IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 January 2024, accepted 30 January 2024, date of publication 5 February 2024, date of current version 12 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3362233

RESEARCH ARTICLE

Additive Cross-Modal Attention Network (ACMA) for Depression Detection Based on Audio and Textual Features

NGUMIMI KAREN IYORTSUUN[1], SOO-HYUNG KIM[1], (Member, IEEE), HYUNG-JEONG YANG[1], (Member, IEEE), SEUNG-WON KIM[1], AND MIN JHON[2]

[1]Department of AI Convergence, Chonnam National University, Gwangju 61186, Republic of Korea
[2]Department of Psychiatry, Chonnam National University Hospital, Hwasun 58128, Republic of Korea

Corresponding author: Soo-Hyung Kim (shkim@jnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] under Grant RS-2023-00219107; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development Grant funded by the Korea Government (MSIT) under Grant IITP-2023-RS-2023-00256629; and in part by the Chonnam National University Hwasun Hospital, Institute for Biomedical Science, under Grant HCRI 23026.

ABSTRACT Detecting depression involves using standardized questionnaires like the Patient Health Questionnaires (PHQ-8/9). Yet, patients might not always provide genuine responses, leading to potential misdiagnoses. Therefore, the need for a means to detect depression in patients without the use of preset questions is of high importance. Addressing this challenge, our study aims to discern telltale symptoms from statements made by the patient. We harness both audio and text data, proposing an Additive cross-modal attention network to learn and pick up the appropriate weights that best capture the cross-modal interactions and relationships between both features using BiLSTM as the backbone of both modalities. We tested our approach on the DAIC-WOZ dataset for depression detection and also evaluated our model performance on the EATD-Corpus. Benchmarked against similar studies on these datasets, our method demonstrates commendable efficacy in both classification and regression models for both unimodal and multimodal approaches. Our findings underscore the potential of our model to effectively detect depression in patients while using textual and speech modalities without the necessary use of preset questions for effective detection.

INDEX TERMS Machine learning, deep learning, depression, healthcare, mental health diagnosis.

I. INTRODUCTION

Various mental disorders such as depression affect a person's emotional, social, and psychological well-being. Depression is one of the most widely known mental disorders and reports have stated that it might as well be the most prevalent mental disorder by 2030 [1]. Depression ranges from mild, temporary episodes of sadness to severe, persistent states often referred to as Clinical or Major depression (MD) [2]. Usually identified through symptoms of deep sadness, and loss of interest in day-to-day social activities, depression, if not properly diagnosed, may result in constant feelings of suicidal ideation and suicide attempts [3]. Therefore, the early detection of depression is of high importance.

For depression detection using machine learning techniques, two main steps are usually employed. Initially, implicit and explicit data are collected in various formats such as visual, acoustic, and textual formats, during the process of answering specific questions. Signal data can also be collected from wearable devices for this purpose. In some cases, like that of the Distress Analysis Interview Corpus (DAIC) [4] one of the most popular open-access depression datasets, text transcripts were extracted from the recorded audio data to enhance diagnostic accuracy, and visual features were also extracted. In the second step, various machine learning/deep learning algorithms are then employed to analyze depression

The associate editor coordinating the review of this manuscript and approving it for publication was Ze Ji.

severity based on the collected data. Although with the high advancement of research in this area, and the development of new machine learning techniques, significant challenges remain in the practical implementation and attainment of diagnostic accuracy.

It is indeed evident that over the last 10 years, massive efforts have been made by researchers to find applicable ways for automatic depression detection which does not involve the typical steps of response collection and manually accessing participants' mental states. However, significant issues pose a challenge to the progress of depression diagnosis research. Firstly, small sample sizes abound in this field mostly as a result of the high expense of data collection which often requires human participants [5] and privacy concerns. While some machine learning (ML) models can maintain their performance accuracy even when trained on a small dataset, the same cannot be said for deep learning (DL) models [3]. Secondly, many patients who are depressed feel the need to hide that part of their lives mostly due to societal stigma. Such patients may provide false responses about their symptoms, which can lead to incorrect diagnoses thereby impacting their treatment outcomes. Thirdly, there is a high demand for further investigation into the effectiveness of combining various features from different data sources which proves the importance of the development of automatic depression detection techniques which could be applied in clinical settings.

In this work, we investigate the effectiveness of attention mechanisms in the role of depression detection by employing the audio and text data of the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) and the Emotional Audio-Textual Depression Corpus (EATD-Corpus) datasets. We aim to effectively detect depression without the use of preset questions. Firstly, we aim to classify depression into states of depressed and not-depressed, and secondly, we aim to estimate the intensity of depression using the provided PHQ-8 scores from the DAIC-WOZ dataset as a regression problem and test our model generalization capacity on the EATD-Corpus dataset.

Our contribution to this research is as follows:

- We designed a model that can be repurposed for 2 tasks; classification to model PHQ-8 Binary outcomes, and regression to model PHQ-8 Score outcomes.
- We propose an Additive cross-modal attention network for depression detection. This method uses Bidirectional Long Short-Term Memory (BiLSTM) with attention layers on unimodal audio and text models to capture representations before multimodal fusion.
- We conduct experiments to show the effectiveness of the attention weights utilized at the BiLSTM levels of both the audio and text models on the entire model architecture.
- Finally, we evaluated our models' generalization capacity on the EATD-Corpus dataset.

The rest of the paper is structured as follows. Section II outlines the background knowledge which breaks down the sections of our research model into research conducted specifically on audio, text, multimodality, and research on attention mechanisms. Section III contains the sub-net architecture of our research methodology and the analysis of our proposed additive cross-modal attention network. Section IV shows details on the datasets used for our experiment and our system implementation. Section V reports the results of our experiments on classification, regression, effects of attention weights on the overall model, and evaluation on the EATD-Corpus dataset. Section VI is the conclusion of the paper.

II. RELATED WORKS

In this section, we present a breakdown of the most common techniques used for the detection of depression using audio and text modalities as well as a fusion of both modalities for the effective diagnosis of depression.

While depression is commonly associated with human experiences, features for this research can be derived from multiple sensory modes by which humans interpret their environment, including auditory, visual, olfactory, and more [6]. In the early stages of depression detection research, essential features were handpicked based on questions considered crucial in understanding depression [7], [8].

