International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 633-649
https://doi.org/10.70917/ijcisim-2026-5629
Article

An Audio-Visual Emotion Recognition Framework for Objective Mental Health Assessment

Rupali Landge[1], Hariram Chavan[2]

[1]Research Scholar, Department of Computer Engineering, K.J. Somaiya Institute of Technology, Sion, Mumbai, India
[1]Assistant Professor, Department of Information Technology, Shree L R Tiwari College of Engineering, Miraroad, Thane, India
[2]Professor, Department of Computer Engineering K.J. Somaiya Institute of Technology, Sion, Mumbai, India
Corresponding author: Rupali Landge
Email: rupali.landge@somaiya.edu

Abstract: Mental health disorders are increasing rapidly due to changes in lifestyle patterns, social pressure and work-related stress. Traditional mental health assessment primarily depends on self-reported questionnaires, which are subjective and does not accurately represent individual psychological condition. To overcome these limitations, this paper proposed an objective behavioural assessment using audio and video emotion analysis for mental health assessment. This proposed framework integrates a deep convolution neural network based VideoEmoNet model and a convolutional neural-network with long short-term memory based AudioEmoNet model. The publically available dataset like FER2013 for the VideoEmoNet model and RAVDESS, TESS, SAVEE, CREMA-D for the AudioEmoNet model are used for experiments. The output of the audio-visual models is integrated using late fusion technique to provide final mental health assessment state. The AudioEmoNet model achieves 97.47% accuracy and VideoEmoNet model achieves an accuracy of 95.01%. The proposed audio-visual framework achieves 98.20% accuracy on benchmark dataset. The proposed framework improves behavioural emotion recognition performance and shows how complementary audio-visual emotion information can be integrated for behavioural evaluation which is important for mental health assessment.

Keywords: Audio-visual emotion recognition, mental health assessment, deep learning, facial emotion recognition, audio emotion recognition

1. Introduction

Mental health disorders have become a global public health concern which affects millions of individuals across all age groups and social-economic backgrounds [1, 3]. Emotional health, productivity and quality of life are negatively impacted by the mental health conditions like depression, anxiety and stress [1, 5]. According to recent mental health report from World Health Organisation, the prevalence of mental health issue has increased considerably due to factors such as social pressure, occupation stress and lifestyle changes. For improving the mental health conditions and reducing the burden on the healthcare system, early identification and interventions are very important [5, 15].

The traditional mental health assessment method primarily relies on self-reported questionnaires, clinical interviews and psychological evaluation scales [15, 21]. These approaches are widely used for clinical practices which are having drawbacks such as subjectivity, response bias, recall problem and dependence on individual willingness to share personal information When people are unable to express their emotional state accurately, the accuracy and reliability of mental health of assessment may be affected [15]. These drawback emphasise the need of more objective and data-driven assessment mechanism which helps mental health professional for analysis and monitoring process.

An individual's emotional and psychological state can be objectively determined by human behavioural signals. Among various behavioural modalities, audio and facial expressions are considered highly informative because they naturally reflect emotional changes associated with mental health conditions. The Facial expression provides visual cues that indicate emotion likes happy, sad, anger, fear and stress [17-20], while variation in audio characteristics such as tone, pitch and energy can reveal the emotional state. The automatic extraction and analysis of these behavioural indicators were made possible by development in computer vision and audio processing.

The rapid advancement in deep learning techniques has significantly transformed the field of emotion recognition. By automatically learning hierarchical feature representation from raw data, deep neural networks have proven better than tradition machine learning techniques. Convolutional neural networks archive remarkable performance in the audio emotion recognition process by extracting important features from audio signals.

Even if there is a progress in facial and audio emotion recognition several challenges are unresolved. Many researchers are focus on unimodal technique either audio or facial expression. These methods are providing good results for emotion classification but fail to give result for comprehensive assessment which is suitable for mental health analysis [18]. As a result, there is growing need for robust deep learning framework that can objectively analyse behavioural indicators lined to psychological well-being [20,25].

Motivated by these challenges, this study proposes a deep learning-based audio-visual behavioural assessment framework for objective mental health analysis. The proposed approach combines multimodal inputs, video-based facial emotion feature and audio-based emotion characteristics to enable more comprehensive and objective evaluation. The integration of these modalities enhances the accuracy and consistency of analysing the mental health state as compared to traditional methods.

