TAB is designed to capture the changing pattern of the facial expression or voice. The MTB is designed to fuse spatio-temporal information between the two modalities, facial expression and speech. This section first presents the structure of the basic transformer block, and then introduces the STAT module separately.

Basic Transformer Block: Typically, the transformer model is a sequential stack of encoder and decoder blocks. The encoder and decoder have a similar network architecture, but they use different weighting parameters and consist of two sub-layers, the Attention layer and the Position-wise Feed Forward Network (FNN). This work uses the encoder block, hence the shortened name "transformer" for the transformer encoder block below. The attention mechanism is calculated as follows:

$$\mathrm{Att}(X) = \mathrm{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the query matrix, the key matrix and the value matrix respectively. d_k is the key dimensionality. When $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are matrix transformations from the same vector, it is called self-attention; when $\boldsymbol{Q}$ and $\boldsymbol{K}$, $\boldsymbol{V}$ are matrix transformations from different vectors, it is called soft-attention. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are calculated as follows:

$$\boldsymbol{Q} = \phi(X), \boldsymbol{K} = \varphi(X), \boldsymbol{V} = \psi(X) \tag{7}$$

where $\boldsymbol{\phi(X)}$, $\boldsymbol{\varphi(X)}$, and $\boldsymbol{\psi(X)}$ are three different trainable linear transformations. X is a matrix representing all elements.

The FFN is a fully connected feedforward network consisting of a linear transformation of two fully connected layers, where the activation function of the first fully connected layer is the ReLU activation function, which can be formulated as:

$$\mathrm{FFN}(X) = \max\left(0, XW_1 + b_1\right)W_2 + b_2 \tag{8}$$

Thus the transformer is calculated as:

$$X = \mathrm{LN}(\mathrm{LN}(\mathrm{A\,t\,t}(X) + X) + \mathrm{FFN}(X)) \tag{9}$$

where LN is a normalization layer.

STAT-Encoder: Inspired by the self-attention mechanism of "spatial tokens" and "channel tokens" in [33], a STAT module that introduces the attention mechanism of "temporal tokens" into unimodal data to extract effective features is proposed. Our model is based on the core idea of transformer, i.e. the use of the attention mechanism. However, we do not directly adopt the structure of transformer as in [17]. Instead, we have designed a new structure, i.e. STAT, to make it more suitable for multimodal depression detection. There are three subdivisions in STAT: SAB, TAB and MTB.

Modelling semantic information has been an effective way of facilitating depression detection tasks. For example, when a person's facial expression is a smile, the facial landmarks of the mouth and eyebrows move together to form specific semantic information. Therefore, we believe that it is important to capture and model the semantic relevance of the training data. The spatial features of the data are extracted using the SAB module with a self-attention in the form of residuals. Because the output of tanh can produce negative relations and is not restricted to positive values [52], [53]. With more flexibility than softmax, we use tanh as an alternative to softmax. Our attention formula is therefore calculated as follows:

$$\mathrm{Att_s} = \tanh\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{10}$$

Equation (11) and (12) are then used to calculate the spatio features X'_{Sm} of the data X_{Sm}.

$$X^{\mathrm{Re}}_{Sm} = \mathrm{Re}\left(X_{Sm}\right) \tag{11}$$

$$X'_{Sm} = \mathrm{LN}\left(\mathrm{Att_s}\left(X^{Re}_{Sm}\right) + X^{Re}_{Sm}\right) \tag{12}$$

Algorithm 1: Training Scheme of Our Proposed Method.

Input: Multimodal depression dataset $\mathcal{D} = \{(\mathbf{X_{Sm}}, \boldsymbol{y})\}$.
Output: Prediction $\hat{\boldsymbol{y}}$.
1: **for** $i = 1$ to Epoches **do**
2: **for** $k = 1$ to K **do**
3: Capture temporal features: $X'_{Sm} = \mathrm{LN}(\mathrm{Att_s}(\mathrm{Re}(X_{Sm})) + \mathrm{Re}(X_{Sm}))$
4: Capture spatial features: $X''_{Sm} = \mathrm{LN}(\mathrm{Att_s}(\mathrm{Re}'(X'_{Sm})) + \mathrm{Re}'(X'_{Sm}))$
5: Capture spatio-temporal fusion features: $X^{\mathrm{out}}_{Sa} = \mathrm{LN}(\mathrm{L\,N}(\mathrm{MultiHead}(\boldsymbol{Q}_a, \boldsymbol{K}_v, \boldsymbol{V}_v) + X''_{Sa}) + \mathrm{FFN}(X''_{Sa}))$ and X^{out}_{Sv}
6: **end for**
7: Compute the final label: $\hat{\boldsymbol{y}} = \mathrm{MLP}(\mathrm{Concat}(X^{\mathrm{out}}_{Sa}, X^{\mathrm{out}}_{Sv}))$
8: Compute loss L.
9: Backward L and update parameters.
10: **end for**

