2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Multi-modal Human Behaviour Graph Representation Learning for Automatic Depression Assessment

Haotian Shen, Siyang Song and Hatice Gunes
AFAR Lab, University of Cambridge

***Abstract*— Automatic depression assessment (ADA) often relies on crucial cues embedded in human verbal and non-verbal behaviors, which exists in video, audio, and text modalities. Although these modalities often show in time-series forms, current research offers limited exploration of the complex intra-modal temporal dynamics inherent to each modality, failing to extract the depression-related cues in a global view. While many methodologies attempt to exploit the multifaceted information encoded across modalities via decision-level or feature-level fusion techniques, they often fall short in effectively representing pairwise inter-modal relationships, which is the key to utilize the distinct complementary relationship between each modality pair. This paper presents a novel graph-based multimodal fusion approach, which can model intra-modal and inter-modal dynamics conveniently using a graph representation. It adopts undirected edges to link not only temporally continuous, pre-extracted features of each modality, but also temporally aligned features across each pair of modalities. This ensures the seamless propagation of global information across temporal dimensions and helps capture the pairwise inter-modal dynamics. We conduct experiments on the E-DAIC dataset to prove our approach's effectiveness, with an RMSE of 4.80 and a CCC value of 0.563, which rival the top-performing method. We also experiment on the AFAR-BSFP dataset to show the generality of our approach. Our code will be made publicly available.**

I. INTRODUCTION

Depression is a highly prevalent mental health disorder that exerts a detrimental influence o n a n i ndividual's feelings and behaviors [36]. Traditional diagnostic methods, primarily reliant on professional interviews, are both time-consuming, subjective, and strain limited mental health resources [27], [45]. Recognizing these challenges, recent research has shifted focus on applying deep learning to automatic depression assessment (ADA). The majority of these ADA studies focuses on analyzing video [57], [78], [19], [28], [32], [25], [58], [1], audio [75], [48], [50], and textual data [17], [41], [12] expressed by the target subject, as these modalities not only can be easily recorded but also contain rich depression-related cues.

Since existing ADA approaches frequently attempt to make predictions based on time-series signals (e.g., video, audio and text), a key challenge is how to properly utilize intra-modal temporal dynamics to extract depression-related cues from each modality. However, most of these approaches fail to consider or deliberately overlook the temporal relationship within each modality. Some of them [31], [33], [64], [32] eliminate the temporal properties within the raw input by extracting hand-crafted features combined with statistical methods (average, sum, frequency, etc.). Others [59], [78], [18] segment the input into many small chunks and make predictions for each chunk, then average them to obtain the prediction. Although there exists several approaches that leverage Temporal Convolutional Networks (TCN) [22], [72] and Recurrent Neural Networks (RNN) [48], [23], [2] to encode temporal dynamics within each modality, they are limited by the one-way induction and long dependency issues, respectively.

To obtain enhanced depression assessment predictions, researchers also investigated how to make predictions from multiple modalities. Consequently, it is important to explore the relationship between different modalities (i.e., modelling inter-modal dynamics) in order to optimally combine them for ADA. To achieve this, feature-level fusion methods that concatenate features from different modalities into a single high-dimensional vector [50], [51] have been frequently employed. However, these approaches mistakenly assume modalities are conditionally independent [46], missing out on capturing important pair-wise inter-modal relationships. Alternatively, decision-level fusion strategies mainly combine the predictions from separate uni-modal models, which overlook the dynamic relationships between modalities [5].

To address the issues discussed above, this paper proposes a novel graph-based multi-modal fusion strategy for the ADA task, which aims to effectively model both intra-modal and inter-modal dynamics via a graph-based strategy. Our approach constructs a multi-modal graph with nodes representing chunk-level depression-relevant features from various modalities. This graph structure includes undirected edges between temporally adjacent nodes within the same modality, allowing each node to consider information from both of its past and future states during the reasoning. We also establish inter-modal edges between temporally aligned nodes from different modalities, enabling the graph to explicitly capture complex inter-modal relationships. These inter-modal edges also act as shortcuts for information sharing, overcoming the limitations of RNN-style models in efficiently transmitting information across temporally nonadjacent nodes. In summary, the main novelties and contributions of this paper are summarised as follows:

- Our study pioneers the use of graphs to explore and represent the intra-modal and inter-modal dynamics of time-series data in ADA, by modelling the temporal dynamics within the same modality or different modalities using graph representations.
- With the proposed graph structure, we address the efficiency and long-dependency limitations typically

associated with RNN-style models.
- Our methodology explicitly models the pair-wise relationship between different modalities within the graph framework, which enables our approach to achieve performance comparable with the SOTA for the E-DAIC dataset.

II. Related Works

A. Automatic depression assessment approaches

In the domain of ADA, the use of video, audio, and text modalities has been extensively researched. To capture the visual cues, Song et al. [59] employs a histogram-based approach to quantify the average occurrences of human facial primitives, using a MLP for training. In another research [57], they apply Fourier transform to facial features, achieving a fixed-length spectral representation conducive for their training process. He et al. [26] develops a comprehensive framework that combines local-attention-based and global-attention-based Convolutional Neural Networks to capture facial features at different scales. Pampouchidou et al. [44] focuses on dynamic facial expressions, utilizing algorithms including local binary patterns on motion history images. Xie et al. [70] designs an end-to-end framework tailored for variable-length videos, integrating a 3D CNN for exploring local temporal patterns and a redundancy-aware self-attention (RAS) scheme for aggregating global features. Melo et al. [19] adopts a two-stream CNN network, separately processing appearance and temporal features, followed by a score fusion method to integrate the predictions from both streams. Wang et al. [69] selects key frames from videos, combines adjacent frames within a certain window, and processes them through separate LSTM networks, with a global max pooling layer aggregating the outputs. Finally, Niu et al. [43] segments videos into fixed-length clips, subsequently analyzing them using a spatio-temporal attention (STA) network.