The primary contribution of this research is summarized as follows:

- Design of video emotion recognition model VideoEmoNet using deep convolutional neural-network for facial emotion recognition using FER2013 dataset
- Design of a convolutional neural network-based audio emotion recognition model AudioEmoNet using MFCC, RMSE and ZCR features for audio emotion recognition.
- Negative emotion aggregation is process for both audio and video modalities, where sadness, anger, and fear values are combined to generate behavioural distress score before audio-visual model.
- A late fusion strategy is investigated for combining complementary audio and video emotional features.
- The proposed model was independently evaluated using multimodal benchmark dataset which is not used for training the model, thereby providing an independent evaluation of the model's generalization capability

The literature survey based on audio and visual emotion recognition models and audio-visual models are covered in section 2. Section 3 presents the proposed system architecture. The methodology and materials are discussed in section 4. The results are discussed in section 5. Section 6 presents conclusion and future scope

2. Literature survey

Mental health disorders have become a major public health concern, so researchers are motivated to develop automated and objective assessment systems using artificial intelligence and deep learning techniques. Form recent studies it has been identified that facial expressions and audio signals are effective behavioural indicators for recognition of human emotions and support for mental health assessment. Prediction performance can be increase by a fusion framework that integrates complementary information from several modalities. The major contributions of recent studies are discussed in this section.

2.1 Literature survey of audio emotion recognition models

Log-spectrograms, Mel-spectrograms, and Mel Frequency Cepstral Coefficients (MFCCs), are typically used as inputs for a variety of deep learning and machine learning models. A.S. Alluhaidan gives enhance feature extraction technique which gives hybrid features. MFCCT is used as input to convolution neural network model to enhance the performance [4]. STFT pre-processing. VGGish and YAMNet pertained models are used by Tae-Wan Kim [5]. J. Singh proposes a self-attention based deep learning model by combining a long short term memory and a two-dimensional convolutional neural network [7]. S. Mishra uses an ensemble deep convolution neural network technique with Bidirectional LSTM-based framework for audio emotion classification and recognition [8]. For data augmentation C. Barhoumi uses a deep learning method along with noise addition and spectrogram shifting techniques. The model accuracy is increase by using convolution neural network techniques with, multilayer perceptron and bidirectional long-short-term memory [9]. Rolinson Begazo propose an architecture with one dimensional and two dimensional convolutions neural network that combines pictures of spectrograms and spectral characteristics [20]. R. Ullah uses Transformer encoder in conjunction with convolutional neural networks [11]. A deep convolution neural networks such as convolution recurrent neural network and gated recurrent unit used by L.T. Van [12].

Table 1: Summary of audio emotion recognition models

Reference	Author	Model	Input Features	Database	Accuracy(%)	Limitations
[4]	Ala et al.	1D CNN	MFCCT(MFCCs with Time Domain feature)	EMO-DB SAVEE RAVDESS	97 93 92	Accuracy can be increased by using RNNs and acoustic features
[7]	J Singh et al.	CNN+LSTM+Attention	MFCCs	RAVDESS,TESS	90.19	Multi-modal setup using audio features
[8]	S Mishra et al.	Hybrid deep CNN + BiLSTM	MFSC	TESS, SAVEE	96.36	Validation can be performed on more datasets
[9]	C Barhoumi et al.	CNN+ BiLSTM	MFCC+ Mel spectrogram + ZCR + Chroma + RMS	RAVDESS, TESS, EMO-DB	90.12	To enhance the performance the system can be extended to other languages using transfer learning
[11]	R Ullah et al.	CTENet(Two parallel CNN)	MFCC Spectrum	RAVDESS, IEMOCAP	82.31 79.42	Accuracy can be increase by refining the model architecture

Table 1 provide summary of various audio emotion recognition approach using deep learning architecture and acoustic features. Highest accuracy is achieved by, hybrid architecture and multi feature extraction methods. Performance of the model varies as per the samples in datasets. This literature summary highlight the need for robust emotion recognition models.