where $X'_{Sm} \in T \times (N \times C)$. The reshape operator $\mathrm{Re}(\cdot)$ represents the data shape $T \times N \times C$ of X_{Sm} reshaped into $T \times (N \times C)$.

It is worth noting that correlations between frames in the temporal dimension are closely related to the length of the time interval. To better construct complex and uncertain correlations, we have developed our framework TAB, which is a powerful mechanism for capturing remote and proximal correlations in the temporal dimension of the input data. The structure of the TAB is shown in Fig. 3. TAB is used to capture single-modality temporal information by using self-attention in the form of a residual. To reduce the number of parameters in our model and to prevent overfitting of our model, the entire transformer framework is not used to extract single-modality temporal information. To obtain temporal information about the data X'_{Sm}, the shape $T \times (N \times C)$ of X_{Sm} needs to be reshaped into $(N \times C) \times T$ of X_{Sm} using the reshape operator $\mathrm{Re}'(\cdot)$, which can be formulated as:

$$X^{\mathrm{Re}'}_{Sm} = \mathrm{Re}'\left(X'_{Sm}\right) \tag{13}$$

The temporal features of the data $X^{\mathrm{Re}'}_{Sm}$ are then extracted using TAB module with a self-attention in the form of residual, which can be formulated as:

$$X''_{Sm} = \mathrm{LN}\left(\mathrm{Att_s}\left(X^{\mathrm{Re}'}_{Sm}\right) + X^{\mathrm{Re}'}_{Sm}\right) \tag{14}$$

We believe that the data from the two single modalities are complementary. For example, when a person is happy, their facial expressions express positive emotions and their tone of voice is lighter. On the other hand, when the person is sad, their facial expressions show more negative emotions and their tone of voice is more subdued. Therefore, multimodal data can be fused using the cross-attention mechanism used in most studies, which facilitates inter-modal information interaction. The MTB is a multi-headed self-attention transform structure that fuses spatio-temporal information from different modalities and adaptively adjusts the weights of each type of feature. If $m = a$, i.e. the input data, are acoustic sequences, based on the X_{Sa} fusion features, which can be computed as:

$$\begin{aligned} \tilde{X}^{\mathrm{out}}_{Sa} &= \mathrm{L\,N}\left(\mathrm{MultiHead}\left(\boldsymbol{Q}_a, \boldsymbol{K}_v, \boldsymbol{V}_v\right) + X''_{Sa}\right), \\ X^{\mathrm{out}}_{Sa} &= \mathrm{LN}\left(\tilde{X}^{\mathrm{out}}_{Sa} + \mathrm{FFN}\left(X''_{Sa}\right)\right) \end{aligned} \tag{15}$$

where $\mathrm{MultiHead}(\boldsymbol{Q}_a, \boldsymbol{K}_v, \boldsymbol{V}_v)$ [20] is a multi-headed self-attentive mechanism. Similarly, we can calculate the fusion features based on $X^{\mathrm{out}}_{Sv} \in T \times (N \times C)$. In our view, the interplay of the data from the two single-modalities with each other so that the varying degrees of fusion of the two single-modality data through the MTB feature layer can highlight the important features of the single-modalities as shown in Fig. 4.

4) Depression Classifier: For the depression classifier, we use the method in [9], consisting of an MLP layer with two fully connected linear layers and a softmax layer. A subject's data can be divided into many data segments of length L by the data preprocessing method mentioned in Section III-B1. In this way, the L data segments will have L predicted values by the classifier. To ensure fairness of comparison with previous studies, the predicted result with the highest proportion of these L predicted values is used as the predicted label for the subject in this study.

