International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 21, No. 3, 2025

Deep Learning Model Integrating Computational Linguistics: Research on Depression Recognition via Facial Expressions and Speech Signals

Hongzhi Pan[1,2], Yungang Wei[3,*]

[1]Zhongji Huasheng Engineering Consulting Co., Ltd., Xizang, China
[2]Beijing Changping District Information Technology Application Association, Changping, China
[3]Tibet Wangding Shensi Digital Technology Co., Ltd., Lasa, China

[*] **Corresponding author:** Yungang Wei

Abstract: To address the challenges of subjective diagnosis and delayed identification caused by imbalanced doctor-patient ratios in depression treatment, this study proposes a deep learning-based multimodal recognition system integrating facial expressions and speech signals. By analyzing characteristic audiovisual patterns in depression patients, we developed a dual-modal feature extraction and fusion framework: 3D Convolutional Neural Networks (3D-CNN) extract spatiotemporal facial expression features, while Causal Convolutional Networks (C-CNN) capture long-term temporal patterns in speech signals. The attention mechanism enables adaptive fusion across modalities. Experimental results on DAIC-WOZ and AVEC2013 datasets demonstrate 83.7% accuracy in classifying major depressive disorder, with an average regression error of 3.52 on the PHQ scale-outperforming single-modal and traditional fusion methods. This research provides a technical foundation for objective and automated depression screening.

Keywords: Computational Linguistics, Depression Recognition, Deep Learning, Facial Expressions, Speech Signals, Multimodal Fusion.

1. Introduction

1.1. Research Background and Significance

Depression, a prevalent mental disorder worldwide, has been identified by the World Health Organization (WHO) as the primary disease burden source by 2030[1]. Characterized by persistent low mood and diminished interest, it may lead to suicide in severe cases, with approximately 50% of suicide incidents directly linked to depression[2]. However, current diagnosis primarily relies on clinicians 'subjective assessments using tools like the Hamilton Depression Rating Scale (HAMD) and Patient Health Questionnaire (PHQ), which are significantly influenced by factors such as physicians' experience and patients' communication abilities. Meanwhile, the global imbalance in doctor-patient ratios results in many patients lacking timely diagnosis and intervention, with 60% of adult psychiatric patients never receiving professional treatment.

Physiological and psychological research has confirmed that individuals with depression exhibit significant abnormalities in auditory and visual behaviors. At the speech level, these manifest as monotonous intonation, slow speech rate, and frequent pauses. Facial expressions typically show reduced movement amplitude, downward gaze, and rigid facial expressions. These objective characteristics provide a basis for automated depression identification. Leveraging its powerful automatic feature extraction capabilities, deep learning technology demonstrates tremendous potential in affective computing. It can uncover hidden depression-related patterns from facial expressions and speech signals, laying the foundation for constructing objective and efficient depression screening systems[3].

1.2. State of Research at Home and Abroad

1.2.1. Research on Unipolar Depression Identification

In speech pattern research, early methods predominantly relied on handcrafted features such as fundamental frequency (F0) and Mel-frequency cepstral coefficients (MFCC), employing traditional machine learning models like support vector machines (SVM) for classification[4]. However, their limited feature representation capabilities resulted in relatively low recognition accuracy. In recent years, deep learning approaches have gradually dominated this field. Convolutional neural networks (CNNs) have been widely used to process speech spectrograms, while recurrent neural networks (RNNs) are employed to capture temporal dependencies. Notably, causal convolutional networks (C-CNNs) demonstrate superior performance compared to RNNs in modeling long-duration speech sequences.

In facial expression recognition, traditional methods rely on manually designed features like Local Binary Pattern (LBP) and Facial Action Units (FAU) to identify depressive states. However, these approaches suffer from time-consuming feature design and insufficient detection of latent patterns. With advancements in deep learning, 2D-CNNs have been employed to extract spatial facial features, while 3D-CNNs demonstrate remarkable progress in expression sequence analysis by simultaneously capturing spatial and temporal dynamics. Some studies have further enhanced the robustness of expression features by incorporating 3D facial scanning data[5].

1.2.2. Multimodal Fusion Research

Multimodal fusion enhances recognition performance by integrating complementary information across different signals. Current fusion methods are primarily categorized into early fusion and late fusion: Early fusion concatenates

audiovisual signals at the feature level, making it susceptible to modal heterogeneity; late fusion integrates outputs across modalities at the decision-making level, neglecting deep feature correlations. In recent years, attention mechanisms and graph neural networks (GNNs) have become focal points in fusion research. For instance, the Local-Global Multimodal Fusion Graph Neural Network (LGMF-GNN) achieves efficient fusion by constructing inter-modal correlation graphs, significantly improving diagnostic accuracy. However, existing methods still face challenges such as unreasonable modal weight allocation and insufficient modeling of long-term dependencies.