A. AUDIO MODALITY

The use of audio/acoustic modality for depression detection involves changing audio data into specific forms. The most widely known forms of speech data are the Mel Frequency Cepstrum Coefficient (MFCC) [9], Log Mel spectrograms, Mel spectrograms, and COVAREP features [10].

Ma et al. [11], proposed a model called DepAudioNet, to mine depression representations from vocal cues, adopting the concepts of Long Short-Term Memory (LSTM) and 1D-CNN to encode a discriminative audio representation for depression recognition. In [12], 1D-CNN was also used to model the spatial feature representations from raw waveforms, and LSTM was used to learn the short-term and long-term feature representations from the Mel-scale filter banks. In addition, to balance the positive and negative samples, a random sampling approach is adopted in the model training stage before using LSTM.

In [13], the authors proposed an end-to-end depression detection method using a Convolutional Neural Network (CNN) auto-encoder for automatic feature extraction out of raw audio signals. In addition, to address the well-known sample imbalance problem of the DAIC-WOZ dataset, they used a cluster-based sampling technique to reduce the risk of bias toward the majority class (not depressed).

B. TEXT MODALITY

Depression detection in machine learning using text modality generally extracts feature vectors during preprocessing and performs an analysis. Over the years, Recurrent Neural Networks (RNN) and CNN have been widely used for this task.

However, the RNN model is used specifically on time-series data and is often a preferred approach to this over CNN. With RNN, various errors such as the vanishing or exploding gradients during backpropagation prove to be a challenge [14]. To solve this problem, models such as Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM) have been developed and are widely used in research involving text modalities. LSTM incorporates memory cells and gating mechanisms, while BiLSTM further enhances the models' capabilities by incorporating bidirectionality, capturing both past and future information. While myriad sources of data can be utilized for depression detection, [15] proposed a study to detect signs of depression from social media posts even when the posted words did not contain phrases like ''depression'' or ''diagnosis.'' They employed three depression-diagnosis-only datasets obtained from Facebook, Reddit, and an electronic diary to assess the performance of the trained models against other social media sources while testing the text properties of two public datasets during the training and testing phases. In another study by [16], the authors were able to attain reasonable results by proposing a deep learning-based hybrid model for an early depression diagnosis. Here, the authors applied BiLSTM with Glove, Fastext embedding techniques, Word2Vec, Linguistic Inquiry and Word Count (LIWC), and Meta-Data features for depression detection.

C. MULTIMODALITY

Reference [17] proposed a hybrid method for estimating and classifying depression that included the fusion of the text, audio, and video data of the DAIC-WOZ dataset. They combined the depression degree estimations obtained from each modality using a multivariate regression model. They developed a depression/non-depression classification system leveraging the text modality and Paragraph Vector (PV), Support Vector Machine (SVM), and Random Forest. The system as a whole produced a 0.667 F-measure for the depressed class in the development set. In another research by [18], a depression detection model was proposed based on sequences of text transcripts and audio of the DAIC-WOZ dataset as well. The performance of their multi-modal Long Short-Term Memory (LSTM) model was evaluated and yielded good performance. Reference [8] won the AVEC-2016 challenge with their multimodal method where they proposed a gender-specific Decision tree classifier based on visual and speech features. Comparing the high performance of the incorporation of various modalities for depression detection.

D. ATTENTION MECHANISM

In the realm of deep learning, attention has emerged as a pivotal concept, garnering significant importance. It draws inspiration from the human cognitive system, which tends to concentrate on salient components when confronted with vast quantities of information. As deep neural networks have progressed, the attention mechanism has found widespread adoption across various domains, owing to its ability to selectively focus on relevant features or elements, enhancing the network's performance and interpretability. Various studies have implemented the Attention mechanism in their research. Reference [19] proposed a model that combines residual thinking and attention mechanism. They designed a depression corpus based on the self-reference effect (SRE) experimental paradigm and labeled the speech dataset; then the attention module was introduced into the residual. Their results proved that the utilization of attention networks for depression detection could show better results compared to traditional machine learning methods.

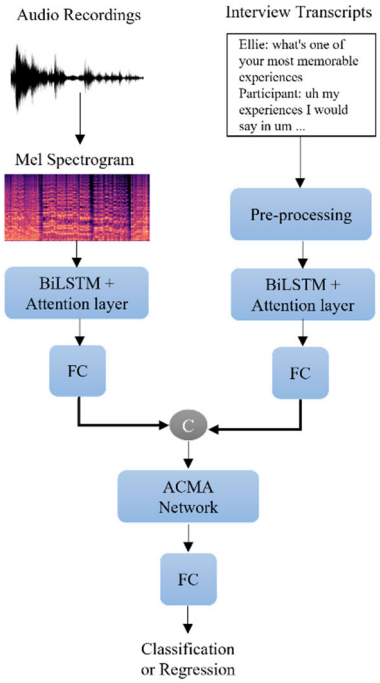


FIGURE 1. Overview of the proposed model architecture. The model performs classification when set up for the classification task and regression when set up for the regression task.

III. PROPOSED METHOD

In this section, we introduce the proposed method for this research – a multi-modal architecture featuring a fusion of extracted representations from the two different data modalities. Our proposed model architecture can be seen in Fig. 1.

A. BILSTM WITH ATTENTION

LSTM, known for its proficiency in managing sequences of varying lengths, has limitations that restrict its ability to effectively leverage future contextual information and extract local contextual details. Additionally, due to the absence of a mechanism to discern varying relevance levels within the data, LSTMs struggle to perceive the differing importance of individual data segments accurately.

