# An Applied Study of Multi-modal Emotion Recognition to Assist in The Diagnosis of Depression

Siyu Wang *

College of Electronic Information and Optical Engineering, Nankai University, Tianjin, China

* Corresponding author: 2211590@mail.nankai.edu.cn

**Abstract.** Multi-modal emotion recognition significantly enhances the reliability and precision of detecting emotions by utilizing complementary information between different modalities and comprehensively analyzing physiological signals such as text, audio, and video. The rapid development of multi-modal emotion recognition technology has led to its wide application in the fields of medical diagnosis, educational feedback, and human-computer interaction, and has provided a new way for the objective diagnosis of depression. The aim of this paper is to review the latest progress of multi-modal affective computing for assisted depression diagnosis technology. The review will cover two aspects: application examples of multi-modal affect recognition in assisted depression diagnosis, future trends and research directions of multi-modal affect recognition. The paper aims to explore and analyze the application and efficacy of the three multi-modal affect recognition models proposed in the past three years for assisted depression diagnosis. The design principles and diagnostic effects of each model will be outlined. The current challenges and potential of their research will be explored. The performance of these models for improving diagnostic accuracy and assisting depression diagnosis will be evaluated and compared. The paper will discuss the current challenges of their research and their potential in practical applications., The paper will explore the contribution of multi-modal data fusion to improving diagnostic accuracy and look forward to future developments.

**Keywords:** Multi-modal emotion recognition, future trends, field of research.

## 1. Introduction

Depression, a serious affective disorder, greatly affects our human mental health and social functioning. Traditional diagnostic methods rely heavily on subjective assessment by clinicians and self-reporting by patients, which has certain limitations. Therefore, the study of accurate and efficient depression detection methods has become an urgent task in current research.

Affective computing is emerging as a bridge between humans and machines. The core of affective computing is the creation of computing systems that recognize, comprehend, and react to human emotions, which is critical to enabling more natural and humane human-computer interactions. Multi-modal fusion technology is especially critical in this field, which provides a more comprehensive and in-depth way of understanding emotions by comprehensively analyzing information from multiple modalities such as text, audio, and video.

Multi-modal emotion recognition captures the complexity and diversity of human emotions more accurately than uni-modal emotion recognition. Uni-modal sentiment recognition, such as sentiment analysis based only on speech or text, is often limited by the incompleteness of the information and the interference of the environment, which hinders the accurate representation of a person's genuine emotional state. In contrast, multi-modal emotion recognition leverages the strengths of various data sources to enhance the precision and robustness of emotion recognition. The development of this technology has not only driven research in the field of affective computing, but has also opened up new opportunities in a number of application areas, including social media analytics, educational feedback, and smart healthcare .For example, through APIs (application programming interfaces) provided by social platforms such as Twitter, Sina Microblog, Facebook, and YouTube, researchers have used strategic queries to access content posted during specific time periods. They searched for textual content containing specific phrases or keywords, or directly from discussion spaces surrounding specific mental health issues Obtain data that reflects an individual's mood, behavior,

and social patterns. It provides multidimensional information for mental health analysis [1]. In addition, the researchers proposed a multi-modal sentiment analysis method combining text and emoji data for assessing the learning status and emotional responses of students using online education [2]. In recent years, research on techniques for multi-modal affective computing for assisted medical diagnosis, especially for assisted depression diagnosis, has made great progress.

The purpose of this paper is to review the application examples, future trends and research directions of multi-modal emotion recognition in assisted depression diagnosis pairs. The paper explores the challenges of applying multi-modal emotion recognition in this field. It discusses the potential of multi-modal emotion recognition in assisted medical diagnosis Through in-depth analysis, this paper not only demonstrates the recent progress of multi-modal emotion recognition technology, but also provides an outlook on its future development, with a view to promoting further research and application in this field.