2.2 Literature survey of video emotion recognition models

Facial feature is used as an input to various Convolution Neural Network and Deep Convolution Neural Network based machine learning models. Dhvanil Bhagata proposed method which immediately detect emotion from real time video stream. This method overcomes the practical constraints of current facial emotion recognition techniques and closes the gap between lab research and practical implementations [13]. E. G Dada proposed a

convolution neural network-10 model with transfer learning methods for facial expression recognition. This model provides a strong tool for accurately identifying and understanding facial expressions [15]. V. Rathod used a method, in which patient's emotional state is taken into account when diagnosing 11 distinct mental health problems [16]. Transfer learning (fine-tuning) with pre-trained Inception- and MobileNet-V2 models used by E. SatrioAgung for creating convolutional neural network models [17]. O.C. Oguine uses two deep learning architectures: Deep convolution neural networks and Harr cascade. Deep Convolution Neural Network used in this study has numerous kernels, additional convolutional layers, and ReLU activation functions which is used for improving the filtering path and facial feature extraction. [18]. J. AINA uses fusion of convolution neural networks and Visual Transformer (ViT) are used by J. AINA for prediction of mental health state [19]. D.Turcian provide a technique for facial expression recognition using a small amount of training data that can be accessed remotely via a web or mobile application [20]. Haibo Xu proposed a mental health analysis method which combines traditional scales with multimodal intelligent recognition technology [22]. Summary of video emotion recognition models study are given in below Table 1

Table 2: Summary of video emotion recognition models

Reference	**Author**	**Model**	**Dataset**	**Training Accuracy(%)**	**Limitations**
[13]	Bhagat et al	Deep CNN using pertained model + Harr Face classifier	FER2013	82.56	The model precision can be extended
[15]	Rathod et al	2D-CNN	FER2013	76.17	Remote Mental Health Analysis not available
[18]	Oguine et al	DCNN+Harr Cascade	FER2013	70.04	Due to insuffient dataset Difficulty in Prediction of two classes
[20]	Turcian et al	CNN	FER2013	70.73	Inadequate dataset make it difficult to develop generalizable algorithm
[22]	Xu et al	DCNN	FER2013	82.3	Accuracy can be improved and Insufficient data sample

Table 2 summarizes the literature summary for video emotion recognition methods and their performance on FER2013 dataset. Various convolution neural networks based architecture show different level of training and validation accuracy. The result indicates that pertained model with deep convolution neural network architecture provides high accuracy, while several approaches suffer from limited generalization. This analysis highlights research gap for improving model performance.

2.3 Literature survey on audio-visual mental health assessment models

L. ZHang et al proposed DepITCM, a multimodal depression detection system that integrates facial and audio information using multi-task learning [23]. Yuchun et al designed a multimodal model based on the audio-visual emotional features using convolutional neural network architecture [25]. Tiwary et al. proposed a multimodal

depression identification model using a convolutional neural network-based audio and video model. The parallel branches of CNN process both audio and video input individually [26]. Zhang et al. proposed a multimodal depression detection model which integrates audio and video signals using long short-term memory network. Audio and video features are extracted separately and the output of both modalities are fused to get the final result [29]. Multimodal depression detection framework proposed by khan et al. which integrates Dynamic Gating Network(DGN), Multi Head Attention Network(MHAN), and Explainable artificial intelligence(XAI). The adaptive fusion mechanism enhanced generalisation and interpretability across several benchmark datasets by dynamically adjusting each contribution of each modality [30]. The model captured emotional features from the audio-visual model and adjusted the contribution of both modalities to optimise the performance of the mental health prediction. Basha et al. developed a real-time multimodal emotion and stress recognition system by integrating facial expression analysis with the audio emotion recognition. This framework uses facial landmark detection along with MFCC-based audio feature extraction and hybrid neural network for classification of emotions. The system gives better results for stress monitoring rather than overall mental health risk assessment. To improve detection accuracy, this method simultaneously learns global and local features [32].