Since we hope to capture the sequential dependencies within each modality, we opted for a BiLSTM network with

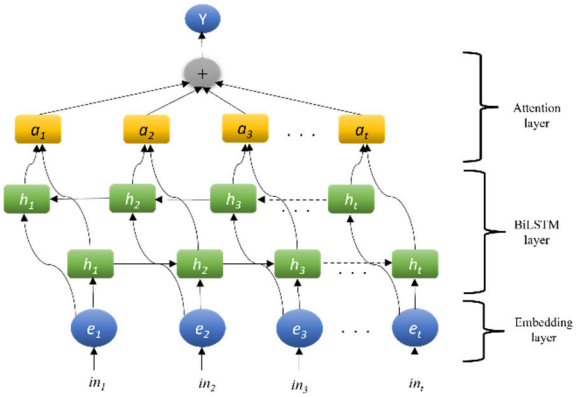


FIGURE 2. Illustration of the BiLSTM with attention applied. in1,...int is the data input sequences passed to model, e1,...et represents the data embeddings, h1,...ht at the BiLSTM layer represent the forward and backward pass of BiLSTM, and a1, ...at are the attention weight. Y represents the output of the BiLSTM with attention.

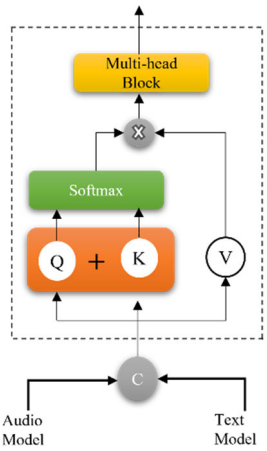


FIGURE 3. Additive cross-modal attention network.

an attention mechanism applied to both text and audio features independently.

While the BiLSTM network assigns weights for feature extraction by considering the forward and backward contextual information of the data and learning the relevance of each feature, the attention mechanism assigns different weights to words to improve comprehension of the sentiment in both modalities.

Overall, the attention mechanism combined with BiLSTM enables the model to selectively attend to and assign weights to the most relevant elements in both the audio and text domains. The architecture of the BiLSTM with attention applied to both modalities is shown in Fig. 2.

B. ADDITIVE ATTENTION

Although the concept of attention has been in existence for a long time and has been used in various areas of research [20], [21], [22], the more recent and widely known concept of attention in the context of deep learning and natural language processing emerged in the early 2010s with the introduction of the "Attention Is All You Need" paper by [23]. This work significantly influenced the development of attention mechanisms and their applications in various domains, including machine translation, language understanding, and image generation. Since then, attention mechanisms have become a fundamental component of many state-of-the-art neural network architectures, providing improved performance in various tasks that require modeling dependencies and interactions between different elements of a sequence or a set of inputs.

In [24], Bahdanau et al. introduced a neural machine translation model with an attention mechanism. The attention mechanism calculates context vectors in the decoder by incorporating the hidden state of the RNN. The context vector for each word is computed as a weighted sum of annotations, where the attention weights are obtained by normalizing energy scores using a SoftMax function. Based on the alignment between the previous hidden state and the annotation, the energy scores are computed. and combining the forward and backward hidden states, the annotations are created. This attentional mechanism aids in the context vectors' ability to transport more selective context data.

C. ADDITIVE CROSS-MODAL ATTENTION NETWORK

Cross-modal attention in machine learning is a mechanism that enables the integration and interaction of information from different modalities. It allows the model to attend to and align relevant features across various modalities such as audio, text, image, or video, to enhance comprehension or performance.

For our experiment, we utilized the cross-modal attention mechanism to enhance our understanding of speech and text representations. Although various attention models exist such as the dot product attention [25], we opt for the additive attention proposed by [24] as described above as the core building block of our model. This mechanism incorporates non-linear transformations to compute attention weights and provides flexibility in modeling complex relationships as well as handling more fine-grained interactions between modalities.

Each audio and text modality after being preprocessed, as shown in Fig. 1, is independently fed into the BiLSTM with an attention layer to capture the sequential dependencies and relevant features within each modality. After contextual representations have been extracted from both modalities and passed through separate fully connected (FC) layers, the Additive Cross-Modal Attention (ACMA) network is then employed as shown in Fig. 3.

This network facilitates the fusion of information from both modalities by considering the dependencies between them. It leverages the attention weights obtained from the individual modalities to compute cross-modal attention weights.

The FC layers of the audio and text modalities, denoted as C, are concatenated as follows:

$$C = [A_f, T_f] \tag{1}$$

TABLE 1. An overview of the PHQ-8 system.

		Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly everyday
Sub-classes	1.	Little interest or pleasure in doing things	0	1	2	3
	2.	Feeling down, depressed, or hopeless	0	1	2	3
	3.	Trouble falling or staying asleep, or sleeping too much	0	1	2	3
	4.	Feeling tired or having little energy	0	1	2	3
	5.	Poor appetite or overeating	0	1	2	3
	6.	Feeling bad about yourself – or that you are a failure	0	1	2	3
	7.	Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
	8.	Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

The sub-classes numbered 1-8 represent the 8 major depression symptoms. PHQ-8 Score is the sum of all PHQ-8 Sub scores, $0-24$, and PHQ-8 Binary is 1 if PHQ-score is ≥ 10 else PHQ-8 Binary is 0.

where A_f and T_f represent feature vectors for audio and text modalities, and $[A_f, T_f]$ denotes the concatenation operation merging the extracted features from both modalities into a single representation.

The concatenated representations are then passed through the ACMA network, which combines the concepts of additive attention and cross-modal interaction.

The concatenated representation C is transformed into a set of three vectors consisting of query (Q), key (K), and value (V) vectors, using linear transformations implemented as FC layers with appropriate weights. We incorporate these linear layers to project the concatenated representation into lower-dimensional spaces, where the resulting vectors can capture specific aspects and relationships between the audio and text features. These transformations are represented as follows:

$$Q_i = W_Q \cdot C + b_Q \tag{2}$$

$$K_i = W_K \cdot C + b_K \tag{3}$$

$$V_i = W_V \cdot C + b_V \tag{4}$$

where W represents weight matrices, and b corresponds to the bias vectors associated with the query, key, and value projections, respectively. These transformations apply matrix multiplications and bias terms to the concatenated representation and allow the network to learn appropriate weights that best capture the cross-modal interactions and relationships between the audio and text features.

Next, we calculate additive attention by applying the SoftMax function denoted as:

$$SoftMax(x) = \frac{exp(x)}{exp\Sigma(x)} \tag{5}$$

to the element-wise sum of the transformed query and key vectors. The resulting attention weights are then multiplied elementwise with the transformed value vectors. The additive attention A_i for the *i-th* attention head is calculated as follows:

$$A_iW = Softmax(Qi + Ki) \tag{6}$$

$$A_iOut = A_iW \odot Vi \tag{7}$$

where A_iW is the attention weight, and A_iOut is the attention output. The resulting attention weights are multiplied elementwise with the value vectors (V_i).