The training process of the entire network is end-to-end, and the loss function used in this study is the focal loss function [54]. The loss can be written as:

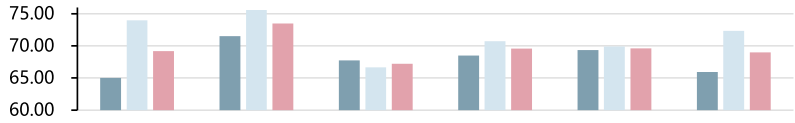
$$L = \begin{cases} -(1-\hat{p})^{\gamma}\log(\hat{p}) & \text{if y} = 1 \\ -\hat{p}^{\gamma}\log(1-\hat{p}) & \text{if y} = 0 \end{cases} \tag{16}$$

where $p_t = \begin{cases} \hat{p} & \text{if y} = 1 \\ 1-\hat{p} & \text{otherwise} \end{cases}$ reflects the proximity to category y. A larger p_t indicates a closer proximity to category y, i.e. a more accurate classification. γ is the adjustable factor. Therefore focal loss function can be written $L = -(1-p_t)^{\gamma}\log(p_t)$. In Algorithm 1, we elaborate the training scheme for DepMSTAT.

IV. Experiment

A. Datasets and Experiment Setting

1) Datasets: Current state-of-the-art methods have been applied to datasets such as AVEC, DAIC-WOZ, etc., however, due to the small sample size of these datasets, we chose Depression Vlog (D-Vlog), which is currently the largest dataset in terms of sample size, to support our study. D-Vlog [19] is a multimodal depression detection dataset for real-world scenarios. This dataset is a balanced dataset that collects multiple vlog videos published by 816 users from YouTube. From these videos, 961 (i.e. approximately 160 hours) vlog video clips were screened for eligibility, including 555 depressive and 406 non-depressive data. To protect user privacy, the dlib [55] and opensmile [56] toolkits are used to extract facial landmarks and acoustic features from the Vlog video data, respectively. The D-Vlog dataset is divided into three parts: train, validation and test sets, as shown in Table 1. We trained our model on the training set, evaluated its performance on the validation set, and tested our model on the test set.



K	1	2	3	4	5	6
Pre(%)	65.00	71.53	67.76	68.50	69.35	65.92
Rec(%)	73.98	75.60	66.66	70.73	69.91	72.35
F1(%)	69.20	73.51	67.21	69.60	69.63	68.99

Fig. 5. Model STAT stacking layer on the value of K for activation selection, where Pre, Rec, and F1 denote Precision, Recall, and F1-Score, respectively.

TABLE I
NUMBER OF SAMPLE DIVISIONS IN THE D-VLOG DATASET [19]

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

2) Implementation Details: DepMSTAT are stacked using 2 Spatio-Temporal Attentional Transformer Blocks, which are composed of four SAB, TAB, and MTB, where the number of attentional heads is set to 4.

Our model is trained on two NVIDIA Tesla V100-PCIE graphics cards of size 32 GB memory. All experiments and code are implemented based on PyTorch [57] framework. We trained our model with a batch size of 64 for a total of 500 epochs; an SGD [58] optimizer with a momentum of 0.5 and a weight decay of 0.001 is used. The initial training learning rate size is set to 0.001 and the learning rate is updated using cosine learning rate decay [59]. To prevent overfitting, we employed dropout in the linear layer, position encoder and transformer encoder, and the dropout ratio [60] is set to 0.5 and flooding [61] is used with a b-value of 0.44. We used the same metrics to evaluate the classification as used in [19], including precision, recall and f1 score.

B. Parameters of the Model

Different parameters in a deep network model can lead to completely different performances, so choosing the right parameters is an important issue. Therefore, we first identified the important parameter in the model, the K-value of the STAT stack number. The other parameters of the model are described before the experiments. The learning rate is 0.001, and the input features for the depression classifier use a combination of different modalities.

To observe the effect of the size of K on the performance of the model with a batch size of 64, we set the search set of K values to {1, 2, 3, 4, 5, 6}, as shown in Fig. 5, and the model performs best when the K value is 2. When K = 3, there is a drop in results, possibly due to changes in the model structure or multimodal data affected by environmental factors. Table II reports the FLOPs and PARAMs for different values of stacking K in this method. This is probably because when the value of K is 1, the model has fewer network layers and is not able to fully fit the features in the data. As the value of K increases, the performance of the model decreases, which may be related to the over-fitting of the model.

