



# DepressGEN: Synthetic Data Generation Framework for Depression Detection

Wenrui Liang<sup>1,\*</sup>, Rong Zhang<sup>1,\*</sup>, Xuezhen Zhang<sup>2</sup>, Ying Ma<sup>3</sup>, Wei-Qiang Zhang<sup>1,#</sup>

<sup>1</sup>Department of Electronic Engineering, Tsinghua University, China

<sup>2</sup>Zhili College, Tsinghua University, China

<sup>3</sup>Shanghai Jiao Tong University, China

lwr24@mails.tsinghua.edu.cn, wqzhang@tsinghua.edu.cn

## Abstract

Automated depression detection is vital for early diagnosis, but ethical and privacy concerns often limit the availability of sufficient training data, hindering research in depression screening. To address this, we introduce DepressGEN, a novel framework that generates synthetic interview dialogue texts and speech simulating depressed patients to improve training for detection models. By inputting linguistic features associated with depression into a large language model, we create dialogue texts and use a TTS system to generate corresponding speech. We also developed a depression modulation module to modify the synthesized speech, as well as a speech verification module to bridge the gap between synthetic and real data distributions. Our results demonstrate that a GRU/BiLSTM-based model trained with additional synthetic data improves F1 scores by 9.9% compared to the same model trained only on original data, outperforming existing methods on the EATD dataset.

**Index Terms:** depression detection, data generation, multimodal

## 1. Introduction

Depression is a widespread mental health disorder that affects millions globally, significantly impacting individuals' quality of life, productivity, and overall well-being [1, 2]. Traditional diagnostic methods, such as clinical interviews and self-reported questionnaires, are often time-consuming and require specialized expertise [3]. This creates an urgent need for machine learning techniques aimed at automated depression detection.

Recent research has significantly advanced depression detection through the integration of multimodal features. These studies leverage information from text, audio, and video to enhance accuracy. For instance, Gong et al. [4] employed an ensemble method combining audio, video, and semantic features with topic modeling to analyze interview data. Building on this, Al Hanai et al. [5] demonstrated the effectiveness of LSTMs in modeling sequential relationships between audio and text, allowing for depression detection with minimal structural information. Wu et al. [6] improved multimodal fusion by incorporating pre-trained foundational models and self-supervised learning for refined emotion and text representations, enhancing classification accuracy. Yang et al. [7] refined audio and text features using expert knowledge before applying late fusion, which boosted overall performance. Similarly, Iyortsuun et al. [8] introduced a cross-modal attention network (ACMA)

with BiLSTMs, capturing complex interactions between audio and text. Zhang et al. [9] advanced this work by integrating acoustic landmarks into a large language model (LLM) framework, transforming speech signals into linguistically meaningful tokens for more robust detection. Most recently, Xue et al. [10] proposed the Channel Attention-based Multimodal Fusion Module (CAMFM), which effectively integrates low-level audio features, wav2vec representations, and BERT-derived embeddings, providing a comprehensive approach to improving depression detection.

Despite the potential of multimodal approaches, obtaining high-quality, large-scale depression datasets remains challenging due to the sensitive nature of the data, which raises ethical concerns and demands strict privacy protections. Additionally, the high costs of data annotation and management further impede the development of suitable datasets. Consequently, depression datasets are scarce [11, 12]. Currently, only a few public databases for depression recognition exist, including EATD [13] and DAIC-WOZ [14]. The EATD dataset features multimodal data (text and speech), while DAIC-WOZ includes text, speech, and video. However, both are limited in size and the number of depression patients is smaller.

The rapid advancement of LLMs and text-to-speech (TTS) systems has made the creation of synthetic training data feasible. Synthetic data methods are widely used in automatic speech recognition [15, 16, 17], and similar approaches have been explored in computer vision to enhance model performance [18, 19, 20]. Inspired by these methods, we aim to leverage generated data from individuals with depression to improve depression detection models.

We propose an innovative approach to generating training data informed by existing medical knowledge about depression. By utilizing LLMs like ChatGPT for text generation and TTS systems like ChatTTS for speech synthesis, we create synthetic samples that realistically reflect the symptoms of depression. To the best of our knowledge, we are the first to generate depression-related data using LLMs and TTS systems.