For audio modality, Ye et al. [72] extract deep features using DeepSpectrum, subsequently integrating these features into a customized Temporal Convolutional Network (TCN), with the final layer employing relational attention classification for output activation. Zhang et al. [76] developed a self-supervised convolutional encoder-decoder network dedicated to extracting features from spectrogram images of audio clips. These features are then processed through a 4-layer, 128-dimensional LSTM network. In a differing approach, Vazquez et al. [66] utilize spectrograms directly as input for their 1D-Convolutional Neural Network (CNN), which draws inspiration from DepAudioNet [40]. Likewise, Lin et al. [38] also engage in spectrogram extraction from audio data, subsequently channeling these into a 1D-CNN for analysis. Finally, Toto et al. [65] segment raw audio into multiple overlapping sub-clips for feature extraction, train these features using Support Vector Machines (SVM), and then employ mean pooling to aggregate the outputs for final prediction.

Regarding the text modality, Chiong et al. [17] extract hand-crafted features using bag-of-words (BoW) and n-gram techniques from a Twitter dataset [55], subsequently utilizing these features in different machine learning classifiers for depression detection. Ray et al. [48] employ the pre-trained Universal Sentence Encoder [14] to derive sentence-level embeddings, which were then padded and processed through a 2-layer Bi-LSTM network. Lin et al. [38] exploit the capabilities of the pre-trained language model, ELMo, to encode textual data, followed by training using a Bi-LSTM network enhanced with an attention layer. In a similar vein, Shen et al. [56] also utilize ELMo for feature extraction, with the training process leveraging a Bi-LSTM network. Amanat et al. [3] adopt a one-hot encoding technique to quantify the frequency of key depressive words in a pre-cleaned dataset, feeding these features into an LSTM-RNN model for depression assessment. Additionally, Ye et al. [72] employed the Continuous Bag of Words (CBOW) method for text feature extraction, followed by training using a customized transformer model.

For enhanced performance and the utilization of complementary information across different modalities, a variety of multimodal fusion methods have been employed in ADA. Rohanian et al. [51] employ a feature-level fusion strategy that concatenates feature vectors from video, audio and text modalities at early stage before feeding the integrated features into a word-level LSTM. In contrast, three other studies [22] [56][38] implement feature-level fusion at a later stage, which is the time after each modality has been processed by the corresponding encoders. Those encoders are typically CNN and LSTM-based networks, and the last layers of them are concatenated horizontally before fed into a dense network for prediction. Ringeval et al. [49] utilize a straightforward decision-level fusion technique, averaging regression outputs from video-based and text-based models to obtain the final depression severity prediction. Conversely, Ye et al.[72] involves the learning process into their decision-level fusion approach, combining predictions from individual modality models into a fully-connected network. Niu et al. [42] implement a novel method involving the concatenation of features extracted from audio and text modalities of each question and answer pair. These concatenated features are treated as vertices in a graph, with edges established between adjacent nodes within a specified context window. This graph is then trained using a customized Graph Attention Network (GAT). While many studies in multimodal fusion for ADA focus predominantly on employing advanced training networks, there remains a lack of research centered on representation learning. Specifically, there is a gap in exploring methodologies for more effectively combining features from different modalities beyond simple concatenation. Addressing this gap could lead to a better performance for the ADA systems.

B. Graph-related techniques in multimodal fusion

Recent research has increasingly focused on the utilization of graph-related techniques for multimodal fusion. One study [74] focuses on the multimodal neural machine translation, which involves translating sentences from a source to a target language within the context provided by an image.

To establish cross-modal relationships, this study introduces edges between the embeddings of nouns in the sentences and the corresponding object embeddings in the images. In the domain of emotion recognition, Hu et al. [29] combine the Memory Fusion Network (MFN) with a Graph Convolutional Network (GCN) to achieve fusion of multimodal data. Li et al. [37] adopt a graph-centric approach, constructing individual graphs for each modality pair among three distinct modalities. Each graph is trained using a Graph Attention Network (GAT), with the outputs subsequently concatenated and processed through a dense network for final analysis. Targeting at action recognition, Duhme et al. [21] employ a unique fusion strategy for data collected from various sensors. They utilize a GCN to integrate sensor data both in the channel dimension and spatial dimension, demonstrating the versatility of graph-related techniques in handling complex multimodal datasets.

III. Methodology

This section introduces a two-module framework for multimodal depression behavior graph representation learning. As shown in Figure 1, this framework comprises a Depression-Related Feature Extraction Module and a Graph-Based Multimodal Fusion Module, which are explained below separately. In short, the framework initiates by extracting time-series features from relevant tools, which are then segmented into multiple chunks. These chunks are processed through respective behavior encoders to extract deep features. Subsequently, these extracted features are represented as nodes in a graph, with edges formed between them based on predefined rules. This constructed graph is then inputted into a GNN aiming to predict the depression severity.

A. Depression-Related Feature Extraction Module

This module is designed to learn depression-related behaviour features from raw input data for subsequent fusion processes, which is illustrated in the first part of Figure 1. Initially, through corresponding preprocessing, toolkits and pre-trained models, raw data from each modality is transformed into multi-channel time series represented as $F_* \in \mathbb{R}^{t \times d}$, where $*$ denotes the type of modality, t for time dimension and d for the feature dimension. The resulting time-series are fed into appropriate behavior encoders to extract deep features. The procedure is detailed for each modality, respectively.

1) Video modality: To summarize complex facial behaviours (facial image sequence) as a set of compact and semantically meaningful representations, we directly employ the widely-used OpenFace 2.0 toolkit [8] to obtain 27 FAU intensities, 6 head poses and 8 gaze directions from each frame. We exclude all the frames where the toolkit fails to capture human face or possess a confidence value below 0.85, then apply min-max normalization for facial representation time-series. Then, we split the time-series into several standardized chunks based on the temporal sequence, where each chunk is treated as an independent and discrete sample. The behavior encoder for extracting depression-related deep features for the video modality draws inspiration from [71], which comprises two main components: the Multi-Scale Behavior Feature Extraction Module (MB) and the Noise Separation (NS) module. The MB module discerns depression-relevant behavior-primitives across varied scales, from small to large are 4, 8, 16, and 31 by adjusting the filter size of the 1D convolutional layers. The NS module eliminates non-depression-related noise from the extracted features from the MB module.