TABLE II
FLOPS AND PARAMETERS OF THE PROPOSED METHOD AT DIFFERENT VALUES OF K

K	FLOPs(G)	PARAMS(M)
1	3.61	1.59
2	7.125	2.91
3	10.64	4.22
4	14.16	5.54
5	17.67	6.86
6	21.19	8.18

C. Comparison With Previous Methods

To validate the overall performance of our models, we have compared a wide range of models, FR [62], K-Nearest Neighbors based Fusion (kNN-Fusion) [64], Bi-directional LSTM (BLSTM), Tensor Fusion Network (TFN) [67], Depression Detector [19], Time-aware Attention Multimodal Fusion Network (TAMFN) [70], SeResNet50 [72], SeResNeXt50 [72], InceptionV3 [75], Xception [77], DenseNet201 [78], CondenseNet74-8 [63], EfficientNet-B7 [65], ViT-B/16 [26], DeiT-S/16 [68], A-ViT [69], MTF [71], TinyViT [73], and T-GCN [74]. The two most recent methods, TBOS [76], and STST [9] respectively. FR and KNN-Fusion are traditional machine learning methods, where KNN-Fusion uses decision level fusion methods to fuse information from audio and video to detect depression. BLSTM [66] have been shown to be effective in extracting time-series features from both data modalities to detect depression. TFN is a network structure in end-to-end trained multimodal sentiment analysis. Depression Detector is an acoustic and visual multimodal fusion network designed to detect depressions based on cross-transformer encoders. TAMFN is a time-aware attention-based multimodal fusion depression detection network that fully mines and fuses multimodal features. SeResNet50, SeResNeXt50, InceptionV3, Xception, DenseNet201, CondenseNet74-8, EfficientNet-B7, ViT-B/16, DeiT-S/16, A-ViT, MTF, TinyViT, and T-GCN are some of the classic deep

TABLE III
DVLOG-PERFORMANCE COMPARISONS BETWEEN BASELINE MODELS AND THE PROPOSED MODEL

Models	Precision(%)	Recall(%)	F1-Score(%)	Models	Precision(%)	Recall(%)	F1-Score(%)
FR [62]	57.69	58.49	57.84	CondenseNet74-8 [63]	63.0	62.0	63.0
KNN-Fusion [64]	57.86	59.43	54.25	EfficientNet-B7 [65]	63.0	63.0	63.0
BLSTM [66]	60.81	61.79	59.70	ViT-B/16 [26]	64.0	63.0	63.0
TFN [67]	61.39	62.26	61.00	DeiT-S/16 [68]	64.0	64.0	64.0
Depression Detector [19]	65.40	65.57	63.50	A-ViT [69]	64.9	64.2	64.6
TAMFN [70]	66.02	66.50	65.82	MTF [71]	65.0	64.3	64.7
SeResNet50 [72]	59.0	60.0	60.0	TinyViT [73]	65.1	64.3	64.7
SeResNeXt50 [72]	60.0	61.0	60.0	T-GCN [74]	68.9	65.6	67.3
InceptionV3 [75]	61.0	61.0	61.0	TBOS [76]	65.4	64.7	65.0
Xception [77]	61.0	61.0	61.0	STST [9]	72.50	77.67	75.00
DenseNet201 [78]	62.0	62.0	62.0	Ours(DepMSTAT)	71.53	75.60	73.51

learning methods. TBOS is proposed as a new temporal convolutional converter with knowledge embedding to solve the joint task of depression detection and emotion recognition. STST explores spatio-temporal features in a multimodal depression detection task, but the large number of parameters in this method results in a less stable model that requires the use of many training techniques.

Table III shows a comparison of DepMSTAT with several classical models, including traditional machine learning and deep learning methods, as well as mainstream approaches. The performance of machine learning models is not as good as that of deep learning models, which could be attributed to the superior fitting capability of deep models to the data. Regarding classical deep learning methods, the approach proposed in this paper takes into account the temporal and spatial features of single-modal data and emphasizes the effective fusion of different modalities through complementary attention. Overall, DepMSTAT shows superior performance compared to other classical methods on the D-Vlog dataset with a precision of 71.53, a recall of 75.60 and an F1-score of 73.51. Furthermore, compared to existing methods, DepMSTAT not only simultaneously captures the spatiotemporal information of single-modal data (in contrast to [19], [70], [76]) but also effectively reduces the parameter count of the model (in contrast to [9]). The proposed SAB, TAB, and MTB modules contribute to the model's attention towards common features associated with depression in multimodal data, potentially enhancing the accuracy of depression diagnosis.