Table 3. Comparative analysis of audio-visual mental health assessment models

Reference	**Author**	**Method**	**modality**	**Dataset**	**Mental Disorder**	**Accuracy**	**Advantages**	**Limitations**
[23]	L.Zhang et al.	LSTM based multimodal	Video + Audio	DAIC-WOZ, E-DAIC	Depression	83.86%	Multimodal model	Noisy data, Limited dataset cause overfitting
[25]	Yichun et al	Spatio-temporal attention mechanism+CNN	Facial + Audio	AVEC2014	Mental health analysis	96.73%	Dynamic fusion strategy	Prediction for only two disorder
[26]	Tiwary et al.	CNN	Facial + Audio	DAIC-WOZ	Mental health monitoring	77%	Parallel branches of CNN	Identify only depression disorder
[29]	Zhang et	Multi-task deep learning	Facial + Audio	AVEC 2017	Depression	82.3%	Learns global and local features	Limited expandability
[30]	Khan et al.	Dynamic Gating+ Multi-head attention + EAI	Facial + Audio	DAIC-WOZ	Depression	93%	Adaptive fusion,Explainabilty	High computation cost
[32]	Basha et al	Hybrid neural network	Facial + Audio	Real time multimodal dataset	Stress	95.2%	Real time emotion and stress monitoring	Not address long-term progression

The Table 3 presents comparative analysis of various existing audio-visual mental health analysis models, which indicates that significant progress has been made in deep learning and multimodal learning for mental health assessment. The studies have focused on facial and audio emotion recognition with integration of audio and video information. The audio-visual frameworks give better results by integrating complementary information from different behavioural modalities.

However, there are certain limitations in the study. Most of the researchers designed a model for depression prediction and identification, which restricted to more general mental health risk assessment. Most researcher work is on computationally demanding fusion techniques which are difficult to implement. The use of complex a multimodal model increases processing cost and decreases interpretability of the model.

From the analysis of the literature survey, it has been identified that there is a need of simple, robust and efficient multimodal framework which is capable of the identification of several mental health risks assessment rather than focusing on a single mental health disorder. To overcome these limitations, the present research proposes an audio-visual framework that integrates audio and video emotion recognition results. The proposed framework provides objective and reliable method for early mental assessment by using complementary information extracted from the facial expressions and audio signals.

3. Proposed audio-visual late fusion architecture

The Figure 1 illustrates the proposed audio-visual fusion framework for mental health assessment.