1.3. Research Content

This study focuses on depression recognition through dual-modal fusion of facial expressions and speech. The research primarily includes: (1) Constructing a preprocessing workflow for dual-modal datasets and designing data augmentation strategies tailored to the characteristics of facial expressions and speech signals; (2) Proposing a dual-modal feature extraction architecture based on 3D-CNN and C-CNN, which respectively capture spatiotemporal features of facial expressions and long-term speech sequence features; (3) Designing an attention-based fusion module to achieve adaptive weighting of modal information; (4) Conducting comparative experiments and ablation tests on public datasets to validate the effectiveness of the proposed method.

2. Theoretical Basis and Data Set

2.1. Deep Learning Basic Model

2.1.1. Convolutional Neural Network

CNN extracts features by alternating convolutional and pooling layers. The convolutional layers utilize local receptive fields and weight-sharing mechanisms to effectively capture spatial correlations. Building upon 2D-CNN, 3D-CNN adds temporal convolution kernels, enabling simultaneous extraction of spatial features and temporal dynamics from video sequences, making it suitable for modeling facial expression sequences[6].

2.1.2. Causal Convolutional Networks

C-CNN maintains temporal causality by limiting the receptive field of convolution kernels to capture only past temporal information, while employing dilated convolutions to expand the receptive field, enabling efficient processing of long sequence data. In clinical interview speech analysis, C-CNN effectively captures the unique pause patterns and speech rate variations characteristic of depression patients.

2.1.3. Attention Mechanism

The attention mechanism achieves adaptive focus on key information by calculating the importance weights of different features. In multimodal fusion, it dynamically allocates weights based on the contribution of each modality, alleviating the challenges of fusion caused by modality heterogeneity.

2.2. Dataset Introduction

2.2.1. DAIC-WOZ Dataset

This dataset contains 189 clinical interviews with 142 participants, totaling 50 hours of multimodal data including 3D facial scans, voice recordings, and transcribed transcripts. Each participant's PHQ-8 score quantifies depressive severity, with scores ≥10 indicating major depression. The dataset was collected through digital avatar interactions, effectively reducing the subjectivity inherent in manual interviews[7].

2.2.2. AVEC2013 Data Set

As a seminal dataset in affective computing, AVEC 2013 integrates multimodal data including audio, facial video, and physiological signals, along with clinical metrics such as the PHQ-9 scale. The video data enables facial action unit analysis, while the speech data contains rich rhythmic and spectral features, providing high-quality support for training dual-modal models. Access to the dataset requires official website registration and signing of a privacy protection agreement.

3. Deep Learning-based Bimodal Recognition Method

3.1. Overall Architecture Design

The proposed dual-modal depression recognition framework consists of four modules: data preprocessing, unimodal feature extraction, attention fusion, and classification regression. The system first preprocesses raw facial videos and speech signals, feeding them into a 3D-CNN facial expression extractor and a C-CNN speech extractor respectively[8]. The attention fusion module then performs weighted integration of multimodal features. Finally, a fully connected layer is employed to classify depression into two categories (severe/depressive vs. non-severe) and predict PHQ scores.

3.2. Data preprocessing

3.2.1. Facial Expression Data Preprocessing

Face detection and alignment: The MTCNN algorithm is employed to detect facial regions in video frames, extract 68 key facial points, and achieve face alignment through similarity transformation to eliminate the influence of pose changes[9].

Sequence sampling and standardization: The video is sampled at 30 frames per second into a fixed length sequence (16 frames), and the pixel values are normalized (mean 0, standard deviation 1) to enhance the generalization of the model.

Data augmentation: Use random horizontal flipping, brightness perturbation and other strategies to expand the data set and alleviate the problem of overfitting[10].

3.2.2. Voice Data Preprocessing

Signal preprocessing: The original speech undergoes pre-emphasis (coefficient 0.97), frame division (25ms frame length, 10ms frame shift), and windowing (Hanning window) to suppress spectral leakage.

Feature extraction: Compute the log-Mel spectral representation (40-dimensional Mel filter bank) and convert it into a 3D tensor (time step × frequency bin × channel) as input for the speech feature extractor.

Uniform duration: The linear interpolation method is used to unify the speech features of different durations into a fixed length (256 time steps) for batch training.

3.3. Unimodal Feature Extraction

3.3.1. Facial Expression Feature Extraction

The system employs an enhanced 3D-CNN architecture to extract spatiotemporal facial expression features, as detailed in Table 1. The input consists of 16-frame × 112×112×3 video sequences, with four 3D convolutional layers progressively capturing spatial features and temporal dynamics. Each

convolutional layer is followed by batch normalization and ReLU activation functions. Maximum pooling layers are applied to reduce feature dimensions, ultimately yielding a 512-dimensional facial expression feature vector through global average pooling.