D. Ablation Study

In this section, we will present ablation experiments of the model that will demonstrate the effectiveness of our model from three different perspectives.

1) Different Connection Type: A simple and effective fusion of data from different modalities is required before inputting them into the depression classifier model. To validate the effectiveness of the feature fusion method used in the depression classifier, experiments are designed to compare the effects of two feature fusion methods, adding and concatenating different features, on the depression detection performance of the model.

TABLE IV
TWO CONNECTIONS TO FEATURES IN DEPRESSION CLASSIFIER

Connection_Type	Precision(%)	Recall(%)	F1-Score(%)
Add	67.66	73.17	70.31
Concat	**71.53**	**75.60**	**73.51**

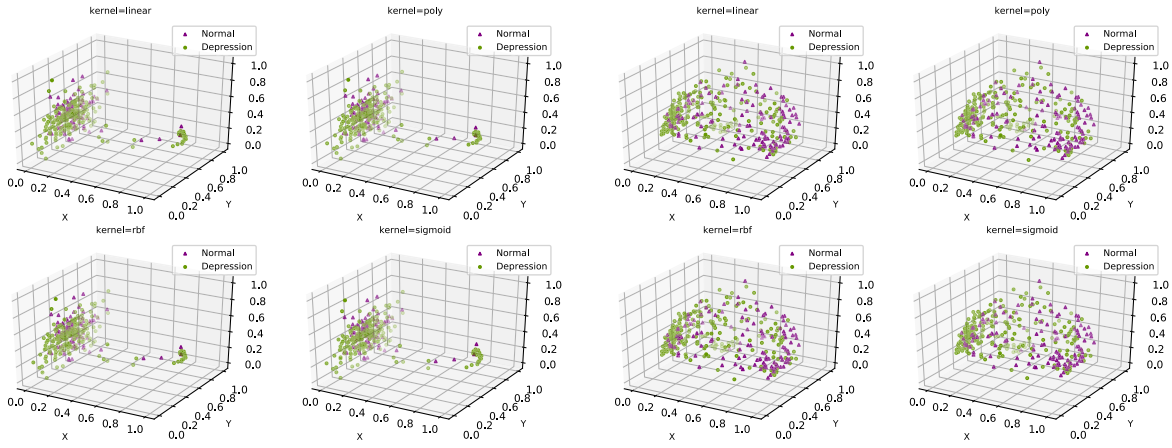
TABLE V
ABLATION STUDIES DIFFERENT SUB-MODULE PERFORMANCE RESULTS

TAB	SAB	MTB	Precision(%)	Recall(%)	F1-Score(%)
×	✓	✓	68.08	**78.04**	72.72
✓	×	✓	67.71	69.91	68.80
×	×	✓	69.17	74.79	71.87
$✓_1$	$✓_2$	✓	69.29	71.54	70.40
$✓_2$	$✓_1$	✓	**71.53**	75.60	**73.51**

"✓" indicates that the corresponding component is in used and "×" indicates that the component is not in use. The subscripts 1 and 2 indicate the order in which the modules appear.

As shown in Table IV, the feature fusion approach using direct adding is lower than the approach of concatenating different modal features for the evaluation metrics of recall, f1 score, and accuracy. This suggests that the direct addition approach may mask the variability of the different modal features compared to the direct concatenation approach, resulting in reduced recognition performance of the model. As illustrated by the features in Fig. 6, the acoustic features are more dispersed than the extracted visual features and are more likely to show differences between patients and normal subjects. Therefore, in this paper, features from different modalities are connected and fed into the depression classifier for depression classification.