Our research focuses on language and speech features linked to depression. Linguistic traits include frequent use of negative vocabulary, expressions of pessimism, and reduced lexical diversity. Key acoustic indicators include slower speech rates, lower volume, monotonous tone, and limited pitch variation. However, synthetic datasets may introduce artifacts not present in real data. To address this, we drew inspiration from Generative Adversarial Networks (GANs) [21] and trained a speech verification module to align the distributions of synthetic and real audio data. The baseline model for depression detection utilizes a GRU/BiLSTM architecture [13]. Moreover, our method could be adapted to other areas of mental health detection, effectively addressing data scarcity.

\*These authors contributed equally.

#Corresponding author.

This work was supported by the National Natural Science Foundation of China under Grant No. 62276153.

## 2. Proposed Method

### 2.1. Data Generation Pipeline

Figure 1 illustrates the workflow of *DepressGEN*, which comprises three key steps. First, we create an outline of questions that include all the questions from the original dataset, as the depression dataset is organized based on clinical interviews. Then, we provide this question outline along with language features related to depression to ChatGPT, so that it can respond in a language style similar to that of individuals with depression, thereby generating simulated text data representing a depressed patient. Second, the ChatTTS module converts the generated text into audio, which is then modified by a depression modulation module to incorporate characteristics such as flat volume, dull voice, and slower speech rate. We also apply speech verification to ensure the high quality of the synthetic data. Lastly, the synthesized text and audio are integrated with the original dataset to train a multimodal depression detection model. The modular design of this detection model offers flexibility for using alternative architectures.

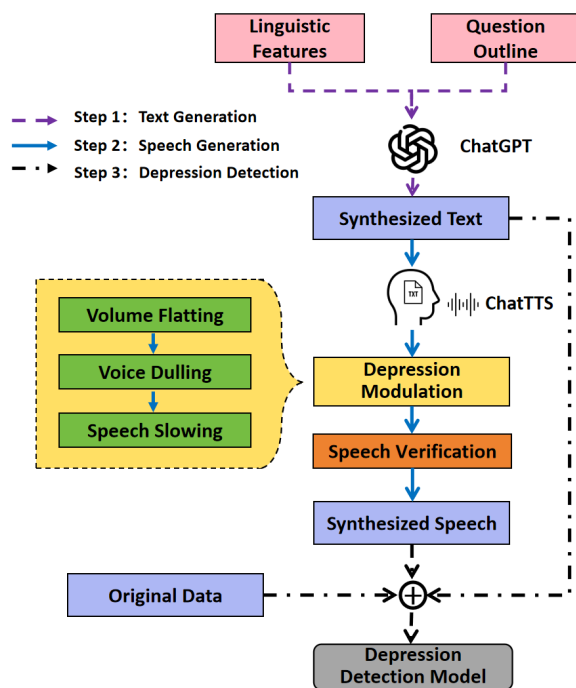


Figure 1: Overview of *DepressGEN*

### 2.2. Text Generation

The text generation process produces realistic depressive text samples by leveraging a knowledge base that encapsulates key linguistic and emotional traits associated with depression—such as negative self-perception, hopelessness, emotional distress, and worthlessness, which is summarized from medical literature. These traits serve as guiding principles and are dynamically embedded into each generated dialogue to simulate authentic depressive language patterns.

We employ OpenAI’s GPT-4o-mini model for text generation. The LLM is prompted with a combination of predefined questions and depression-related linguistic cues from the

knowledge base. The LLM is instructed to answer the preset questions while integrating elements from the knowledge base of depression, which includes examples of depression characteristics such as:

- **Self-Reflective Monologues:** These monologues mirror the introspective and self-critical nature of individuals with depression, often expressing feelings of guilt, regret, and negative self-evaluation [22].
- **Dreading the Unknown:** Individuals with depression often exhibit pessimistic or fatalistic reactions to everyday situations, emphasizing cognitive patterns like rumination and catastrophizing [23, 24].
- **Emotionally Negative Dialogues:** These dialogues express sadness, frustration, hopelessness, and emotional isolation, reflecting the emotional depth of depressive speech [25].
- **Absolutist Language:** Individuals with depression frequently use absolutist terms such as “always”, “never”, and “nothing”, reflecting cognitive distortions associated with all-or-nothing thinking. This black-and-white thinking reinforces feelings of worthlessness and hopelessness, common in depressive states [26].
- **Poor Physical Symptoms:** Depression often manifests through significant physical symptoms, which are frequently emphasized in the speech of affected individuals. This focus on somatic complaints, such as persistent fatigue and sleep disturbances, often overshadows emotional symptoms [27].