Lastly, a concatenated output from all attention heads is acquired and directed to an FC layer:

$$A = Concat[A_1, A_2, \ldots, A_M] \tag{8}$$

With Msymbolizing the number of attention heads.

IV. EXPERIMENTS

A. DATASETS

1) DAIC-WOZ DATASET

The DAIC-WOZ [4] dataset is utilized for this research. This dataset is one of four interviews all under the Distress Analysis Interview Corpus (DAIC) which comprises Face-to-face, Teleconference, Wizard-of-Oz, and Automated interviews. The DAIC-WOZ interview was conducted by an animated virtual interviewer (Ellie), controlled by a human interviewer in another room. The dataset itself contains clinical interviews designed to support the diagnosis of mental disorders such as depression, PTSD, and anxiety. The dataset contains 189 participants/sessions, and various forms of data have been collected including audio, verbal, and non-verbal cues.

Measures of psychological distress were also included and the Patient Health Questionnaire depression scale (PHQ-8) [26], was used as the standardized validation method for assessing and diagnosing the severity measure of depression. To calculate the PHQ-8 Score, participants were asked about the frequency of specific depressive symptoms experienced over the previous 14 days. The individual sub-scores from PHQ-8 are then combined to yield a total score ranging from 0 to 24 points. This total is used to determine the presence of Major Depression (MD). Specifically, a PHQ-8 Score of 10 or higher indicates a positive diagnosis of MD, while a score below 10 suggests the absence of the condition. Table 1 shows an overview of the PHQ-8 system for further understanding.

Although this dataset indeed contains various data formats, various issues accompany it, e.g., imbalance as only 56 participants are annotated as ''depressed'' and 133 participants are annotated as ''not depressed'', labeling errors, and it is rather a small-scale dataset.

2) EATD-CORPUS DATASET

Emotional Audio-Textual Depression Corpus (EATD-Corpus) is a publicly available Chinese depression dataset. This dataset was collected from 162 volunteer students from Tongji University, and it consists of audio and text transcripts where participants were made to answer three randomly selected questions and also complete a Self-rating Depression Scale questionnaire (SDS). With EATD-Corpus, a standard SDS point equivalent to or over 53 implies a state of depression while an SDS score below 53 implies no depression. Therefore, for the EATD-Corpus dataset, its statistics imply a high level of data imbalance as 30 participants were depressed while 132 participants were not depressed.

3) IMBALANCE HANDLING ON THE AUDIO AND TEXT DATA

As proven in the survey conducted by [3] regarding the data imbalance issue with the DAIC-WOZ dataset, this issue is rather a global problem with most mental health datasets proving to be difficult to access due to ethical and privacy concerns. With the issue of data imbalance comes the challenge of prediction bias from classes with the highest number of samples.

Reference [27] proposed a data resampling technique where the participant recordings are cropped into a number of slices. For audio feature analysis, the audio data of depressed samples are resampled without duplication until their quantity is equivalent to that of not-depressed samples. A similar approach is taken for the text features. However, in this case, every 10 responses from each participant are grouped and text samples are randomly selected from different groups of responses from the depressed samples. This process is repeated until the number of text samples balances out between the depressed and not-depressed samples. For this experiment, we adapted this resampling method to balance out the training samples between the depressed and not-depressed classes of the DAIC-WOZ dataset.

With EATD-corpus, the dataset has no specific split labels of depressed or not-depressed therefore, we utilized the SDS labeling scale where an SDS score ≥ 53 implies depression and an SDS score < 53 implies non-depression to manually label the provided dataset. More about this dataset is discussed in the results section below.

B. SYSTEM IMPLEMENTATION

This experiment was conducted using Pytorch version 1.12.1+cu102 with Python programming language version 3.10.6. Features of the text and audio data of this experiment are processed separately following the structure of the architecture shown in Fig. 1.

TABLE 2. BiLSTM with attention model parameter setting.

Layers	Parameter
BiLSTM layer	Hidden LSTM units: 128 Layers: 2 Dropout: 50%
Attention layer	
Dropout	50%
Dense layer (FC 1)	Output features: 128 Activation function: ReLU
Dropout	50%
Dense layer (FC 2)	Output features: 128 Activation function: ReLU

1) TEXT MODEL ANALYSIS

The DAIC-WOZ dataset consists of text transcripts that contain sequences of responses between the interviewer and the participants. First, we identified and extracted topic-related information based on pre-defined topics from the provided 'queries' file. Then, the transcript for the given participant ID was processed line by line and participant responses were extracted and grouped based on the identified topics. While iterating through the participant responses, irrelevant data were scrubbed out before further processing and feature extraction. We then utilized the Universal Sentence Encoder-large [28] to encode these responses into dense vector representations. Unlike other embedding techniques that encode text at the word level, the Universal Sentence Encoder is an encoder of greater-than-word texts trained on a variety of data using the Transformer architecture to encode text that is more significant than word lengths such as sentences. Hence, it has been adapted as the sentence embedder for this experiment.

For this experiment, emphasis was placed specifically on the responses derived from the participants while completely ignoring or cutting out the questions from the virtual interviewer ''Ellie''. This is to enhance the possibility of detecting depression symptoms in patients in common situations where preset questions are not necessarily asked.

After performing feature extraction as described above, 128 feature embeddings were extracted and passed through two BiLSTM layers followed by an attention layer to assign weights to the most relevant elements. This is then fed into an FC network consisting of two linear layers and the ReLu activation function is applied to the output of the second linear layer to identify participants' binary and PHQ-8 scores on the text model. A summary of the parameter setting of the BiLSTM with attention model is shown in Table 2.