## 2. Examples of Applications

Emotion recognition is a technique for identifying emotional states by analyzing the physiological and behavioral responses triggered when emotions are expressed. Much progress has been made in recognizing human emotional states by analyzing single-modal data such as audio, text, visual, and other physiological signals. However, single-modal emotion recognition does not conform to the human perception pattern of emotion, and has limitations and instability, and is highly susceptible to the interference of other modal signals, so that the accuracy of emotion recognition decreases dramatically. In view of the complementary nature of different modalities, multi-modal fusion of emotion recognition research is receiving increasing attention, and the construction of a multi-modal emotion recognition system can effectively improve the accuracy and robustness of emotion recognition.

Traditional diagnostic methods for depression mainly rely on the clinical experience of psychiatrists and health questionnaires filled out by patients, such as the Personal Health Questionnaire (PHQ), the Beck Depression Inventory (BDI-II), and the Hamilton Depression Scale (HAMD)[3].These methods are characterized by a strong reliance on self-reporting of patients' conditions. It is difficult for them to detect specific signs of depression, and there is an absence of objective assessment methods. This can easily lead to over-diagnosis or misdiagnosis, especially under the imbalance of the doctor-patient ratio and the high work pressure of doctors, the efficiency of the test and the reliability of the results are difficult to ensure. Therefore, a detection system based on multi-modal emotion recognition makes the process more objective and interpretable and can solve the problems encountered with traditional methods of diagnosing mental disorders.

### 2.1. Model Based on Deep Learning Techniques

The model based on deep learning techniques uses a 3D convolutional neural network (3D-CNN) and VGG facial model to extract facial features, k-nearest neighbor algorithm is applied to linguistic analysis of dynamic textual descriptions to differentiate between different types of mental disorders. Dimensionality reduction and regression is performed by the Random Forest algorithm to predict depression scales. The model was trained for 2,000 cycles, which improved the accuracy of the depression prediction results. The model was trained using the AVEC 2016-17 dataset to develop depression scale prediction models, mood classification, and facial feature extraction. The dataset employs an audio signal processing technique that uses a deep learning algorithm to calculate the patient's anxiety BDI-II score, and the feature extraction process for the video modality is shown in Figure 1. Criteria for assessing model performance included sensitivity, specificity, precision, and accuracy, while RMSE and MAE were used to measure the accuracy of predictions.