2) Audio Modality: For the audio modality, we use the Hugging Face pre-trained speaker diarization model to separate the sound of interviewee from the raw audio clip for further processing [9]. Each prcoessed audio clip is filtered with 4th order band pass filtering to filter out sound between 85-3400 Hz (typical human sound frequency range) [6], [13], [63]. Then the resulting audio clip is fed into the DeepSpectrum tookit [4] to get the 4096-d time-series form deep features, which is obtained by applying the pre-trained VGG16 model to the spectrogram of the audio – an image created by generating 128 mel-frequency bands for the window size of 4 seconds and the hop size of 1 second. Similary to the video modality, the time series is segmented into chunks for further feature extraction. The feature extraction backbone for the audio modality is a CNN-RNN architecture, as it is powerful in managing time-series data. These networks commence with multiple CNN blocks, each containing a 1D convolutional layer succeeded by a batch normalization layer, a Rectified Linear Unit (ReLU) activation, and a max-pooling layer. Following the CNN blocks is a 32-d LSTM layer and a fully connected layer with 32 neurons, where the deep extracted features are obtained.

3) Text Modality: we use the Huggin Face DistilBERT [53] to extract the sentence-level features from the transcript of each interview, resulting in a 768-d representation. The behavior encoder of the text modality empirically shares the similar CNN-RNN architecture with the RNN part is a 32-d GRU layer.

Motivation of using chunk-level feature: In our study, the time series video and audio signals are divided into multiple chunks based on the fact that indications of depression are discernible in short-term interval regardless of the conversational content [35], [39]. The similar feature processing strategies have also been adopted in [20], [54], which is particularly advantageous in processing variable lengths of the video and audio time series. Segmenting these into fixed-length chunks both facilitates data augmentation and simplifies the training process. On the contrary, for the text modality in our study, we employ a different strategy. The text-based behavior encoder combines all sentence embeddings of the target speech transcript, which generates only one chunk-level representation. This approach stems from the premise that textual data might reveal depressive indicators primarily within specific conversational contexts. For instance, mundane discussions, such as those about an individual's residence, are less likely to provide insights into depressive states when analyzed in isolation. In contrast,

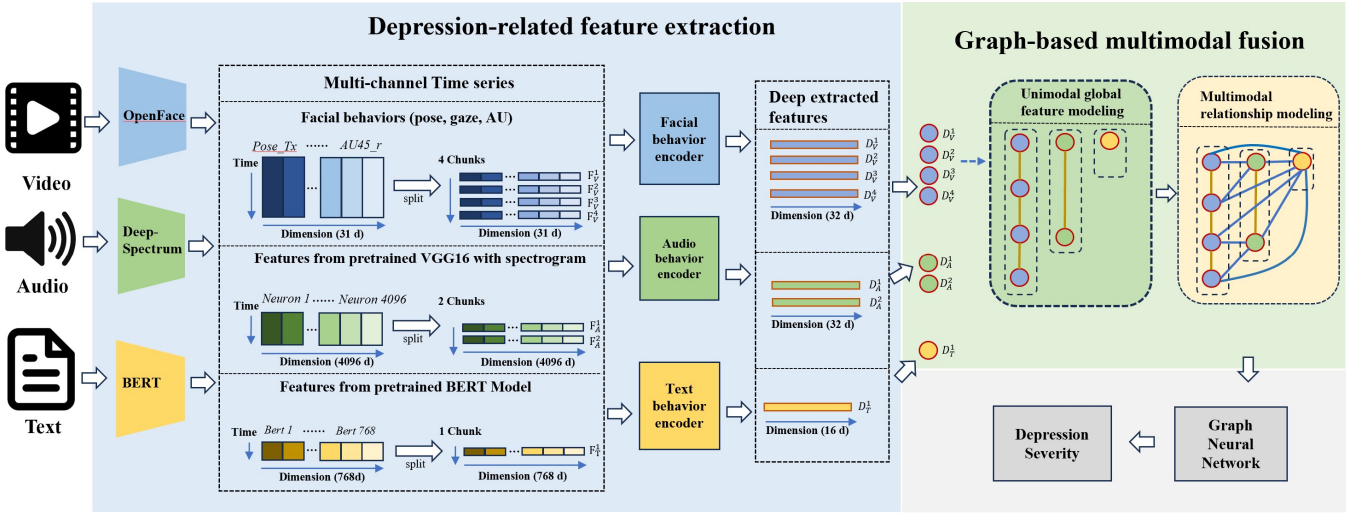

Fig. 1: Graph-Based Multimodal Fusion Pipeline. Note: The chunk numbers (4, 2, and 1) for each modality depicted in the figure are for illustrative purposes only and do not represent actual case scenarios.

considering the entire speech transcript enables the inclusion of conversational topics and global contexts, which can serve as significant indicators of depression [11], [60], [47].

B. Graph-Based Multimodal Fusion Module

As illustrated in the second part of Figure 1, this module constructs a graph representation to fuse the extracted audio, visual and text features. This is achieved by establishing the intra-modal and inter-modal edges to model the intra-modal and inter-modal dynamics. Before defining these edges, we first encode each extracted chunk-level depression-related feature as a node of the graph. For simplicity, we denote i_{th} chunk feature in the time dimension of video, audio and text modality as D_V^i, D_A^i, D_T^i, where V, A, T represent video, audio, and text modality, respectively. It should be noted that the i-th nodes from different modalities do not necessarily correspond to the same time interval, since the number of chunks varies for different modalities.

1) Intra-modal dynamics modelling: To model the intra-modal dynamics, we consider the temporal order of each chunk in the original modality. To encapsulate the inherent temporal inter-relationships among continuous features, undirected edges are established between temporally adjacent node pairs of the same modality. The set of intra-modal edges E_{intra} can be formulated as:

$$E_{\text{intra}} = \left\{ \{D_*^i, D_*^{i+1}\} \mid 1 \leq i < N_* \right\} \tag{1}$$

where $*$ represents modality V,A or T and N_* represents the number of nodes defined for the modality $*$. These intra-modal edges serve a dual purpose: (i) they enable seamless access to information from both preceding and succeeding nodes compared to the TCN-based methods which only allows one-way information passing; and (ii) despite that the introduction of shortcuts between nodes facilitated by the inter-modal edges (to be discussed subsequently), these intra-modal edges preserve the temporal induction inherent in the relationships between nodes. As a result, a comprehensive and detailed understanding of temporal dynamics within each modality can be achieved.