### 2.3. Speech Generation

The speech generation process produces realistic audio that mimics the acoustic features of depression, such as flat intonation, dull voice, slower speech rate, and frequent pauses, indicative of diminished emotional expression, energy, and cognitive fluency [28, 29]. By utilizing ChatTTS for speech synthesis, we incorporated parameters to control pauses and generate voices from various speakers. To further replicate depressive speech, we created a depression modulation module and implemented a speech verification module to ensure the synthesized speech closely aligns with the data distribution of real speech.

#### 2.3.1. Speech Synthesis with ChatTTS

ChatTTS is a deep learning model that produces natural, contextually relevant speech from preprocessed text by integrating prosodic features such as pitch, tone, and rhythm. The model can emulate different speaker characteristics through speaker embedding and control pauses in the generated speech by pause insertion parameters.

- **Speaker Embedding:** Speaker embeddings capture a speaker’s unique vocal characteristics, enabling personalized and diverse speech. They are created by sampling Gaussian noise to form a fixed-length speaker vector, which is integrated into the network and stored as feature vectors in JSON files. During speech generation, the relevant embedding is retrieved and incorporated into the synthesis model to convey the speaker’s vocal style.
- **Pause Insertion:** ChatTTS improves speech delivery by adding pauses that mimic natural speech patterns and emotional nuances. In our speech synthesis process, we set the pause insertion parameter to a high value to generate speech with more pauses, which better aligns with the speech characteristics of individuals with depression.

### 2.3.2. Depression Modulation

We enhance the speech generated by ChatTTS by incorporating a depression modulation module, giving it a more depressive tone. This module incorporates flat intonation, dull voice, and slower speech rates through dynamic range compression, low-pass filtering, and speech rate adjustments, working alongside built-in pause control to produce slower, deeper, and more melancholic speech.

- **Volume Flattening:** Individuals with depression often exhibit flattened vocal volume. This effect can be replicated using dynamic range compression, which lessens the difference between maximum and minimum volume levels. First, we compute the gain reduction  $G(t)$  with a threshold  $T$  and a compression ratio  $R$  as follows [30]:

$$G(t) = \begin{cases} 0, & L(t) \leq T \\ (1 - \frac{1}{R})(L(t) - T), & L(t) > T \end{cases} \quad (1)$$

where  $L(t)$  is the log-RMS loudness of each frame. Then, the gain reduction  $G(t)$  is applied to the waveform, and the processed signal  $y(t)$  is obtained:

$$y(t) = x(t) \cdot 10^{-G(t)/20} \quad (2)$$

- **Voice Dulling:** A common vocal characteristic of patients with depression is that their voice are dull, which can be simulated using a low-pass filter (LPF). The LPF filters out the high-frequency components of the voice signal, causing the voice to sound deeper. The frequency response of the LPF is defined as:

$$H(f) = \begin{cases} 1, & |f| \leq f_c \\ 0, & |f| > f_c \end{cases} \quad (3)$$

where  $f_c$  is the cutoff frequency. The filtered signal  $y(t)$  is computed as:

$$y(t) = \mathcal{F}^{-1} [H(f) \cdot \mathcal{F}[x(t)]] \quad (4)$$

- **Speech Slowing:** Depression is also marked by a slower speech rate. We employ time-stretching to adjust audio duration. Faster speech segments, identified through short-time energy analysis, are selectively lengthened, while slower parts remain intact. This dynamic adjustment effectively simulates speech patterns associated with depression. The effect is achieved by modifying the audio signal's speed using a factor  $\alpha$ , which changes the duration according to:

$$y(t) = x(\alpha t) \quad (5)$$

where  $x(t)$  is the original signal,  $y(t)$  is the adjusted signal, and  $\alpha$  is the speed factor.