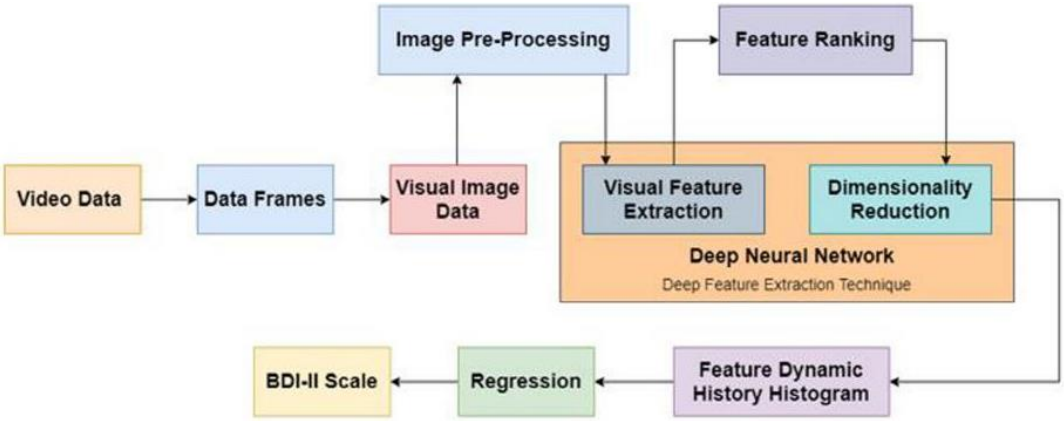


**Figure 1.** Block structure diagram of video data processing and feature extraction [4]

Experimental results show that the model improves facial detection and feature extraction by 2.7% over existing frameworks, proving its effectiveness in diagnosing depressive symptoms [4]. The model provides a new approach for early diagnosis and treatment of depression by combining deep learning and multi-modal data. The model combines facial images, user-generated data, and textual descriptions to help analyze and diagnose depression from different perspectives, and uses a convolutional neural network (CNN) model to automatically extract facial features, replacing traditional manual feature extraction methods and improving feature representation and diagnostic accuracy. However, because the dataset used in the experiment primarily reflects levels of depression in German ethnicity, the applicability of the model to other cultural and racial groups may be limited. Moreover, the accuracy of the model is dependent on how the BDI-II scale is asked and how patients respond, which may affect the generalizability of the model. Future research on the model will focus on expanding and diversifying the training dataset by including more diverse populations and cultural backgrounds. The research will validate the model's accuracy and reliability across different cultures and ethnic groups. The aim is to improve the model's generalization ability, interpretability, and cross-cultural applicability. Also, there is consideration to add a convolutional neural network layer to improve facial feature extraction.

### 2.2. The AVTF-TBN Model

Due to the small size scale of the public depression risk detection dataset and in order to improve the strength and accuracy of single-modality features and seeks effective ways to combine features from multiple modalities, the researcher designed the AVTF-TBN model and the emotion elicitation paradigm based on the reading and interviewing tasks to build the depression risk detection dataset. The model uses three separate branches to extract features from video, audio, and text. Then, combining the three modal features through an attention and residual mechanism in the MMF module. Finally predicts the labels of the samples through the fully connected layer. The detailed framework flow of the AVTF-TBN model is shown in Figure 2.