2) Inter-modal dynamics modelling: To model inter-modal dynamics among features extracted from text, audio and video, we also propose to add inter-modality edges allowing nodes from temporally aligned modality pairs (e.g., a pair of audio-visual chunk-level node features) to be interconnected. Specifically, nodes from different modalities are interconnected if they represent equivalent temporal intervals. The inter-modal edges E_{inter} of a pair of modalities $m1$ and $m2$ can be formulated as:

$$E_{\text{Inter}} = \bigcup_{j=1}^{N_{m2}} \left\{ D_{m1}^i, D_{m2}^j \mid i \in \left[1 + (j-1) \left\lfloor \frac{N_{m1}}{N_{m2}} \right\rfloor, j \left\lfloor \frac{N_{m1}}{N_{m2}} \right\rfloor \right] \right\} \tag{2}$$

where $m1$, $m2$ represent two distinct modalities while N_{m1}, N_{m2} represent the number of nodes for the modality $m1$ and $m2$, respectively. This formula suggests dividing the nodes with a larger count into segments equivalent to the ratio of N_{m1} to N_{m2}, and then linking each segment to the corresponding node from the modality with fewer nodes.

The design stems from the assumption that features in close temporal proximity across different modalities can offer compensatory or complementary insights [7] for various tasks. Contrary to methodologies introduced in [15], [51], [42], the proposed multi-modal graph representations does not require manual temporal-alignment in the pre-processing phase, which simplifies the data preparatory process. This is because the inter-modal edges ensure each node can accept and share information to all other nodes within the graph, which allows information contained in temporally associated nodes of different modalities propagate to each other wherever they are positioned in the graph.

Moreover, the inter-modal edges function as efficient conduits for information exchange between distantly located

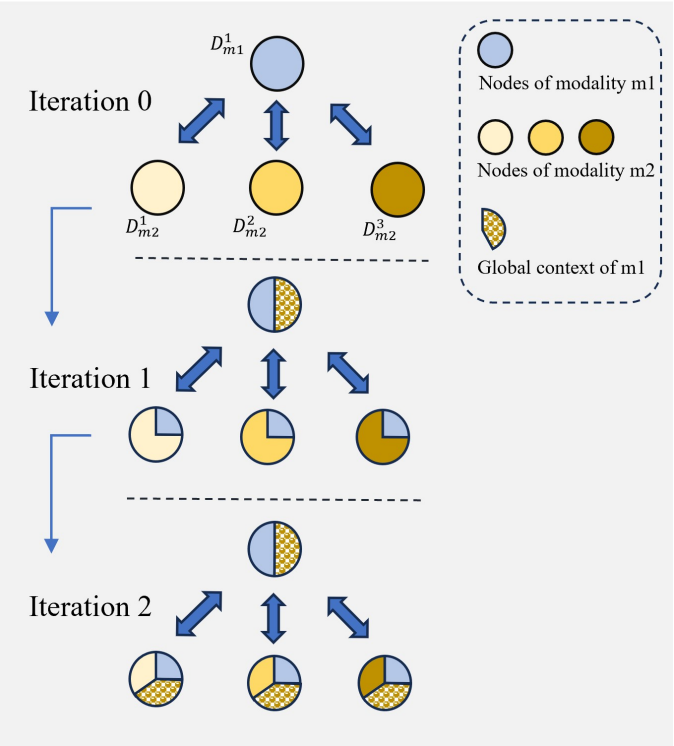

Fig. 2: Information sharing with shortcut of inter-modal edges

nodes of the same modality, thereby obviating the need to traverse through all intermediate nodes chronologically. This mechanism is important in aggregating the global context of information from a specific modality. To illustrate, consider the hypothetical scenario depicted in Figure 2, comprising three nodes from modality $m1$ and one node from modality $m2$. In the initial iteration, all nodes from modality $m1$ relay their information to the node of $m2$, which in turn encodes the overarching context of $m1$. In the subsequent iteration, this global context is propagated to all nodes of $m1$, thereby granting access to the information of other nodes, regardless of their temporal distance. This efficient information sharing mechanism requires merely two iterations. In contrast, a scenario relying solely on intra-modal edges, without the facilitation provided by inter-modal edges, would necessitate $N-1$ steps for information from the initial node to reach the final node in a sequence of N nodes of the same modality.

Importantly, unlike decision-level fusion approaches that make predictions for each modality separately or naive feature-level methods that only concatenate features across all modalities into a high-dimension vector, our graph representation learning-based fusion strategy not only utilizes the relationship between features of different modalities to encode rich task-related information but also can explicitly and specifically model the relationship between each pair of modalities.

After modelling both intra-modal and inter-modal dynamics, the edge set is defined: $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} = E_{\text{intra}} \cup E_{\text{inter}}$. Upon defining the node set $\mathcal{V}$ and edge set $\mathcal{E}$, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed. To better capture the complementary relationship between features from different modalities, the nodes are classified into different types based on the modality, and the edges are classified into different types based on the type of nodes they link, which result in a heterogeneous graph.

C. Prediction with Heterogeneous Graph Transformer

Upon the construction of the graph, we employ the Heterogeneous Graph Transformer (HGT) [30] for training. HGT is uniquely configured to recognize and interpret diverse node and edge types, thereby facilitating a deeper understanding of the interactions among various relationships and entities within the graph [30] Node representations are refined by sequentially stacking HGT layers, with the representation of node t at the $(l-1)$-th layer, $H^{l-1}[t]$, being updated to the l-th layer, $H^{l}[t]$, through the application of the general update formula [30]:

$$H^{l}[t] \leftarrow \underset{\forall e \in E(s,t), \forall s \in N(t)}{\text{Aggregate}} \left(\text{Extract}\left(H^{l-1}[s]; H^{l-1}[t], e\right)\right) \tag{3}$$

Here, $N(t)$ denotes all neighboring nodes of t, and $E(s,t)$ represents all edges directed from node s to t. The function Extract(.) is responsible for extracting information through the pair of nodes and the edge within, while Aggregate(.) compiles the neighborhood information for the target node. Specifically, the model computes heterogeneous mutual attention, inspired by transformers, using meta-relations—represented as $\langle \tau(s), \phi(e), \tau(t) \rangle$ —which encode the type of source node s edge e, and target node t. This attention value is then utilized to update the message extracted from the source node. Subsequently, the neighboring information is aggregated through a weighted average, with each attention vector serving as a weight, to yield the updated representation $H^{l}[t]$ [30].