2) AUDIO MODEL ANALYSIS

Just like the text feature preprocessing, audio data corresponding to each participant were read, and audio segments were extracted based on information from the transcripts of the dataset. However, the audio data was split into

15-second segments using its sampling rates to ensure that each segment spans 15 seconds, a duration that aligns with relevant patterns in the audio data. During the training phase, overlapping segments of 15 seconds with adjacent 14-second overlap are generated to allow for the inclusion of diverse audio patterns, thereby providing a richer representation of depression-related features for the depression class. Using the Librosa library [29], a powerful Python package for audio signal processing, Mel spectrogram features were then extracted from each audio segment with 80 Mel filters. The extracted audio features were then fed into two BiLSTM layers and an attention layer similar to the implementation conducted with the text features.

Adam optimizer was used during the training process with a batch size of 8 for 100 epochs, and a learning rate of 1e-4. On the multimodal network, a batch size of 2 was utilized. The 128 feature embeddings produced by the BiLSTM layers of both the text and audio models are concatenated and fed into the ACMA network as implemented above.

V. RESULTS

Depression moves across the spectrum, so deriving a binary state (depressed not depressed) from a single test (PHQ-8) might seem a bit unrealistic. As people experience varying moods of depression daily, this does not directly translate to such a person being clinically depressed or having MD. Considering that the DAIC-WOZ dataset is originally a binary labeled dataset, and it is relatively small, with a higher percentage of the participants being not depressed, it suffers greatly from the well-known data imbalance problem in this area. A possible solution to this problem is to bin such datasets into the required number of classes and randomly generate samples to be used for the experiment. However, with the DAIC-WOZ dataset, binning and generating data samples essentially synthesizes new samples that follow the same distribution as the original dataset and are not representative of the underlying true data distribution, which potentially leads to unrealistic and irrelevant redundant samples.

For this purpose, we conducted classification and regression experiments using the provided PHQ-8 binary labels, and PHQ-8 scores respectively to identify a patient's depression state as well as estimate their levels of depression.

A. CLASSIFICATION

The classification model was trained specifically using the PHQ-8 binary labels to predict the depressive state of a patient (depressed/not-depressed) from the provided training set of the DAIC-WOZ dataset evaluated on the development set as the test set had no publicly available labels at the time of this experiment. Various research has implemented the use of different types of data modalities and features for depression detection [6], [30], [31]. However, we opted for the text and audio modalities as they are the most common forms of data models used daily. Table 3 shows the results of our comparison with baseline models and closely related research. We compared our results with three text models and four audio models. In the end, our multi-modal ACMA network performance was also compared with two results where multimodality of audio and text features was utilized. For fairness, it is worth noting that all the experimental results used for comparison in this study were conducted with the DAIC-WOZ dataset.

TABLE 3. Comparison with closely related research on the DAIC-WOZ dataset.

		Classification			Regression
Model	Method	F1-score	Precision	Recall	MAE
T	Al Hanai et al. [18]	0.44	**1.00**	0.29	7.32
	Sun et al. [7]	0.55	0.89	0.40	4.98
	Lam et al. [32]	0.45	0.37	0.58	-
	Our BiLSTM + Attn.	**0.78**	0.80	**0.76**	**3.96**
A	Valstar et al. [10]	0.46	0.32	**0.86**	5.36
	Al Hanai et al. [18]	0.67	**1.00**	0.50	7.60
	Wei et al. [30]	0.61	0.59	0.66	5.19
	Lam et al. [32]	0.56	0.44	0.78	-
	Our BiLSTM + Attn.	**0.73**	0.78	0.72	**5.08**
A/T	Al Hanai et al. [18]	0.77	**0.83**	0.71	5.10
	Lam et al. [32]	0.67	0.60	0.75	-
	Our ACMA Net.	**0.82**	0.79	**0.86**	**4.65**

T= Text, A= Audio, A/T= Audio and Text, MAE= Mean Absolute Error

The table shows the comparison results for the classification and regression experiments conducted. The best performances are shown in bold text.

In the experiment with the classification model, we scored its performance using the most common metrics: Precision, Recall, and F1-Score which were used by the other researchers for easy comparison. The values of precision, recall, and F1-score are computed from the True positive, False positive, False negative, and True negative values of the confusion matrix therefore, Fig. 4 shows our confusion matrix for the results of the experiment with the DAIC-WOZ dataset. From Table 3, it can be seen that in the text feature experiment, our BiLSTM with attention model performs best with an F1-Score of 78% and Recall of 76%. Compared to the text feature experiment, our proposed BiLSTM with attention model for the audio feature experiment performance dwindles back, however still outperforms the other results on F1-Score with 73%.

Compared with the experiment conducted by [18] and [32], our proposed multi-modal fusion network performs exceptionally well on F1-Score and Recall, achieving performance levels of 82% and 86%, respectively. It is worth noting that our models achieved accuracies of 72.7% in text, 69.7% in audio, and 75.8% in audio and text fusion modalities, on the DAIC-WOZ dataset (Fig. 6).

B. REGRESSION

With the regression model, we considered the threshold for a patient regarded as depressed or not depressed using the PHQ-8 score labels provided by the authors of the dataset. With the PHQ-8 score, a score from 1-9 is regarded as