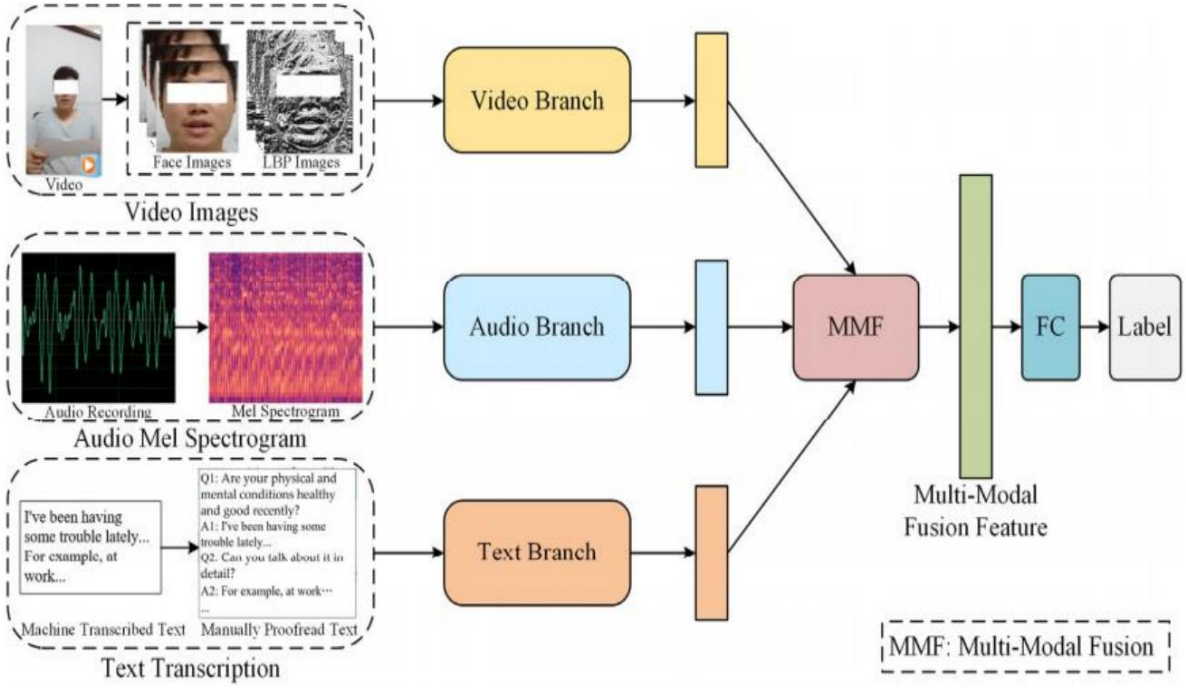


**Figure 2.** Detailed framework of the AVTF-TBN model [5]

The video branch is performed by extracting and fusing facial features from the original facial images and their associated LBP maps via the RST-DBN module, using Tanh and Softmax functions to measure the importance of each frame, and further extracting the facial features by using the BiLSTM model and the MHA module. The audio branch uses the NetVLAD module to obtain the Mel spectrogram vector, the GRU model and an MHA module to extract the features in depth to obtain the audio feature vector. The text branch uses the BERT Chinese pre-training model to extract text features from the original input text data, and an MHA module to deeply extract multi-level text features. The MMF module combines features from video, audio, and text. The MMF module combines features from video, audio, and text. Performance is evaluated using three key metrics: Precision, Recall, and the F1 Score, all derived from the confusion matrix. The experimental results show that the AVTF-TBN model performs best when data from both tasks are used for detection, confirming the effectiveness of the experimental paradigm and the AVTF-TBN model in detecting the risk of depression and demonstrating the important role of sensor-based data in mental health detection [5].

The AVTF-TBN model provides a new effective tool for depression risk detection through its innovative multimodal fusion approach and deep learning techniques. The model captures various behavioral and physiological features of depressed patients more comprehensively by combining data from different modalities of audio, video, and text, and enhances feature fusion by being able to identify and emphasize key information in different modalities through the MHA module. It is worth mentioning that the researchers designed an effective mood-evoking task, which helps to more accurately detect an individual's mood state, improves the capacity to extract useful features directly from raw data, and addresses the problem of the small size scale of the public depression risk detection dataset and the simplicity of the data collection paradigm. However, the model still has limitations; the multimodal fusion and deep learning features require high computational resources, which may limit its application in resource-constrained environments, and the model's ability to process data in real-time is unclear, which is an important factor for clinical applications. Nonetheless, its performance demonstrated in experiments suggests its potential for application in mental health. Future research will concentrate on enhancing the model's ability to generalize, real-time performance and interpretability, as well as ensuring data security and privacy.

### 2.3. The MFM-Att Model

In order to address the dependency that the model cannot fully exploit the textual context, the researchers proposed a multimodal fusion model with a multi-level attention mechanism (MFM-Att). The model not only extracts valuable intra-modality depression features, but also learns inter-modality correlations, thus enhancing its overall performance by diminishing the effects of redundant information. The model leverages data across three modalities, including audio, video and text, to fully extract depression features by utilizing intra-modality feature diversity and inter-modality information complementarity. Depression features are first learned from the audio, visual and text modalities, and then the output features from the three modalities are effectively fused using the attention fusion network (AttFN) and finally classified. The framework diagram of modal feature extraction for the MFM-Att model is shown in Figure 3.