2) Effectiveness of TAB and SAB: To verify the effectiveness of the two modules, TAB and SAB, in STAT, we designed three sets of ablation experiments: omitting the TAB module, omitting the SAB module, and omitting the TAB+SAB module. As can be seen from the first three rows of the ablation experiment in Table V, the recall and F1 scores for SAB+MTB are higher than those for TAB+MTB and MTB.The Precision (68.08%), Recall (78.04%) and F1 scores (72.72%) for TAB+MTB are



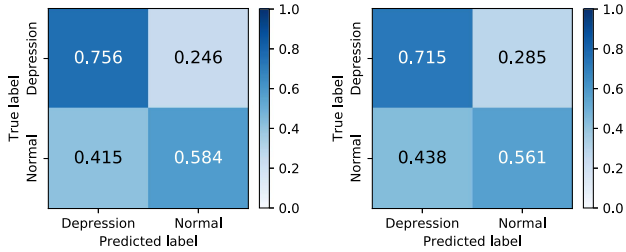
(a) Visualisation of visual features (b) Visualisation of acoustic features

Fig. 6. Each feature uses four different kernel PCA dimensionality reduction methods, and the kernel functions are linear kernel, poly kernel, rbf kernel and sigmoid kernel respectively. The triangle (purple) represents data from normal individuals, and the circle (green) represents data from depressed individuals.

lower than those evaluated for SAB+MTB, indicating that, compared to temporal features, spatial features better reflect the differences between normal and depressed individuals. The accuracy (71.53%), recall (75.60%) and F1 score (73.51%) of the DepMSTAT model were all higher than those of the three experimental groups, indicating that the introduction of the TAB and SAB modules could make the model performance more stable and the ability to extract valid spatiotemporal features with the ability to discriminate between depressed and normal controls.

3) Order Between SAB and TAB: To verify the effect of the stacking order of the TAB and SAB modules on the performance of the model, we designed ablation experiments with the opposite stacking order of these two modules as in DepMSTAT. As shown in the last two rows of Table V, all evaluation metrics for DepMSTAT are higher than those for this ablation experiment. It is noteworthy that the evaluation metrics of both sets of experimental results are basically more stable compared to the results of the previous ablation experiments. This to some extent indicates that the early fusion of the spatio-temporal characteristics of the different data modalities is beneficial to the stability of the model performance. In addition, the spatial information in the DepMSTAT model is more conducive to the extraction of discriminative features than the extraction of temporal features first. Fig. 7 shows the classification confusion matrix for the DepMSTAT model (Fig. 7(a)) and the model with the opposite stacking order (Fig. 7(b)) on the test set, showing that both models have more limited discriminative power for normal individuals than for depressives. Overall, the DepMSTAT model offers new ideas for real-world depression detection research.

Compared to the other baseline models, the DepMSTAT model achieved the best performance with a precision of 71.53%, a recall of 75.60%, and an f1 score of 73.51%. This is due to the ability of the DepMSTAT model to simultaneously perform spatial-temporal feature extraction and to fuse information from different unimodalities, thus enabling early feature fusion. In addition, the depression classifier used



(a) Confusion matrix for DepMSTAT (b) Confusion matrix for models with opposite stacking order to DepMSTAT

Fig. 7. Normalized confusion matrices of the fusion methods. Each row of the confusion matrices represents the true label and each column represents the predicted label. The element (i, j) is the percentage of samples in class i that is classified as class j.

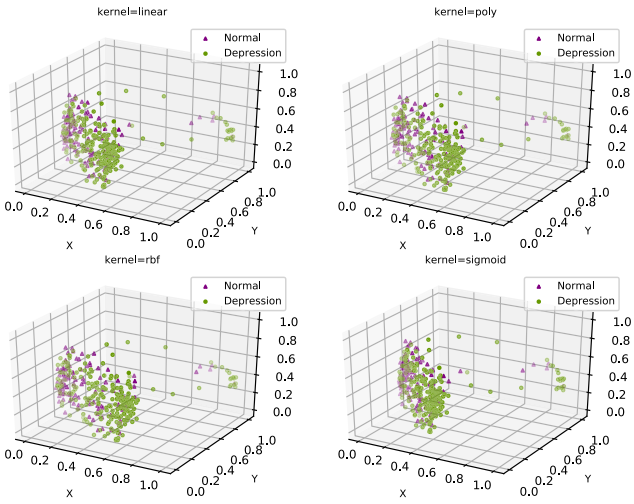


Fig. 8. Visualization of fusion feature distribution. Each fusion feature uses four different kernel PCA dimensionality reduction methods, and the kernel functions are linear kernel, poly kernel, rbf kernel and sigmoid kernel respectively. The triangle (purple) represents data from normal individuals, and the circle (green) represents data from depressed individuals. The feature visualisation after multimodal fusion is more compact in spatial distribution and has better discriminability.