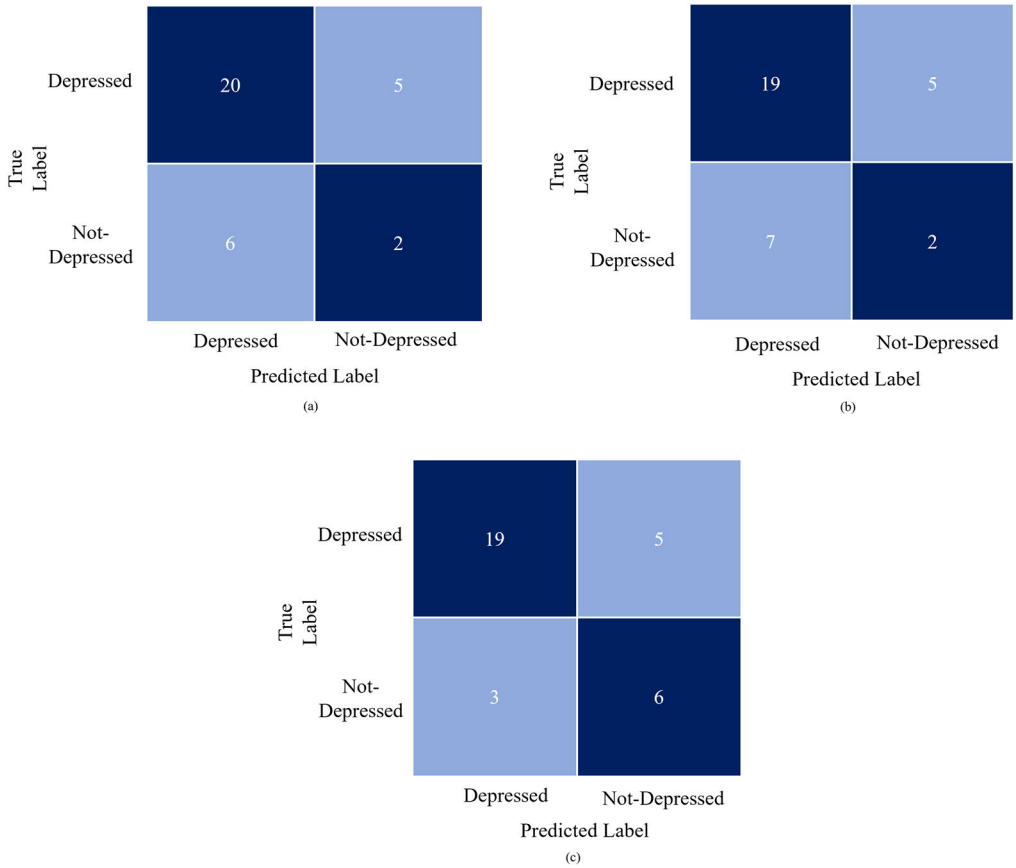


FIGURE 4. Confusion matrix of experiment on DAIC-WOZ dataset. (a) Our Text BiLSTM + attn. (b) Our Audio BiLSTM + attn., and (c) Our ACMA net.

TABLE 4. Performance with Attention and Without Attention on the BiLSTM Models with the DAIC-Woz dataset to Emphasize Model Complexity.

	Modality	F1-score	Precision	Recall	WA vs WoA
WA	Text	0.79	0.81	0.77	WA > WoA in all cases
	Audio	0.73	0.78	0.72	
	ACMA Net	0.80	0.76	0.85	
WoA	Text	0.78	0.80	0.76	
	Audio	0.71	0.76	0.71	
	ACMA Net	0.78	0.75	0.83	

WA = BiLSTM with Attention, WoA = BiLSTM without Attention

not depressed, and a score between 10-24 is regarded as depressed. However, logically, depression rates can vary between the determined depression score of 10-24, as well as the not-depressed score of 1-9. For this reason, we extended our experiment to include a regression technique to predict participants' PHQ-8 scores, thereby scoring the regression model based on Mean Absolute Error (MAE).

Table 3 also shows the result of our experiment in this regard. Following the same model architecture our audio model showed a good regression performance of 5.08 MAE compared to the research conducted by [10], [18], [30], and [33]. When only the text features are considered, our method outperforms the rest with a prediction of 3.96 MAE. The proposed ACMA multi-modal fusion network produces a result with an MAE of 4.65 also outperforming the result recorded by the two compared research.

These results indicate that our proposed ACMA network can effectively detect depression in patients with high levels of accuracy.

C. EFFECTS OF ATTENTION WEIGHTS ON THE OVERALL MODEL

The purpose of the attention mechanism added to the BiLSTM layer is to mainly identify the influence of each word on every sentence noted. This is done by assigning attention weights to each word to capture important components of the sentence semantics thereby improving model accuracy.

To validate this, using the classification model alone, we trained the audio and text models separately, and then the entire ACMA network by selectively excluding the attention mechanism from BiLSTM on both the audio and text models. The result of this experiment can be seen in Table 4. Compared to our model performance comparison results shown in Table 3, it is evident that the inclusion of the attention mechanism has a great influence on both the single and multimodal

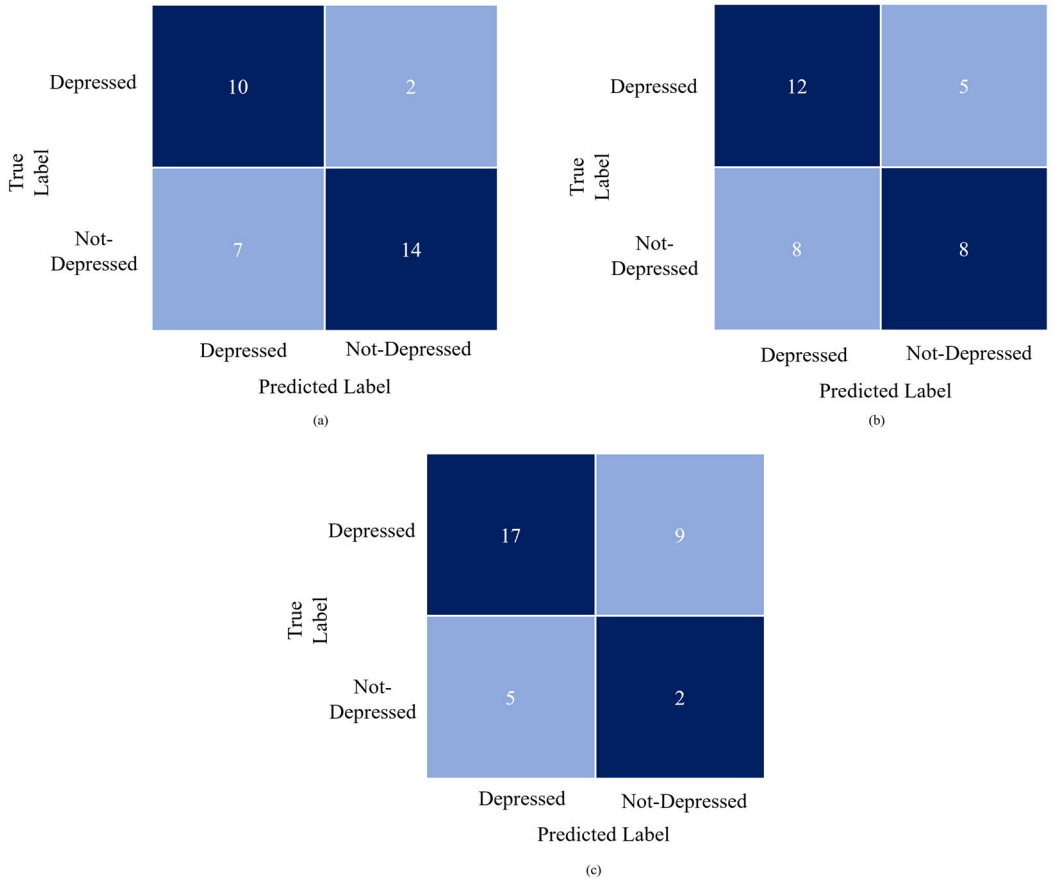


FIGURE 5. Confusion matrix of experiment on EATD-corpus. (a) Our Text BiLSTM + attn. (b) Our Audio BiLSTM + attn., and (c) Our ACMA Net.