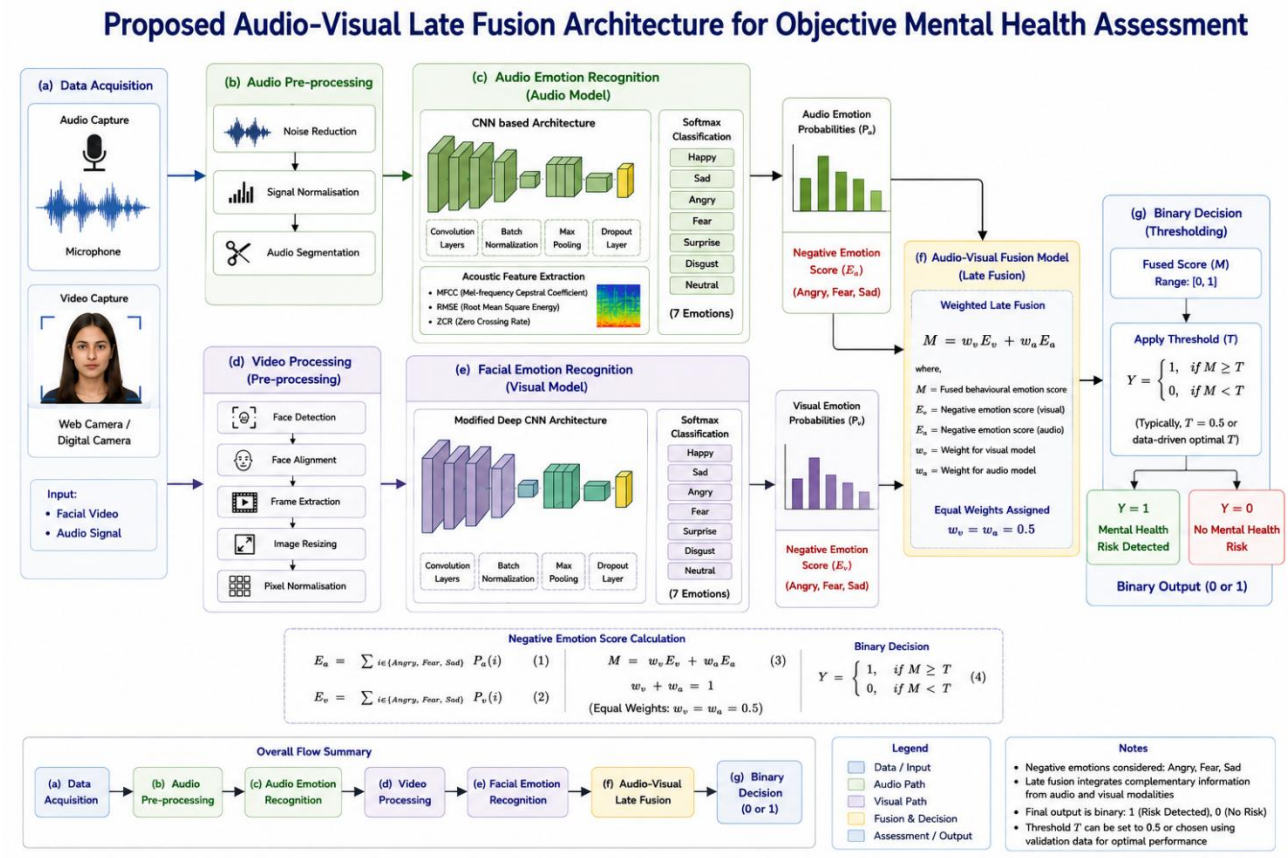


Figure 1. Proposed audio-visual framework for objective mental health assessment

The proposed architecture performs objective mental health assessments by integrating audio and visual emotion recognition score using late fusion technique. The individual mental health risk is analysed using behavioural

cues extracted from facial expression and audio signals. The complete architecture consists of six stages which are describe below.

a) **Data Acquisition**: Facial videos of the user are captured using a web camera or a digital camera and audio signals are recorded using microphone. These behavioural modality provides complementary information, where facial expression represents visual emotional cues and audio signals reflects vocal characteristics such as tone. Pitch and intensity.

b) **Audio pre-processing**.: The recorded audio is pre-processed before the emotion classification. Noise reduction, signal normalisation, audio segmentations are the steps of audio pre-processing.

c) **Audio emotion recognition**: The convolution neural network based model is designed to extract acoustic features like Mel-frequency cepstral coefficient, Root Mean Square Energy, Zero Crossing Rate for classification of emotion using convolution neural networks.

d) **Video processing**: Video prepressing is going through various stages before emotion recognition to improve the quality of feature extraction. Face detection, face alignment, frame extraction, image resizing and pixel normalisation are the steps for video pre-processing. These steps reduce background noise, illumination variation and unnecessary information before emotion recognition.

e) **Facial emotion recognition**: After pre-processing, facial images are input to the modified deep neural network architecture. The network extracted facial features through various convolution layers followed by batch normalisation, the max pooling and a dropout layer. The facial features are classified into seven basic emotions using softmax activation function.

f) **Audio-visual fusion Model**: The predicted output of the video and audio modality are fused to determine mental health analysis of individuals. Both audio and video modalities are assigned equal importance because each provides complementary behavioural information. The fused emotion representation captured complementary information from both audio and video modalities, which gives a more robust and reliable assessment than the unimodal models.

The proposed framework used negative emotions state like anger,fear and sadness because this emotion are used as a behavioural indicators for the proposed framework. The negative emotion score is calculated for each modality.

$$Ev = sad + fear + angry \quad (1)$$

$$Ea = sad + fear + angry \quad (2)$$

$$M = Wv.Ev + Wa.Ea \quad (3)$$

$$Wv = 0.5, Wa = 0.5 \;\; and \; Wv + Wa = 1$$

Where

M=Fused behavioural emotion score

Ev=Negative emotion score for visual model

Ea=negative emotion score for audio model

Wv=Weight assigned to visual model

Wa=Weight assigned to audio model

Equal weights were assigned to audio and visual modalities because both modalities give complementary behavioural information.

g) **Mental health assessment**: The mental health risk score is calculated using fusion output M. If M=1 then mental health risk detected else no risk of mental health. The fused behavioural score M is converted into a binary mental-health score decision using a threshold T. If M> T then output is assigned to class 1, indicating the presence of mental-health risk; otherwise, the output is assigned to class 0, indicating no detected risk. The threshold T=0.5 is determined using the validation set to obtain the optimal classification performance.

$$M = \{yes \quad M = 1\ No \quad M = 0 \tag{4}$$

This behavioural method can be used as a decision support tool for early assessment of mental health state.

4. Material and Methodology

4.1 Dataset used for audio and visual model

This section discusses the dataset used by video and audio models. The FER2013 dataset is used for training and testing by the video emotion recognition model. The audio emotion recognition model used RAVDESS, TESS, SAVEE and CREMA-D datasets.