### 2.3.3. Speech Verification

We use a speech verification module to evaluate whether the generated voice samples authentically exhibit characteristics of depression. To determine whether to retain or discard these features, we employed a convolutional neural network (CNN) for depression prediction [31], which was trained on the original dataset, namely the EATD dataset. Once trained, we input the generated audio data into the model to obtain the probability that each sample indicates depression. We then generate a random number  $\theta$  between 0 and 1; if this number is less than the

predicted probability, we keep the sample; otherwise, we discard it. The algorithm is defined as follows:

$$\begin{cases} p(x_g) = F_{\text{cnn}}(x_g) \leq \theta & \text{keep } x_g, \\ p(x_g) = F_{\text{cnn}}(x_g) > \theta & \text{discard } x_g. \end{cases} \quad (6)$$

Here,  $p(x_g)$  represents the probability that the generated audio sample  $x_g$  suggests depression, as predicted by the CNN model  $F_{\text{cnn}}$ . The process concludes after accepting  $N$  synthetic samples.

## 2.4. Depression Detection Model

The depression detection model utilized in this study comprises a GRU model and a BiLSTM model [13], enhanced by an attention layer that summarizes both audio and text representations. These summaries are concatenated and fed into a single-layer fully connected (FC) network. Modal attention, indicated by a trained weight vector, evaluates the significance of each modality, ultimately generating a binary label for the presence of depression. Text features are derived from high-dimensional embeddings through ELMo [32], while audio features are extracted from Mel spectrograms. In line with the approach in [13], we rearranged and resampled responses from depressed volunteers in the training set.

We also employed three-fold cross-validation, dividing volunteers into three groups: two groups for training and one group for testing. To ensure fairness, the generated data is included only in the training set, while the test set consists entirely of original data.

## 3. Experiment

### 3.1. Dataset

The EATD dataset [13] comprises audio and transcript recordings from 162 student volunteers in China. Participants answered three randomly selected questions and completed the Self-rating Depression Scale (SDS) questionnaire [33], with scores of 53 or higher indicating depression [34]. Of the 162 participants, 132 were classified as non-depressed, while 30 were identified as depressed.

### 3.2. Experimental Settings

We trained the depression detection model on an NVIDIA GeForce RTX 3090 GPU, using the F1 score as the evaluation metric, where a higher score indicates better performance. We generated text and audio data for 48 patients with depression, 1.6 times the number in the original dataset. Table 1 outlines the parameters used in the speech synthesis process, with parameters  $T$  and  $R$  defined in formulas (1), and  $f_c$  and  $\alpha$  specified in formulas (3) and (5) of section 2.3.2.

Table 1: *The Parameters for Synthesizing Depression Speech*

Params	Baseline	Our Experiment
Threshold $T$	–	–20 dB
Compression ratio $R$	–	3
Cutoff frequency $f_c$	–	3000 Hz
Speed factor $\alpha$	–	0.8
Pause insertion level	–	7
Number of speakers	162	162 + 48

## 4. Results and Discussion

### 4.1. Comparison with Original Data

Table 2 presents a comparison of F1 scores across three features: text, audio, and fusion. The baseline reflects the performance of the GRU/BiLSTM-based model trained solely on the original dataset, while “Ours” denotes the performance of the model trained with both synthetic and original data. For text features, the baseline achieved an F1 score of 0.65, which our method improved to 0.75, indicating a significant increase of 15.4%. In audio features, the baseline score was 0.66, with our method raising it to 0.69, reflecting an improvement of 4.6%. Though this increase is modest, it still signifies progress across modalities. In fusion feature, the baseline score of 0.71 was enhanced to 0.78 by our method, marking a 9.9% increase. These results suggest that combining information from different modalities allows synthetic data to improve overall performance. Overall, the comparisons reveal varying degrees of improvement in text, audio, and fusion features, with text showing the most significant gain. These findings highlight the effectiveness of synthetic data in enhancing the performance of depression detection models and its potential for training applications.

Table 2: F1 Score Comparison with Original Data

Features	Baseline	Ours	Improvement
Text	0.65	0.75	15.4%
Audio	0.66	0.69	4.6%
Fusion	0.71	0.78	9.9%

### 4.2. Comparison with Other Works

The performance of the GRU/BiLSTM-based model trained with our additional synthetic data is compared to that of other models trained with the original dataset, as illustrated in Table 3. This performance comparison provides valuable information on the effectiveness of different methods in terms of audio, text, and fusion features.

When evaluated on the audio modality, CAMFM stands out with the highest F1 score of 0.73. Our method achieves a competitive F1 score of 0.69, positioning it as the second-best option in this category and outperforming the ACMA method and multimodal LSTM, which scored 0.65 and 0.49, respectively. Overall, in the audio modality, although our method did not achieve the best performance, we still obtained decent results.