models and shows a high model complexity compared to other research with plain BiLSTM models. The main reason for the less high performance noted with only BiLSTM is that LSTM models are biased models. Although they can address challenges related to long-time lags thanks to their gating mechanism, the prominence of more recent words over earlier ones makes it challenging to recognize longer sentences [34].

D. MODEL EVALUATION ON A SECONDARY DATASET

To provide a more robust estimate of our model's performance generalization ability to new unseen data, we performed a model evaluation on the EATD-corpus dataset. However, the EATD-corpus dataset also suffers greatly from imbalanced class problem which tends to greatly affect the performance of machine learning algorithms. We adopt the Stratified 3-Fold cross-validation technique to further evaluate our model with EATD-Corpus and then we observe the split results and perform random oversampling to improve the class imbalance problem of the dataset.

3×256 and 3×512 embeddings of audio and text are extracted respectively from the training and test sets of the dataset and our proposed BiLSTM models for both audio and text are trained on the extracted embeddings. Following this, the generated representations are concatenated and passed to the fusion model as described in section IV above. The results of our model evaluation on the EATD-corpus can be seen in Table 5 and the confusion matrix in Fig. 5 below.

TABLE 5. Evaluation Performance on the EATD-Corpus with Baseline Research.

Model	Method	F1-score	Precision	Recall
T	Al Hanai et al. [18]	0.57	0.53	0.63
	Shen et al. [35]	0.65	0.65	**0.66**
	Our BiLSTM + Attn.	**0.66**	**0.79**	0.58
A	Al Hanai et al. [18]	0.49	o.44	0.56
	Shen et al. [35]	**0.66**	0.57	**0.78**
	Our BiLSTM + Attn.	0.65	**0.70**	0.60
A/T	Al Hanai et al. [18]	0.57	0.49	0.67
	Shen et al. [35]	**0.71**	0.62	**0.84**
	Our ACMA Net.	0.70	**0.65**	0.7

In the text modality, our proposed BiLSTM + Attn. model outperforms the two baseline models with an F1-score of 66%, demonstrating a good balance between precision (79%) and recall (58%) with an accuracy of 60.6%. This proves that our text model can effectively identify relevant instances while minimizing both false positives and false negatives.

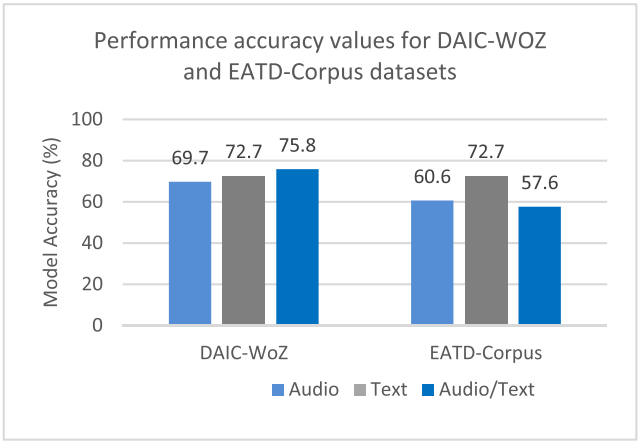


FIGURE 6. Performance accuracy values for DAIC-WOZ and EATD-Corpus datasets.

For the audio modality, an accuracy of 72.7% was realized, and our BiLSTM + Attn. model shows competitive performance with an F1-score of 65%, comparable to the approach, which achieves a slightly higher F1-score of 66%. Our model however exhibits a higher precision of 70% but a slightly lower recall of 60%, reflecting a trade-off similar to that observed in the text modality.

With the audio and text fusion modality, our proposed ACMA Net. model achieves an F1-score of 70%, trailing slightly behind Shen et al.'s approach at 71%. However, our model demonstrates a notable improvement in precision by 65% compared to Shen et al.'s 62%, and 49% by Al Hanai et al. thereby suggesting a reduction in false positives, while maintaining a competitive recall of 78% and an overall accuracy of 57.6%.

The results from the evaluation of our method indicate that our model possesses a high generalization ability, therefore it can be applied to different depression datasets which are characterized by diverse demographics, clinical features, and data collection sources. The performance demonstrated across varied datasets as can be seen in Fig. 6 suggests that our model is not overly sensitive to specific nuances present in the training data, making it adaptable to a broader range of scenarios. This capability enhances the practical utility of our model, opening avenues for its application in real-world settings where depression data may exhibit inherent heterogeneity. The versatility of our approach positions it as a valuable tool for researchers, clinicians, and healthcare professionals seeking reliable and generalized insights into depression across diverse populations and contexts.

VI. CONCLUSION

For various reasons, patients tend to give false reports on their mental status when answering questionnaires during depression screening, this tends to result in wrong diagnosis thereby putting such a patient at greater risk of the highly destructive state of suicidal ideations and suicide attempts. It is therefore important to develop methods that can be used to counter such negatives that come with using preset depression questionnaires. It is in this regard that we propose a variant of the cross-modal attention network to effectively capture telltale symptoms of depression from patients' speech and text.

Our proposed method follows a simple encoding of audio and text features into embeddings while completely excluding the questions asked during the interview conducted while performing data collection. The outputs of both modalities are fused and passed to our proposed Additive Cross-modal attention network which employs additive attention to calculate the attention weights before performing cross-modal fusion. While our model showed encouraging results that are comparable to other research in this area, we believe that our proposed method fulfills our primary goal, can effectively detect depression in patients, and can be applied in both unimodal cases (text or audio) as well as multimodal cases (text and audio) without the use of preset questionnaires.