4.1.1 Audio emotion datasets: Multiple benchmark datasets are combined to train the Audio emotion recognition model, which helps to improve data diversity and robustness of the model. The audio datasets consist of recordings with various emotion states.

The Ryerson Audio-visual Database of Emotional Speech and Song (RAVDESS) dataset consists of recordings from 12 men and 12 females. These recordings are of varying intensity levels. The Toronto emotional speech set (TESS) dataset consists of 2800 recordings of seven different emotion states which are in wav format. SAVEE (Surrey Audio-visual Expressed Emotion) dataset consist of recordings from male speakers which are partition into training, testing and validation for systemic evaluation. A crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) consists of emotion audio recording of 91 actors**.** This dataset provides high quality of recordings with consistent sampling rate and supports reliable feature extraction and classification [34]**.**

The model generalization can be improved by combining different audio datasets**.** All audio recordings are resampled to a common sampling frequency, normalised and process using feature extraction technique. To improve robustness across different recording conditions data augmentation technique like noise addition, pitch shifting and time stretching were applied.

4.1.1 Facial emotion dataset: FER2013 is a large-scale dataset used for facial expression recognition. This dataset consists of 48X 48 pixel images of seven basic emotions (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). It is widely used for benchmarking FER algorithm. training deep neural network and studying cross dataset generalisation [35]. The training set consists of 28,709 examples and the public test set consists of 3,589 examples.

The FER2013 is used as benchmark dataset for facial expression analysis for independently training the VideoEmoNet model. This dataset is used to learn a strong mapping between facial appearance patterns and basic emotional state.

4.2 Dataset used for audio-visual fusion model

The Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset is a multimodal dataset used by audio-visual model. This dataset includes audio, video and transcript data used to support the diagnosis of physiological conditions like depression and anxiety [36]. It is not publically available due to privacy restriction, access can have requested through the officially dataset portal. The dataset consists of 189 interaction sessions with predefined division of training and testing samples.

4.3 Methodology for AudioEmoNet: Proposed audio emotion recognition model:

The Figure 2 illustrate the proposed AudioEmoNet architecture for audio emotion recognition. The model combines the convolution neural network for acoustics feature extraction with long short-term memory network for learning temporal dependencies in audio signals. Initially, input audio signal is pre-processed. The handcrafted acoustic features such as Mel-frequency cepstral coefficient, root mean square energy, zero crossing rate are extracted to represent temporal and spectral characteristics of audio.

The CNN architecture consist of three one-dimensional convolutional layers with 1024,512 and 256 filter respectively. Each CNN layer used the ReLU activation function and a kernel of size 5 to learns the acoustic patterns from extracted features. Max pooling reduces feature dimension, batch normalisation after each convolutional block accelerate convergence and the drop out layers are used to minimize overfitting.