When evaluated on the text modality, our method outperforms all competitors with an impressive F1 score of 0.75. This surpasses CAMFM, which obtained a score of 0.72, as well as ACMA and the multimodal LSTM, which achieved scores of 0.66 and 0.57 respectively. Compared to the closest competitor CAMFM, our method shows a notable improvement, solidifying its state-of-the-art performance in this modality. Similarly, in the fusion category, our method again leads with the highest F1 score of 0.78, slightly outperforming CAMFM which achieved 0.77. The consistent performance gain over all other baselines demonstrates the effectiveness of our multimodal integration strategy.

Overall, the findings indicate that our method consistently yields high performance across all feature categories, particularly in text and fusion. This suggests promising avenues for the further application of our model in relevant fields.

Table 3: Performance Comparison of Different Methods

Features	Methods	F1-Score	Recall	Precision
Audio	multimodal LSTM [5]	0.49	0.56	0.44
	ACMA [8]	0.65	0.60	<b>0.70</b>
	CAMFM [10]	<b>0.73</b>	<b>0.84</b>	0.64
	Ours	0.69	0.80	0.61
Text	multimodal LSTM [5]	0.57	0.63	0.53
	ACMA [8]	0.66	0.58	<b>0.79</b>
	CAMFM [10]	0.72	<b>0.80</b>	0.66
	Ours	<b>0.75</b>	0.78	0.72
Fusion	multimodal LSTM [5]	0.57	0.67	0.49
	ACMA [8]	0.70	0.70	0.65
	TAMFN [35]	0.75	0.85	0.69
	CAMFM [10]	0.77	0.84	<b>0.73</b>
	Ours	<b>0.78</b>	<b>0.86</b>	0.71

### 4.3. Ablation Study

Table 4 presents results from an ablation study on speech verification, showing the performance of the F1 score for audio and fusion features under “w/o Verif.” and “w/ Verif.” conditions. The only distinction between these conditions is whether the speech was subjected to the verification module, while the text data and the depression modulation parameters remained consistent. In the audio modality, the F1 score under “w/o Verif.” conditions is 0.66, consistent with the baseline, indicating a sub-par quality of the synthesized speech data. When transitioning to “w/ Verif.” conditions, the F1 score improved from 0.66 to 0.69, a 4.6% gain that highlights the contribution to verification accuracy. The fusion modality also shows significant improvement. The F1 score of fusion modality starts at 0.75 under “w/o Verif.” condition and rises to 0.78 under “w/ Verif.” condition, marking a 4.0% improvement. The speech verification module thus enhanced the final fusion effect. Overall, the performance improvement in both speech and fusion modalities underscores the importance of the speech verification module in speech generation.

Table 4: F1 Score with and without Speech Verification

Features	w/o Verif.	w/ Verif.	Improvement
Audio	0.66	0.69	4.6%
Fusion	0.75	0.78	4.0%

## 5. Conclusions

In this article, we presented *DepressGEN*, a framework designed to synthesize text and speech data related to depression to address the issue of data scarcity. We input linguistic features associated with depression into ChatGPT to generate dialogue text and utilize ChatTTS for speech synthesis. We developed a depression modulation module to fine-tune synthesized speech, while a speech verification module makes the distribution of the synthesized data closer to that of real data. Ablation experiments demonstrated the effectiveness of the speech verification module. By augmenting the EATD dataset with synthetic samples, we achieved significant improvements in model performance. Specifically, a GRU/BiLSTM-based model trained on synthetic data and original data exhibited a 9.9% increase in F1 scores compared to the same model trained exclusively on original data, outperforming existing methods on the EATD dataset.