For future research, more focus will be laid on solving the data imbalance problem of depression datasets to enhance system accuracy and performance using methods such as the Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) to generate synthetic instances in underrepresented classes in the dataset. In the future, we also aim to apply our proposed model in a real-world application that can be used by therapists and psychological professionals in the proper diagnosis of depression by explicitly identifying depressed patients from speech and texts without the necessary use of preset questionnaires.

REFERENCES

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006.

[2] M. D. Daniel and K. Hall-Flavin. (2023). *Clinical Depression: What Does that Mean?* Accessed: Aug. 10, 2023. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/depression/expert-answers/clinical-depression/faq-20057770#:~: text=What%20does%20the%20term%20%22clinical,depression%20or%20major%20depressive%20disorder

[3] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, "A review of machine learning and deep learning approaches on mental health diagnosis," *Healthcare*, vol. 11, no. 3, p. 285, Jan. 2023.

[4] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, 2014, pp. 3123–3128.

[5] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0224365.

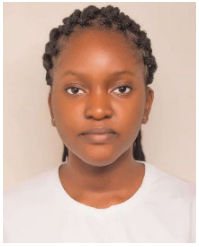
[6] F. Ceccarelli and M. Mahmoud, "Multimodal temporal machine learning for bipolar disorder and depression recognition," *Pattern Anal. Appl.*, vol. 25, no. 3, pp. 493–504, Aug. 2022.

[7] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 61–68.

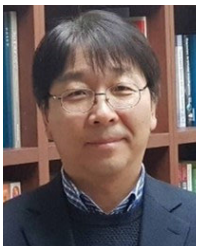
[8] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 89–96.

[9] W. Oh, "Comparison of environmental sound classification performance of convolutional neural networks according to audio preprocessing methods," *J. Acoust. Soc. Korea*, vol. 39, no. 3, pp. 143–149, 2020.

[10] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 3–10.
[11] 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*, Oct. 2016, pp. 35–42.
[12] B. J. Shannon and K. K. Paliwal, "A comparative study of filter bank spacing for speech recognition," in *Proc. Microelectronic Eng. Res. Conf.*, 2003, pp. 310–312.
[13] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using convolutional autoencoder," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116076.
[14] K. N. Phan, N. K. Iyortsuun, S. Pant, H.-J. Yang, and S.-H. Kim, "Pain recognition with physiological signals using multi-level context information," *IEEE Access*, vol. 11, pp. 20114–20127, 2023.
[15] 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," (in English), *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104499, doi: 10.1016/j.compbiomed.2021.104499.
[16] F. M. Shah, F. Ahmed, S. K. S. Joy, S. Ahmed, S. Sadek, R. Shil, and M. H. Kabir, "Early depression detection from social network using deep learning techniques," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jun. 2020, pp. 823–826.
[17] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 45–51.
[18] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.
[19] X. Lu, D. Shi, Y. Liu, and J. Yuan, "Speech depression recognition based on attentional residual network," *Frontiers Bioscience-Landmark ed.*, vol. 26, no. 12, pp. 1746–1759, Dec. 2021, doi: 10.52586/5066.
[20] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, Oct. 1995.
[21] F. Cutzu and J. K. Tsotsos, "The selective tuning model of attention: Psychophysical evidence for a suppressive annulus around an attended item," *Vis. Res.*, vol. 43, no. 2, pp. 205–219, Jan. 2003.
[22] J. Amudha and K. Soman, "Selective tuning visual attention model," *Int. J. Recent Trends Eng.*, vol. 2, no. 2, p. 117, 2009.
[23] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[25] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[26] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disorders*, vol. 114, nos. 1–3, pp. 163–173, Apr. 2009.
[27] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: A BiLSTM/1D CNN-based model," *Appl. Sci.*, vol. 10, no. 23, p. 8701, Dec. 2020.
[28] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," 2018, *arXiv:1803.11175*.
[29] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.
[30] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Multi-modal depression estimation based on sub-attentional fusion," 2022, *arXiv:2207.06180*.
[31] G. Sun, S. Zhao, B. Zou, and Y. An, "Multimodal depression detection using a deep feature fusion network," in *Proc. 3rd Int. Conf. Comput. Sci. Commun. Technol. (ICCSCT)*. Bellingham, WA, USA: SPIE, Dec. 2022, pp. 1571–1576.
[32] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 3946–3950.
[33] 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*.
[34] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, Apr. 2018.
[35] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model," 2022, *arXiv:2202.08210*.



NGUMIMI KAREN IYORTSUUN received the B.S. degree from the Department of Mathematics and Computer Science, University of Mkar, Nigeria, in 2019. She is currently pursuing the integrated M.S./Ph.D. degree with the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea. She joined the Pattern Recognition Laboratory, in March 2022. Her research interests include depression intensity and treatment outcome estimation, pattern recognition, and computer vision.



SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Artificial Intelligence, Chonnam National University, South Korea. His research interests include video understanding, multi-modal emotion recognition, medical image analysis, and document image processing.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chungbuk National University, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

SEUNG-WON KIM received the bachelor's and master's degrees from the University of Tasmania, in 2008 and 2010, respectively, and the Ph.D. degree from the HIT Lab NZ, New Zealand, in 2016, under the supervision of Prof. Mark Billinghurst. He is an Assistant Professor with Chonnam National University. Currently, he is also running the Empathic Computing Laboratory in South Korea (ECL KR). He has more than 20 publications in remote collaboration studies at several notorious journals and conferences. His research interests include remote collaboration using augmented virtual communication cues and sharing experience/emotion between distance users.

MIN JHON received the B.S. and M.S. degrees in medicine from Chonnam National University, South Korea, in 2013 and 2017, respectively. Since 2020, she has been a Clinical Professor with the Chonnam National University Hwasun Hospital, South Korea. She was a Chief of the Metropolitan Mental Health Welfare Center, Gwangju, South Korea, from 2019 to 2022. Her research interests include digital biomarker research in psychiatry, geriatric psychiatry, and psychotherapy.