In our HGT design, we employ a 64-dimensional linear projection layer to unify the feature dimensions across all modalities. Subsequently, we stack eight PyTorch Geometric (PyG) [24] HGT layers, each with a filter size of 64 and 2 attention heads. A linear regressor is then applied to each node, reducing its feature dimension to 1. To obtain a graph-level prediction, we employ a customized weighted mean pooling layer that computes the average of all the post-regression node values, with larger weights for modalities that have lower node count. We adopt the Mean Square Error (MSE) as the loss function, which is defined as follows:

$$\begin{aligned} F_{loss} &= \frac{1}{n}\sum_{i=1}^{n}(y^i - \hat{y}^i)^2 \\ \text{where } \hat{y}_i &= \frac{1}{3}\left(\sum_{i=1}^{N_V}\frac{\hat{y_V^i}}{N_V} + \sum_{i=1}^{N_A}\frac{\hat{y_A^i}}{N_A} + \hat{y_T^i}\right) \end{aligned} \tag{4}$$

IV. Experiments

A. Experimental settings

Modality	E-DAIC		
	Reference	RMSE	CCC
A+V	Baseline [49]	6.37	0.111
A+V	Sun et al. [61]	6.22	0.331
A+T	Kaya et al. [34]	-	0.344
A+V+T	Rodrigues et al. [50]	6.11	0.403
A+V+T	Suggu et al. [52]	5.36	0.457
A+V+T	Zhao et al. [77]	**4.11**	-
A+T	Fan et al. [22]	5.91	0.430
A+V+T	Yin et al. [73]	5.50	0.442
A+V+T	Fang et al. [23]	5.17	-
A+V+T	Sun et al. [62]	-	**0.583**
A+V+T	**Ours**	<u>4.80</u>	<u>0.563</u>

TABLE I: Performance comparison of mutimodal fusion depression severity assessment on the E-DAIC. A: Audio; V: Video; T: Text. Best result is highlighted in bold; second best result is highlighted with underline.

a) Datasets: Our study employs two datasets: the Extended Distress Analysis Interview Corpus (E-DAIC) for analyzing mental health issues such as depression, and the AFAR-Brief Solution-Focused Practice (BSFP) dataset for classifying mental well-being states. The E-DAIC comprises 275 semiclinical interviews (totaling 73.2 hours), where each involves an English-speaking participant answering a series of open-ended questions. PHQ score that ranges from 0 to 27 is used as the ground truth for assessing the depression severity. The AFAR-BSFP dataset is a proprietary dataset owned by the AFAR Lab at the University of Cambridge [16]. It contains data from 41 sessions each recorded as 20-minute mental wellbeing coaching sessions between a human coach and 11 coachees, practising Brief Solution Focussed Practice (BSFP) over a 4-week period. The metal-wellbeing state of the participant of each session was labelled as either positive or negative based on self-report questionnaires. There are 17 negative samples and 24 positive samples.

b) Metrics: We follow previous studies [49], [71] to employ Root Mean Square Error (RMSE) and Concordance Correlation Coefficient (CCC) as metrics for the E-DAIC dataset. For the BSFP dataset, we use accuracy and F1 score in accordance with the baseline established by the dataset owner.

c) Implementation Details: In our approach, feature segmentation is important for the temporal analysis. For the video modality, we configure each chunk with 30 timestamps, equivalent to approximately one second of raw video at a 30Hz sampling rate. In the audio modality, chunks also consist of 30 timestamps, each timestamp representing one second of audio, aligning with our spectrogram generation technique that uses a one-second hop size. For training our graph-based model, we optimize performance through specific parameters: a batch size of 10, a learning rate of 10^{-3}, and the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, applied consistently across both datasets. The Mean Square Error (MSE) loss function is employed for the regression task in E-DAIC, while Cross Entropy is used for the classification task in BSFP dataset. To ensure balanced training in the BSFP dataset, we utilize a weighted random sampling approach, addressing potential data imbalances. Our evaluation for the E-DAIC dataset is on its test set, following the train-test split predefined by the dataset owner. For the AFAR-BSFP dataset, we implement a 5-fold cross validation approach to evaluate the performance of our methodology.

Modality	AFAR-BSFP-DB		
	Reference	Acc	F1
A+V	Baseline [16]	0.760	0.800
A+V	**Ours**	**0.780**	**0.830**

TABLE II: Performance comparison of mutimodal fusion general mental-wellbeing state classification on the AFAR-BSFP-DB. No other result except the baseline from [16] is available.

B. Comparison with existing methods

Table I shows that our graph-based multimodal fusion methodology yields competitive results. Specifically, it achieved the second-best result in terms of both RMSE and CCC metric on the E-DAIC dataset, trailing by less than 3.5% from the current SOTA. This underscores the efficacy of our graph representation approach in integrating multiple modalities, positioning it as one of the leading methodologies for multimodal depression recognition. Notably, the work of Zhao et al. [77], which records the best RMSE, also utilizes a transformer-based approach that investigates the pairwise relationships between modalities, thereby validating our method's emphasis on establishing pairwise inter-modal edges. Furthermore, all competitive methods, including Sun et al.'s study [62]—the best approach in terms of CCC on modeling the complexities inherent in multimodal interactions, moving beyond simplistic concatenation-based feature-level fusion and decision-level fusion strategies. Moreover, as presented in the table II, the robust performance of our approach on the BSFP dataset further shows its general applicability, indicating it can be safely used with different dataset and tasks.

C. Ablation studies

We conduct a series of rigorous ablation studies to examine the reliability and validity of our multimodal fusion methodologies. Notably, these investigations are exclusively conducted on the E-DAIC dataset, while preserving universal applicability.