Table 1. 3D-CNN Facial Feature Extractor Architecture

Layer type	Kernel size	step	Output dimensions	activation function
3D Convolutional Layer 1	3×3×3	1×1×1	16×112×112×64	ReLU
3D pooling layer 1	2×2×2	2×2×2	8×56×56×64	-
3D Convolution Layer 2	3×3×3	1×1×1	8×56×56×128	ReLU
3D pooling layer 2	2×2×2	2×2×2	4×28×28×128	-
3D Convolution Layer 3	3×3×3	1×1×1	4×28×28×256	ReLU
3D pooling layer 3	2×2×2	2×2×2	2×14×14×256	-
3D Convolution Layer 4	3×3×3	1×1×1	2×14×14×512	ReLU
Global average pooling	-	-	512	-

3.3.2. Speech Feature Extraction

The C-CNN architecture extracts long-term speech features using a network comprising four causal convolutional blocks. Each block consists of a dilation convolution, batch normalization, and an ELU activation function, with parameters specified in Table 2. The input is a 256×40×1 log-Mel spectrogram. The receptive field is expanded through causal convolution with dilation coefficients 2^n (where n denotes the block index), and the final 512-dimensional speech feature vector is obtained via global max pooling.

Table 2. Architecture of C-CNN speech feature extractor

Layer type	Kernel size	coefficient of expansion	Output dimensions	activation function
Causal Convolution Block 1	3×3	1	256×40×64	ELU
Causal Convolution Block 2	3×3	2	256×40×128	ELU
Causal Convolution Block 3	3×3	4	256×40×256	ELU
Causal Convolution Block 4	3×3	8	256×40×512	ELU
Global max pooling	-	-	512	-

3.4. Attention Fusion Module

The design of the multimodal attention fusion module enables adaptive learning of weights for facial expression and speech features. First, the dual-modal features are concatenated into a 1024-dimensional vector. Through a fully connected layer with a Sigmoid activation function, attention weights α (facial expression feature weight) and 1-α (speech feature weight) are calculated. The feature fusion formula is as follows:

$$Ffusion=\alpha*Fvisual+(1-\alpha)*Faudio$$

Here, Fvisual represents the facial expression feature vector, while Faudio denotes the speech feature vector. The attention mechanism dynamically adjusts the modal contribution based on sample characteristics. For instance, when facial expression features are more prominent, the α value approaches 1; conversely, it approaches 0.

3.5. Classification and Regression Head

The fusion features are fed into two heads: a classification head and a regression head. The classification head performs binary classification of major depression using two fully connected layers with Softmax activation, while the regression head outputs PHQ score predictions through three fully connected layers. During training, a combined loss function is employed, which combines the cross-entropy loss (for classification) and mean squared error loss (for regression) with weighted summation.

$$Loss=\lambda*LossCE+(1-\lambda)*LossMSE$$

In this case, λ is the equilibrium coefficient, which is set to 0.6 for optimal performance after experimental verification.

4. Experiment and Result Analysis

4.1. Experimental Environment and Parameter Settings

The experiment was conducted using the PyTorch 1.13

framework on an Intel Xeon Gold 6338 processor with NVIDIA A100 GPU (40GB VRAM). The dataset was split into training, validation, and test sets in a 7:2:1 ratio. We employed the Adam optimizer with an initial learning rate of 0.001, a 0.5 decay rate every 10 epochs, batch size of 16, and trained for 50 epochs.

4.2. Evaluation Indicators

For classification tasks, accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the ROC curve (AUROC) were used as evaluation metrics. For regression tasks, mean absolute error (MAE) and relative error (RE) were employed to assess the prediction accuracy of PHQ scores.

4.3. Comparison of Experimental Results

4.3.1. Comparison with Single-Mode Methods

Table 3 presents the comparison results between the DAIC-WOZ dataset and single-modal approaches. Our proposed dual-modal method outperforms single-modal methods across all metrics, with 4.2% higher accuracy (ACC) and 3.8% higher AUROC compared to the optimal single-modal approach (C-CNN speech recognition). This demonstrates that dual-modal fusion effectively leverages complementary information to enhance recognition performance.

Table 3. Comparison of Unimodal and Bimodal Methods (DAIC-WOZ Dataset)

method	modality	ACC(%)	TPR(%)	TNR(%)	AUROC(%)	MAE	RE(%)
3D-CNN	countenance	72.3	74.1	70.5	76.2	4.21	17.8
C-CNN	phonetic	79.5	80.2	78.8	82.5	3.89	16.3
This method	bimodal	83.7	84.5	82.9	86.3	3.52	14.8

4.3.2. Comparison with Existing Fusion Methods

Table 4 presents the comparison results with existing fusion methods on the AVEC2013 dataset. Our proposed method achieves 81.2% ACC and 84.7% AUROC, outperforming LGMF-GNN (without physiological signals) by 2.5% and 3.1% respectively. This demonstrates that the attention fusion module effectively captures modal correlations, thereby enhancing fusion efficiency.