employs a voting mechanism in order to avoid to some extent the problem of classification uncertainty due to data quality issues. The experimental results show that the proposed DepMSTAT method is effective and can learn the differences between normal subjects and depressed patients. In this work, we used visualised multimodal spatio-temporal fusion features to illustrate the representational learning capability of the model as shown in Fig. 8.

V. Conclusion and Future Work

To analyse the accuracy of multimodal data for depression detection in social scenarios, a depression detection framework, DepMSTAT, is proposed in which the STAT module is able to extract temporal and spatial features of the data and fuse them effectively. Experimental results show that effective fusion of spatio-temporal information from multimodal data and the use of a classifier with a voting mechanism can better classify depression. The method achieves better performance on the latest social media based dataset D-Vlog with lower number of parameters and faster computation speed. Based on the experimental results, we know that for the D-Vlog dataset, spatial features are more important than temporal features for detecting depression, but complementing each other can make the model performance more stable. Overall, the method proposed in this paper is objective and effective for the detection of depression. In addition, the use of multimodal fusion to improve recognition rates may provide some research ideas for future researchers. However, studying the use of primary datasets and the fusion of more modalities (facial expressions, speech, text, etc.) need to be further addressed by designing and optimising the network structure.

There are currently some issues that remain to be addressed in this study; the D-Vlog dataset only provides facial landmarks and processed speech data that has been processed to prevent disclosure of personal privacy, which to some extent lacks the original authenticity of the data. Therefore, the task of detecting depression is made more real by examining raw data from real-life scenarios. Simple data segmentation can result in redundant and invalid data, which not only increases computation, but can also interfere with the model's learning of valid features. In the future, it is necessary to design data segmentation with weights to effectively extract key sequence segments for depression detection.

References

[1] G. S. Malhi and J. J. Mann, "Depression," *Lancet*, vol. 392, no. 10161, pp. 2299–2312, 2018.

[2] K. Hawton, C. C. I. Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: A systematic review," *J. Affect. Disord.*, vol. 147, no. 1-3, pp. 17–28, 2013.

[3] H. Berenbaum, "Posed facial expressions of emotion in schizophrenia and depression," *Psychol. Med.*, vol. 22, no. 4, pp. 929–937, 1992.

[4] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[5] B. Hu, Y. Tao, and M. Yang, "Detecting depression based on facial cues elicited by emotional stimuli in video," *Comput. Biol. Med.*, vol. 165, 2023, Art. no. 107457.

[6] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Trans. Multimedia*, vol. 21, pp. 1476–1486, 2019.

[7] M. Yang, Y. Ma, Z. Liu, H. Cai, X. Hu, and B. Hu, "Undisturbed mental state assessment in the 5G era: A case study of depression detection based on facial expressions," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 46–53, Jun. 2021.

[8] L. He et al., "Deep learning for depression recognition with audiovisual cues: A review," *Inf. Fusion*, vol. 80, pp. 56–86, 2022.

[9] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, "Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer," *Digit. Commun. Netw.*, 2023, doi: 10.1016/j.dcan.2023.03.007.

[10] S. Xu et al., "Emotion recognition from gait analyses: Current research and future directions," *IEEE Trans. Computat. Social Syst.*, early access, Dec. 05, 2022, doi: 10.1109/TCSS.2022.3223251.

[11] A. Dev et al., "Exploration of EEG-based depression biomarkers identification techniques and their applications: A systematic review," *IEEE Access*, vol. 10, pp. 16756–16781, 2022.

[12] X. Zang, B. Li, L. Zhao, D. Yan, and L. Yang, "End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal," *J. Med. Biol. Eng.*, vol. 42, no. 2, pp. 225–233, 2022.

[13] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10140–10156, 2022.

[14] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[15] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3D facial expressions," 2018, *arXiv:1811.08592*.

[16] Y. Cao, Y. Hao, B. Li, and J. Xue, "Depression prediction based on BiAttention-GRU," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 11, pp. 5269–5277, 2022.