1) Number of modalities: Table III shows the performance of our multimodal methods via different combination of modalities. It can be observed that the more modalities used in the fusion, the better performance obtained. This is because the inclusion of more modalities in the fusion process implies the incorporation of a broader spectrum of

potentially complementary information, which in turn provides a richer set of depression-related cues. This also proves that out graph-based fusion method can exploit the subtle inter-relationship between modalities for better prediction.

Modality	E-DAIC	
	RMSE	CCC
A	5.65	0.413
V	5.83	0.324
T	5.68	0.418
A+V	5.23	0.552
A+T	5.01	0.526
V+T	5.16	0.516
A+V+T	**4.80**	**0.563**

TABLE III: Performance comparison of our mutimodal fusion depression severity assessment on the E-DAIC via different set of modalities. Best result is highlighted in bold.

2) Comparison with other fusion techniques: We also examine the performance of other three fusion techniques, including two decision-level fusion and one feature-level fusion strategies. The compared fusion approaches are detailed as follows:

- **Decision-level fusion (Average)**: This fusion approach computes an average of the individual predictions from each behavior encoder to generate the final prediction. No training incurred within this method.
- **Decision-level fusion (MLP)**: This fusion technique involves concatenating the predictions from all behavior encoders to form an input feature, which is then fed into a 2-layer MLP, consisting of 8 and 4 neurons respectively, to generate the final prediction.
- **Feature-level fusion (concat)**: This fusion method concatenates the features extracted from various behavior encoders and inputs the resulting feature into a 2-layer MLP, which with the hidden dimension of 32 and 16 respectively, for prediction.

Fusion level	Name	RMSE	CCC
Decision	Average	5.07	0.445
	MLP	5.26	0.425
Feature	Concat	6.97	0.351
	Ours	**4.80**	**0.563**

TABLE IV: Performance comparison of our depression severity assessment on the E-DAIC between different fusion methods. Best result is highlighted in bold.

The comparative performance of the various fusion methods is presented in Table IV. As expected, the proposed graph representations yield the best performance. This can be attributed to the generation of the graph representation, which finely captures the interplay of both intra-modal and inter-modal relationships. Visualization of predictions on Figure 3 shows our proposed method has the best performance, as more predictions approximate to the ground truth.

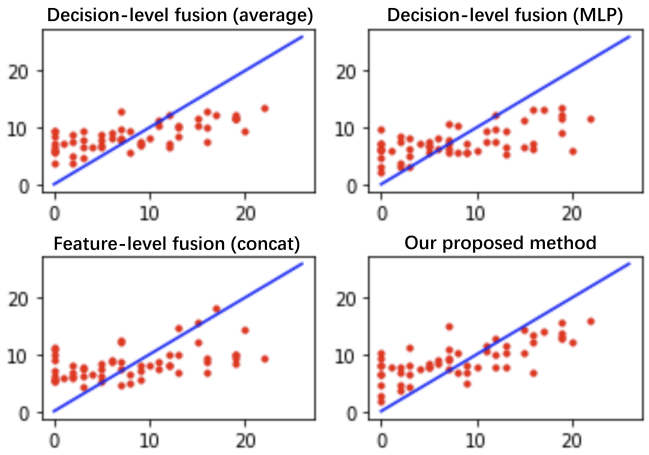

Fig. 3: Predictions of different fusion methods on the E-DAIC dataset. The x-axis denotes ground truth, y-axis denotes predictions.

3) Comparison with other GNNs: We conducted a comparative analysis of various Graph Neural Networks (GNNs) applied to our graph representation, utilizing the PyTorch Geometric (PyG) implementations. The GNN architectures examined include the Graph Attention Network (GAT) [67], Graph Attention Network version 2 (GATv2) [10], Heterogeneous Attention Network (HAN) [68], and Heterogeneous Graph Transformer (HGT) [30]. GAT is characterized by its utilization of linear attention mechanisms between a source node and its neighbors. In contrast, GATv2 extends this concept by incorporating a dynamic attention mechanism. HAN represents a heterogeneous adaptation of GAT, designed to process graphs with diverse node and edge types. Meanwhile, HGT draws inspiration from the Transformer model, catering specifically to the nuances of heterogeneous graphs. The first two models, GAT and GATv2, are tailored for homogeneous graphs, where there is no differentiation between node and edge types, whereas HAN and HGT are engineered to handle the complexity of heterogeneous graphs. The results of this comparative study are shown in the table V.

GNN	RMSE	CCC
GAT	7.23	0.428
GATv2	5.04	0.549
HAN	<u>4.99</u>	<u>0.557</u>
HGT	**4.80**	**0.563**

TABLE V: Performance comparison between GNNs. Best result is highlighted in bold. Second best result is underlined.

In Table V, it is evident that the GNNs designed for homogeneous graphs exhibit comparatively lower performance. This outcome is due to the inherent limitations of these models in discerning and categorizing the meta-relations between nodes. Conversely, GNNs specifically developed for

heterogeneous graphs are pre-equipped with this relational information, facilitating more effective learning. Among the heterogeneous graph-focused GNNs, the HGT exhibits superior performance across both metrics. This result can be ascribed to its Transformer-based architecture, which is adept at capturing the intricate nuances present among various modalities. Furthermore, the HAN also shows satisfactory performance. This outcome underscores the versatility and general applicability of our graph representation framework, indicating its robustness across different GNN architectures.

4) Comparison between different chunk size: A significant challenge inherent in our methodology is related to the determination of an appropriate chunk size for the time series data derived from various modalities. Given that the text modality necessitates a single chunk configuration to capture the global conversational context essential for accurate prediction, as previously discussed, our focus is on the chunk size settings for the video and audio modalities. We conduct the grid search to find the optimal chunk-size. The best performance is selected based on the prediction of the extracted features from corresponding behavior encoders.

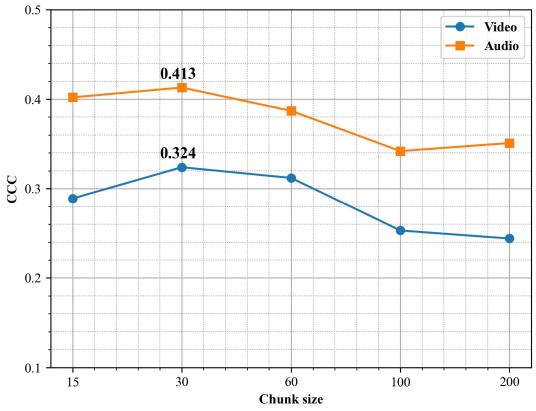

Fig. 4: Predictions of corresponding modality encoder with different chunk size. The best results are presented.