Table 4. Comparison with Existing Fusion Methods (AVEC 2013 Dataset)

method	Integration Strategy	ACC(%)	AUROC(%)	MAE
Early feature splicing	Feature-level fusion	74.6	77.3	4.05
Late-stage decision-making	Decision-level integration	76.8	79.5	3.92
LGMF-GNN (Simplified)	Image Fusion	78.7	81.6	3.71
This method	Attention fusion	81.2	84.7	3.58

4.4. Fusion Test Results

To validate the effectiveness of each module, we conducted ablation experiments, with results shown in Table 5. Removing 3D-CNN's spatiotemporal modeling capability caused an 3.6% decrease in ACC; removing C-CNN's long sequence modeling resulted in a 4.1% drop; and eliminating attention fusion led to a 5.3% decline. These findings demonstrate that all three core modules are crucial for performance improvement, with attention fusion contributing the most significantly.

Table 5. Data from the DAIC-WOZ dataset

Experiment Configuration	ACC(%)	AUROC(%)	MAE
complete pattern	83.7	86.3	3.52
Remove 3D-CNN spatial modeling	80.1	82.7	3.78
Remove C-CNN long sequence modeling	79.6	82.2	3.85
Replace with simple concatenation	78.4	81.0	3.97

4.5. Model Explainability Analysis

The Gradient-CAM method visualizes attentional regions in facial expression features. The model primarily focuses on key areas like the eyes and mouth, consistent with clinical observations of depression patients exhibiting 'drooping eyes and mouth corners.' Voice feature analysis reveals the model shows highest attention to the 200-500Hz low-frequency band (corresponding to intonation variations) and 1-2 second intervals (corresponding to pause patterns), validating the rationality of feature extraction.

5. Conclusion

This study proposes a deep learning-based depression recognition method that integrates facial expression and speech modalities. The approach employs 3D-CNN to capture spatiotemporal facial features, C-CNN to extract long-term speech sequence features, and an attention mechanism to achieve modal adaptive fusion. Experimental results demonstrate superior performance on both DAIC-WOZ and AVEC2013 datasets, achieving 83.7% accuracy in classifying major depressive disorder with an average PHQ score regression error of 3.52, significantly outperforming existing methods. The research confirms the complementary value of facial expression and speech signals in depression identification, providing a novel solution for constructing objective and efficient screening systems. Future work

focusing on expanding modal types and optimizing model lightweighting could enable practical clinical applications.

References

[1] Davidson F K G. DEFINITION AND EPIDEMIOLOGY OF TREATMENT-RESISTANT DEPRESSION[J]. Psychiatric Clinics of North America, 1996. DOI:10.1016/S0193-953X(05)70283-5.

[2] Smith M, Dietrich B J, Bai E W, et al. Vocal pattern detection of depression among older adults[J]. International Journal of Mental Health Nursing, 2019. DOI:10.1111/inm.12678.

[3] Lacroix A, Calvet B, Laplace B, et al. Predictors of clinical response after rTMS treatment of patients suffering from drug-resistant depression[J]. Translational Psychiatry, 2021, 11: 1-7. DOI:10.1038/s41398-021-01555-9.

[4] Di Y, Wang J, Li W, et al. Using i-vectors from voice features to identify major depressive disorder[J]. Journal of Affective Disorders, 2021, 288(10). DOI:10.1016/j.jad.2021.04.004.

[5] Simon T L, Smith K J. The issues in autologous transfusion[J]. 2019. DOI:10.1016/0046-8177(89)90196-2.

[6] Sverdlov O, Curcic J, Hannesdottir K, et al. A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar Depression[J]. Frontiers in Psychiatry, 2021, 12: 640741. DOI:10.3389/fpsyt.2021.640741.

[7] Nuez N A, Joseph B, Pahwa M, et al. Augmentation strategies for treatment resistant major depression: A systematic review and network meta-analysis[J]. Journal of Affective Disorders, 2022, 302: 385-400. DOI:10.1016/j.jad.2021.12.134.

[8] Zhu T, Li K, Herrero P, et al. Deep Learning for Diabetes: A Systematic Review[J]. IEEE Journal of Biomedical and Health Informatics (J-BHI), 2021, 25(7). DOI:10.1109/JBHI.2020.3040225.

[9] Zhang L, Duvvuri R, Chandra K K L, et al. Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative[J]. Depression & Anxiety (1091-4269), 2020, 37(7). DOI:10.1002/da.23020.

[10] Hansen L, Zhang Y, Wolf D, et al. A Generalizable Speech Emotion Recognition Model Reveals Depression and Remission[J]. Acta Psychiatrica Scandinavica, 2021. DOI:10.1111/acps.13388.