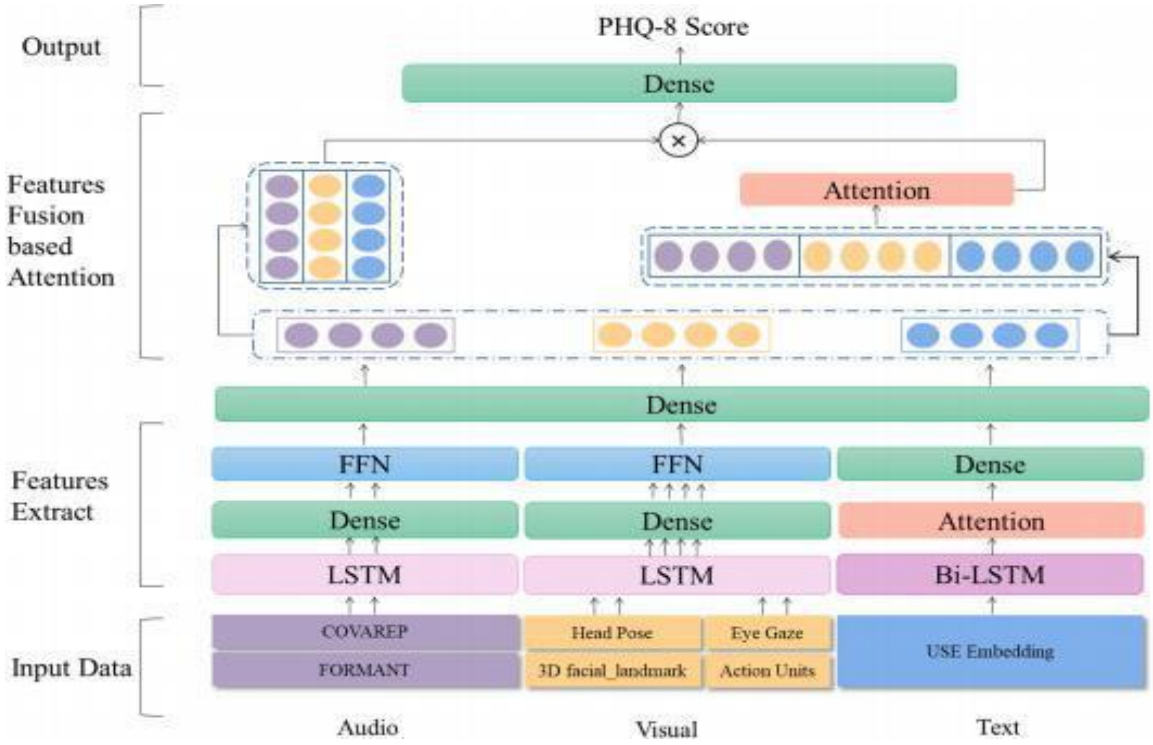


**Figure 3.** Overall framework of the MFM-Att model [3]

The audio modality sends the preprocessed COVAREP and FORMANT of speech signals to extract long-term audio features using LSTM. And then sends them to FFN for feature-level fusion. Finally, after several stacked dense layers and dropout layers, a value is output, which is mapped to the PHQ-8 scale to obtain the depression level. The visual modality includes four types of visual data, namely, 3D facial landmarks (3D FL), eye gaze (EG), head pose (HP) and facial action unit (AU), which are divided into two groups (3D FL, HP) and (AU, EG), and they are sent to FFN for feature integration. The text modality inputs the preprocessed text data embedding into Bi-LSTM to extract text features and learn depression features using the intra-modality attention mechanism to obtain more valuable text features. DAIC-WOZ dataset is used for supporting the identification of clinical depression, and adopts the MAE and RMSE as evaluation indicators. MAE is the average of the absolute differences between the predicted and actual values. RMSE is the root mean square differences between the predicted and actual values, serving as a measure of their dispersion. The MFM-Att model was evaluated on the DAIC-WOZ dataset and the results outperformed the state-of-the-art model in terms of RMSE [3].

The MFM-Att model provides an objective and accurate method for depression detection by combining three modalities of audio, video, and text data and a multilevel attention mechanism to capture depression features from different aspects. The model employs an intra-modality attention mechanism to identify key depression features and a cross-modal attention mechanism to assess the

significance of each modality and to reduce the redundancy of information due to feature diversity. The design of the model takes into account the complexity and diversity of depression features, improves the accuracy of detection by fusing information from different modalities, and shows strong robustness and accuracy compared with existing unimodal and multimodal approaches. However, the MFM-Att model still has limitations. For example, the model was trained and evaluated on a specific DAIC-WOZ dataset, which means that the model may not be able to achieve the same performance on new or different datasets. A variety of deep learning network architectures were used, including LSTM, Bi-LSTM, and FFN, which may increase the computational complexity of the model and affect its real-time performance and scalability. Also, the researcher mentioned the data privacy importance, but did not detail how to protect subjects' privacy during data collection and processing. Therefore, the MFM-Att model needs to be improved in future research in terms of model structure optimization, real-time performance improvement and data privacy protection. Future research will further optimize the model structure in order to improve the recognition accuracy and reduce the computational complexity of the network as much as possible. The aim is to boost the real-time performance of the network. What's more, data collected should be encrypted to protect subjects' privacy.