As illustrated in Figure 4, empirical results indicate that both the video and audio modalities exhibit optimal performance with a chunk size of 30. However, despite the same chunk size, the temporal coverage of information differs significantly between these two modalities. Specifically, a single chunk in the video modality encompasses one second of raw video data, whereas in the audio modality, it covers 30 seconds of audio data. The difference in its temporal coverage can be attributed to the nature of each modality. Visual cues related to depression, such as a frown, can manifest instantaneously and therefore can be effectively captured within short temporal spans. Conversely, vocal indicators of depression, which include a reduced speech rate and a monotonous tone, require a longer time frame for accurate detection. Additionally, the presence of pauses between sentences in the raw audio further necessitates longer chunks. Short temporal chunks in audio might fall on the pause part, failing to capture meaningful patterns. Therefore, the determination of chunk size for each modality reflects a careful consideration of the inherent characteristics of each modality in presenting the depression-related cues.

5) Alternative graph structure: In our research, we rigorously evaluated two alternative configurations for the graph structure utilized in multimodal fusion. The first alternative involved the introduction of additional edges between nodes of the same modality, where these nodes were separated by one intermediate node, effectively creating a 'skip' connection. The second structural variation was the incorporation of self-loops for each node, a design intended to facilitate capturing of an explicit self-attention within the graph's framework. However, the performance metrics reveal that neither of these alternative structures outperform the original graph design. This observation could be attributed to the additional, and potentially erroneous, assumptions introduced by these modifications. Specifically, the creation of 'skip' edges between non-adjacent nodes presupposes a strong correlation between these nodes across the intervening gap. However, this assumption may not universally apply across all timestamps, potentially leading to inaccuracies in the model's interpretation of the temporal dynamics. Furthermore, the introduction of additional self-attention through self-loops, while theoretically promising for emphasizing individual node characteristics, may not effectively contribute to the overall fusion process in the context of multimodal data, where inter-node and inter-modality relationships are more crucial.

V. Conclusion and Limitations

In this research, we have introduced a graph-based multimodal fusion approach for automatic depression assessment, addressing challenges of modeling both intra-modal and inter-modal behavioural dynamics. Our approach achieved competitive results in the E-DAIC dataset, providing an RMSE of 4.80 and a CCC value of 0.563, and outperform the baseline of the AFAR-BSFP dataset. Despite these promising results, our study has certain limitations, notably the utilization of a fixed graph structure, which may not be universally optimal across diverse datasets. Additionally, our investigation was confined to the integration of three behavioral modalities: video, audio, and text. Future research directions could focus on dynamic graph construction methods and extend the exploration of our fusion approach to other modalities, including Electroencephalography (EEG), thereby broadening the scope and applicability of our findings in the realm of mental health assessment.

Ethical Statement

In this research, we strictly adhere to the highest ethical standards, ensuring all data were collected from individuals who have provided informed consent for their use in research. The datasets are exclusively utilized for the intended research purposes and are not shared beyond the research team. Moreover, all data processing is conducted anonymously, guaranteeing that individual participants could not be identified, thus upholding the privacy and confidentiality of each participant.

ACKNOWLEDGEMENTS

Haotian Shen undertook this research work as part of his MPhil in ACS degree at the Department of Computer Science and Technology, University of Cambridge. **Funding:** Siyang Song and Hatice Gunes have been supported by the EPSRC project ARoEQ under grant ref. EP/R030782/1. **Open Access:** For the purpose of *open access*, the authors have applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising. **Data Access:** This study involves secondary analyses of existing datasets, that are described and cited in the text. Licensing restrictions prevent sharing of the datasets.

REFERENCES

[1] N. I. Abbasi, S. Song, and H. Gunes. Statistical, spectral and graph representations for video-based facial expression recognition in children. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1725–1729. IEEE, 2022.

[2] T. Al Hanai, M. M. Ghassemi, and J. R. Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[3] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin. Deep learning for depression detection from textual data. *Electronics*, 11(5):676, 2022.

[4] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller. Snore sound classification using image-based deep spectrum features. In *Interspeech 2017*, pages 3512–3516. ISCA, Aug. 2017.

[5] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.

[6] R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[8] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[9] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[10] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[11] W. Bucci and N. Freedman. The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334, 1981.

[12] S. G. Burdisso, M. L. Errecalde, and M. Montes y Gómez. Using text classification to estimate the depression level of reddit users. *Journal of Computer Science & Technology*, 21, 2021.

[13] J. C. Catford et al. *A practical introduction to phonetics*. Clarendon Press Oxford, 1988.

[14] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[15] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 163–171, New York, NY, USA, 2017. Association for Computing Machinery.

[16] J. Cheong, M. Spitale, and H. Gunes. " it's not fair!"–fairness for a small dataset of multi-modal dyadic mental well-being coaching. In *IEEE International Conference on Affective Computing and Intelligent Interaction (IEEE ACII'23)*, pages 1–8, 2023.

[17] 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. *Computers in Biology and Medicine*, 135:104499, 2021.

[18] W. C. de Melo, E. Granger, and A. Hadid. Combining global and local convolutional 3d networks for detecting depression from facial expressions. In *2019 14th ieee international conference on automatic face & gesture recognition (fg 2019)*, pages 1–8. IEEE, 2019.

[19] W. C. De Melo, E. Granger, and M. B. Lopez. Encoding temporal information for automatic depression recognition from facial analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1080–1084. IEEE, 2020.

[20] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 22(2):525–536, 2017.

[21] M. Duhme, R. Memmesheimer, and D. Paulus. Fusion-gcn: Multimodal action recognition using graph convolutional networks. In *DAGM German Conference on Pattern Recognition*, pages 265–281. Springer, 2021.

[22] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu. Multi-modality depression detection via multi-scale temporal dilated cnns. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 73–80, New York, NY, USA, 2019. Association for Computing Machinery.