## 6. References

- [1] R. Peveler, A. Carson, and G. Rodin, "Depression in medical patients," *British Medical Journal*, vol. 325, no. 7356, pp. 149–152, 2002.
- [2] C.-S. Wu, C.-J. Kuo, C.-H. Su, S.-H. Wang, and H.-J. Dai, "Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records," *Journal of Affective Disorders*, vol. 260, pp. 617–623, 2020.
- [3] L. L. Craft and D. M. Landers, "The effect of exercise on clinical depression and depression resulting from mental illness: A meta-analysis," *Journal of Sport and Exercise Psychology*, vol. 20, no. 4, pp. 339–357, 1998.
- [4] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, 2017, pp. 69–76.
- [5] T. A. Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.
- [6] W. Wu, C. Zhang, and P. C. Woodland, "Self-supervised representations in speech-based depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
- [7] B. Yang, M. Cao, X. Zhu, S. Wang, C. Yang, R. Ni, and X. Liu, "MMPF: Multimodal purification fusion for automatic depression detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 6, pp. 7421–7434, 2024.
- [8] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang, S.-W. Kim, and M. Jhon, "Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features," *IEEE Access*, vol. 12, pp. 20 479–20 489, 2024.
- [9] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, and J. Epps, "When LLMs meet acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," in *Proc. Conf. Empirical Methods in Natural Language Processing*, Nov. 2024, pp. 146–158.
- [10] J. Xue, R. Qin, X. Zhou, H. Liu, M. Zhang, and Z. Zhang, "Fusing multilevel features from audio and contextual sentence embedding from text for interview-based depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 6790–6794.
- [11] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, and T. Gedeon, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8022–8026.
- [12] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Proc. Interspeech*, 2013, pp. 847–851.
- [13] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6247–6251.
- [14] J. Gratch, R. Arstein, G. M. Lucas, G. Stratou, and S. Scherer, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, 2014, pp. 3123–3128.
- [15] T.-Y. Hu, M. Armandpour, A. Shrivastava, J.-H. R. Chang, H. Koppula, and O. Tuzel, "Synt++: Utilizing imperfect synthetic data to improve speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4558–4562.
- [16] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7382–7386.
- [17] B. Hilmes, N. Rossenbach, and R. Schlüter, "On the effect of purely synthetic training data for different automatic speech recognition architectures," in *Proc. Workshop on Synthetic Data's Transformative Role in Foundational Speech Models*, Aug. 2024, pp. 46–50.
- [18] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2107–2116.
- [19] A. Antoniou, A. Storkey, and H. Edwards, "Augmenting image classifiers using data augmentation generative adversarial networks," in *Proc. Int. Conf. Artif. Neural Netw.*, ser. Lecture Notes in Computer Science, vol. 11141, 2018, pp. 594–603.
- [20] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 819–828.
- [21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [22] A. T. Beck, *Depression: Causes and Treatment*. Philadelphia: University of Pennsylvania Press, 1967.
- [23] D. D. Burns, *Feeling Good: The New Mood Therapy*. New York: William Morrow and Company, 1980.
- [24] S. Nolen-Hoeksema, "The role of rumination in depressive disorders and mixed anxiety/depressive symptoms," *Journal of Abnormal Psychology*, vol. 109, no. 3, pp. 504–511, 2000.
- [25] J. W. Pennebaker and C. K. Chung, *The Psychological Functions of Function Words*. New York: Psychology Press, 2007, pp. 343–359.
- [26] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.
- [27] G. E. Simon, M. V. Korff, M. Piccinelli, C. Fullerton, and J. Ormel, "An international study of the relation between somatic symptoms and depression," *New England Journal of Medicine*, vol. 341, no. 18, pp. 1329–1335, 1999.
- [28] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA: American Psychiatric Publishing, 2013.
- [29] H. I. Kaplan and B. J. Sadock, *Kaplan and Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry*, 8th ed. Baltimore, MD: Lippincott Williams & Wilkins, 1998.
- [30] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design—A tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 399–408, 2012.
- [31] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: A BiLSTM/1D CNN-based model," *Applied Sciences*, vol. 10, no. 23, Art. no. 8701, 2020.
- [32] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, and C. Clark, "Deep contextualized word representations," in *Proc. NAACL: Human Language Technologies*, 2018, pp. 2227–2237.
- [33] W. W. K. Zung, "A self-rating depression scale," *Archives of General Psychiatry*, vol. 12, no. 1, pp. 63–70, 1965.
- [34] C. Wang, Z. Cai, and Q. Xu, "Evaluation analysis of self-rating disorder scale in 1,340 people," *Chinese Journal of Nervous and Mental Diseases*, vol. 12, pp. 267–268, 2009.
- [35] L. Zhou, Z. Liu, Z. Shangquan, X. Yuan, Y. Li, and B. Hu, "TAMFN: Time-aware attention multimodal fusion network for depression detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 669–679, 2022.