## 3. Future Trends and Research Directions

In current multi-modal emotion recognition research, researchers face a number of challenges involving several key aspects of the system. First, the heterogeneity between modalities leads to a diversity of data types and feature representations, which increases the complexity of data fusion across modalities. Second, the effectiveness of feature extraction directly affects the performance of the model, and dealing with high-dimensional data and achieving feature dimensionality reduction are key issues in this step. In addition, limited labeled multi-modal data limits the depth and breadth of model training and evaluation due to the time-consuming and costly creation of large-scale and diverse labeled datasets. Meanwhile, real-time processing and latency issues are critical for application scenarios that require fast responses, and models must reduce processing time while ensuring high accuracy. Finally, with the widespread use of AI in multi-modal emotion recognition research, responsible AI development becomes particularly important, which includes ensuring data privacy, fairness and transparency of models, and dealing with algorithmic bias and security issues [6].

Future research in multi-modal sentiment recognition will move towards improving cross-cultural adaptability and real-time analysis, while focusing on personalization and context-awareness to enhance model robustness and generalization. In addition, the research will be extended to domain-specific applications such as educational feedback and customer service, and to address the ethical and privacy issues associated with emotion recognition. Technological advances will include deep learning, transfer learning, and hardware optimization for more efficient model deployment, as well as building more diverse and high-quality multi-modal sentiment databases to support research and application development. Future research on multimodal emotion recognition computation is to expand the sample capacity and diversity to further train the model. It should explore the correlation between different modal features and reduce the model computation to further improve the detection performance. Meanwhile, it should consider the differences in emotional expression and develop models adaptable to multiple cultural features. These are starting points to enhance the accuracy and robustness of emotion recognition models. It is also important to pay attention to the complexity of the algorithms and to solve the problem of excessive consumption of computational resources.

## 4. Conclusion

Multi-modal emotion recognition, as a cutting-edge technology, has shown great potential for application in the field of assisted depression diagnosis. This paper reviews three adjunctive

depression diagnostic models that provide a more accurate and comprehensive assessment of affective states by integrating data from video, audio, and textual modalities. These models employ deep learning techniques, such as convolutional neural networks (CNNs), and advanced multi-modal fusion techniques to effectively capture and parse the nuances of emotional expression, thereby improving the accuracy of depressive mood recognition. The integrated application of these models not only improves the recognition of depressive symptoms, but also offers the possibility of early diagnosis and intervention through real-time monitoring and analysis. However, their development still faces challenges such as data scarcity, modal synchronization, real-time requirements, and cross-cultural differences. Therefore, to realize the application of these models in clinical practice, there is a need to improve the generalization ability of the models, ensure the privacy and security of the data, and enhance the interpretability of the models.

With the advancement of technologies such as deep learning, future research will focus more on the scalable design of affective features, the application of transfer learning, and the challenges of cross-cultural affect recognition in order to enhance the generalization capability of multi-modal affective computing and to enable integration with existing healthcare systems. In addition, researchers should continue to explore new data fusion methods and machine learning algorithms to further improve the performance and reliability of the model. The goal is for multimodal emotion recognition technology to play a more critical role in the early identification and treatment of depression. It should provide technical support for assisting in the diagnosis of depression, and contribute to the improvement of the quality of life of patients, and bring more well-being and innovation to the human society.

## References

[1] Khoo, L. S., Lim, M. K., Chong, C. Y., McNaney, R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors, 2024, 24 (2): 348.

[2] Qin, X., Zhou, Y., Li, J. Multi-Modal Emotion Recognition for Online Education Using Emoji Prompts. Applied Sciences, 2024, 14 (13): 5146.

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

[4] Meshram, P., & Rambola, R. K. Diagnosis of depression level using multimodal approaches with deep learning techniques and selective features. Expert Systems, 2023, 40 (4): 2933.

[5] Zhang, Z., Zhang, S., Ni, D., Wei, Z., Yang, K., Jin, S., Huang, G., Liang, Z., Zhang, L., Li, L., Ding, H., Zhang, Z., Wang, J. Multimodal sensing for depression risk detection: Integrating audio, video, and text data. Sensors, 2024, 24 (12): 3714.

[6] Geetha A.V., Mala T., Priyanka D., Uma E. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. *Information Fusion, 2024, 105, 102218.