[17] H. Sun et al., "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors*, vol. 21, no. 14, 2021, Art. no. 4764.

[18] Y. Guo, C. Zhu, S. Hao, and R. Hong, "A topic-attentive transformer-based model for multimodal depression detection," 2022, *arXiv:2206.13256*.

[19] J. Yoon, C. Kang, S. Kim, and J. Han, "D-Vlog: Multimodal vlog dataset for depression detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, pp. 12226–12234, 2022.

[20] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[22] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[23] G. Synnaeve et al., "End-to-end ASR: From supervised to semi-supervised learning with modern architectures," 2019, *arXiv:1911.08460*.

[24] C. Lüscher et al., "RWTH ASR systems for LibriSpeech: Hybrid vs attention," in *Proc. Interspeech*2019, pp. 231–235.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28.

[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[28] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, and D. Yu, "Recurring the transformer for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14063–14073.

[29] C. Wang and Z. Wang, "Progressive multi-scale vision transformer for facial action unit detection," *Front. Neurorobot.*, vol. 15, 2021, Art. no. 824592.

[30] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10448–10457.

[31] S. Yao and X. Wan, "Multimodal transformer for multimodal machine translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4346–4350.

[32] W. Zhang, Y. Ying, P. Lu, and H. Zha, "Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 9571–9578.

[33] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 74–92.

[34] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, Oct. 2023.

[35] M. K. Hasan et al., "Humor knowledge enriched transformer for understanding multimodal humor," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 12972–12980.

[36] X. Zhan et al., "Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11782–11791.

[37] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2019, vol. 2019. Art. no. 6558.

[38] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3D dance generation with AIST ," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13401–13412.

[39] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 65–72.

[40] F. Ringeval et al., "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, 2019, pp. 3–12.

[41] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 588–601, Mar. 2020.

[42] W. Ragheb, J. Aze, S. Bringay, and M. Servajean, "Negatively correlated noisy learners for at-risk user detection on social networks: A study on depression, anorexia, self-harm, and suicide," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 01, pp. 770–783, Jan. 2023.

[43] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.

[44] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Jul.–Sep. 2020.

[45] W. C. D. Melo, E. Granger, and A. Hadid, "A deep multiscale spatiotemporal network for assessing depression from facial dynamics," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1581–1592, Jul.–Sep. 2022.

[46] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[47] M. A. Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 262–268, Jan.–Mar. 2021.

[48] L. He, C. Guo, P. Tiwari, R. Su, H. M. Pandey, and W. Dang, "Depnet: An automated industrial intelligent system using deep learning for video-based depression analysis," *Int. J. Intell. Syst.*, vol. 37, no. 7, pp. 3815–3835, 2022.

[49] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer Bi-LTSM," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 864–870, Apr.–Jun. 2022.

[50] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proc. Audio/Vis. Emotion Challenge Workshop*, 2018, pp. 23–30.

[51] Y. Dong and X. Yang, "A hierarchical depression detection model based on vocal and emotional cues," *Neurocomputing*, vol. 441, pp. 279–290, 2021.

[52] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 38–53.

[53] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 10, 2023, doi: 10.1109/TNNLS.2023.3247103.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[55] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[56] 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, Apr.–Jun. 2016.

[57] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[58] L. Bottou, "Stochastic gradient descent tricks," in *Proc. Neural Netw.: Tricks Trade*, 2012, pp. 421–436.

[59] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[61] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4604–4614.

[62] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Berlin, Germany: Springer, 2009.

[63] G. Huang, S. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2752–2761.

[64] A. Pampouchidou et al., "Facial geometry and speech analysis for depression detection," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 1433–1436.

[65] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[66] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, 2019, pp. 65–71.

[67] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[68] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[69] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-Vit: Adaptive tokens for efficient vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10818.

[70] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "TAMFN: Time-aware attention multimodal fusion network for depression detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 669–679, 2023.

[71] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12186–12195.

[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[73] K. Wu et al., "TINYVIT: Fast pretraining distillation for small vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 68–85.

[74] L. Zhao et al., "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[76] W. Zheng, L. Yan, and F.-Y. Wang, "Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2595–2613, Oct.–Dec., 2023.

[77] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[78] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.