[23] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561, 2023.

[24] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[25] M. Gavrilescu and N. Vizireanu. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors*, 19(17):3693, 2019.

[26] L. He, J. C.-W. Chan, and Z. Wang. Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, 422:165–175, 2021.

[27] L. He, D. Jiang, and H. Sahli. Multimodal depression recognition with dynamic visual and audio cues. In *2015 International conference on affective computing and intelligent interaction (ACII)*, pages 260–266. IEEE, 2015.

[28] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022.

[29] J. Hu, Y. Liu, J. Zhao, and Q. Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*, 2021.

[30] Z. Hu, Y. Dong, K. Wang, and Y. Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[31] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12, 2018.

[32] S. Jaiswal, S. Song, and M. Valstar. Automatic prediction of depression and anxiety from behaviour and personality attributes. In *2019 8th international conference on affective computing and intelligent interaction (acii)*, pages 1–7. IEEE, 2019.

[33] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Computational and mathematical methods in medicine*, 2018, 2018.

[34] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 27–35, 2019.

[35] D. Keltner and A. M. Kring. Emotion, social function, and psychopathology. *Review of General Psychology*, 2(3):320–342, 1998.

[36] Y. Lee, R.-M. Ragguett, R. B. Mansur, J. J. Boutilier, J. D. Rosenblat, A. Trevizol, E. Brietzke, K. Lin, Z. Pan, M. Subramaniapillai, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of affective disorders*, 241:519–532, 2018.

[37] J. Li, X. Wang, G. Lv, and Z. Zeng. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, page 126427, 2023.

[38] L. Lin, X. Chen, Y. Shen, and L. Zhang. Towards automatic

depression detection: A bilstm/1d cnn-based model. *Applied Sciences*, 10(23):8701, 2020.

[39] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

[40] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42, 2016.

[41] K. Milintsevich, K. Sirts, and G. Dias. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14, 2023.

[42] M. Niu, K. Chen, Q. Chen, and L. Yang. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4235–4239. IEEE, 2021.

[43] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE transactions on affective computing*, 2020.

[44] A. Pampouchidou, M. Pediaditis, E. Kazantzaki, S. Sfakianakis, I.-A. Apostolaki, K. Argyraki, D. Manousos, F. Meriaudeau, K. Marias, F. Yang, et al. Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation. *Machine Vision and Applications*, 31(4):30, 2020.

[45] V. Patel, S. Saxena, C. Lund, G. Thornicroft, F. Baingana, P. Bolton, D. Chisholm, P. Y. Collins, J. L. Cooper, J. Eaton, et al. The lancet commission on global mental health and sustainable development. *The lancet*, 392(10157):1553–1598, 2018.

[46] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[47] N. Ramirez-Esparza, C. Chung, E. Kacewic, and J. Pennebaker. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In *Proceedings of the international AAAI conference on web and social media*, volume 2, pages 102–108, 2008.

[48] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88, 2019.

[49] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12, 2019.

[50] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63, 2019.

[51] M. Rohanian, J. Hough, M. Purver, et al. Detecting depression with word-level multimodal fusion. In *Interspeech*, pages 1443–1447, 2019.

[52] G. S. Saggu, K. Gupta, K. Arya, and C. R. Rodriguez. Depressnet: A multimodal hierarchical attention mechanism approach for depression detection. *Int. J. Eng. Sci.*, 15(1):24–32, 2022.

[53] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[54] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund. Audio based depression detection using convolutional autoencoder. *Expert Systems with Applications*, 189:116076, 2022.

[55] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[56] Y. Shen, H. Yang, and L. Lin. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE, 2022.

[57] S. Song, S. Jaiswal, L. Shen, and M. Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020.

[58] S. Song, Y. Luo, T. Tumer, M. Valstar, and H. Gunes. Loss relaxation strategy for noisy facial video-based automatic depression recognition. *ACM Transactions on Computing for Healthcare*, 2024.

[59] S. Song, L. Shen, and M. Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 158–165. IEEE, 2018.

[60] S. W. Stirman and J. W. Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522, 2001.

[61] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen. Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors*, 21(14):4764, 2021.

[62] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3722–3729, 2022.

[63] J. Sundberg. Articulatory interpretation of the "singing formant". *The Journal of the Acoustical Society of America*, 55(4):838–844, 1974.

[64] M. Tlachac and E. Rundensteiner. Screening for depression with retrospectively harvested private versus public text. *IEEE journal of biomedical and health informatics*, 24(11):3326–3332, 2020.

[65] E. Toto, M. Tlachac, F. L. Stevens, and E. A. Rundensteiner. Audio-based depression screening using sliding window sub-clip pooling. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 791–796. IEEE, 2020.

[66] A. Vázquez-Romero and A. Gallardo-Antolín. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688, 2020.

[67] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[68] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019.

[69] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, and S. Li. Automatic depression detection via facial expressions using multiple instance learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1933–1936. IEEE, 2020.

[70] W. Xie, L. Liang, Y. Lu, C. Wang, J. Shen, H. Luo, and X. Liu. Interpreting depression from question-wise long-term video recording of sds evaluation. *IEEE Journal of Biomedical and Health Informatics*, 26(2):865–875, 2021.

[71] J. Xu, S. Song, K. Kusumam, H. Gunes, and M. Valstar. Two-stage temporal modelling framework for video-based depression recognition using graph representation. *arXiv preprint arXiv:2111.15266*, 2021.

[72] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu. Multi-modal depression detection based on emotional audio and evaluation text. *Journal of Affective Disorders*, 295:904–913, 2021.

[73] S. Yin, C. Liang, H. Ding, and S. Wang. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 65–71, New York, NY, USA, 2019. Association for Computing Machinery.

[74] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo. A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint arXiv:2007.08742*, 2020.

[75] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi. Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 47–53, 2019.

[76] P. Zhang, M. Wu, H. Dinkel, and K. Yu. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 135–143, 2021.

[77] Z. Zhao and K. Wang. Unaligned multimodal sequences for depression assessment from speech. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 3409–3413, 2022.

[78] X. Zhou, K. Jin, Y. Shang, and G. Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 11(3):542–552, 2018.