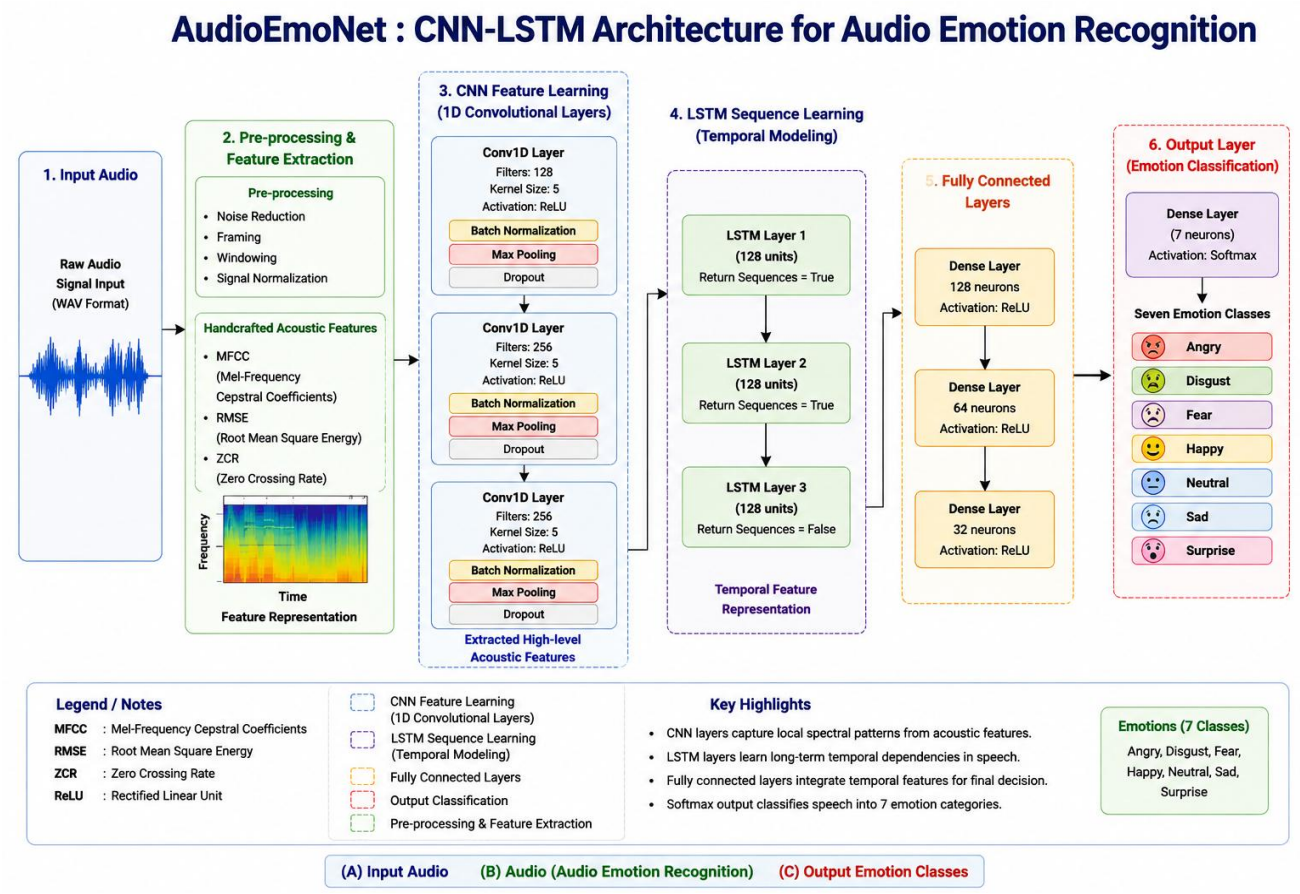
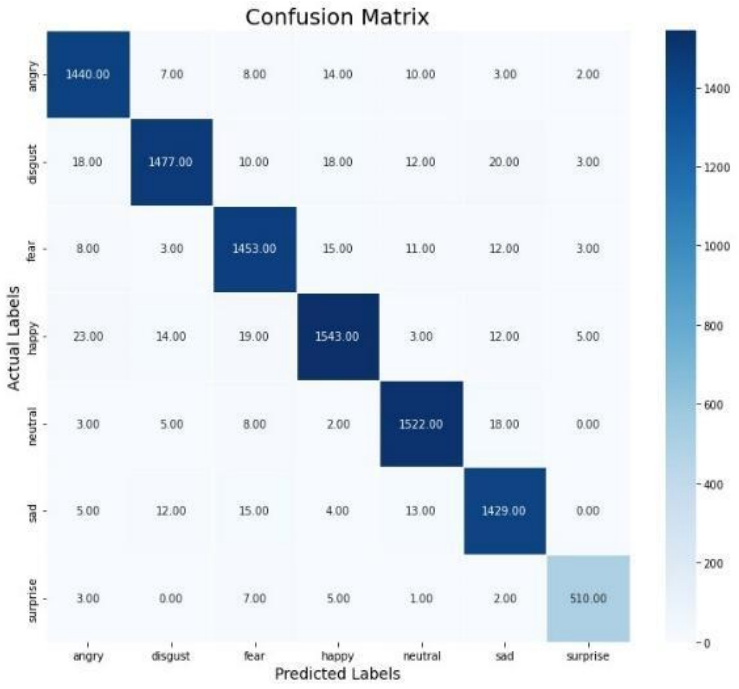


Figure 2: Methodology for AudioEmoNet model

The features extracted from the CNN are passed to the three stacked layers of LSTM each consisting of 128 hidden units. The first two LSTM layers return a complete sequence to preserve temporal information whereas, the final LSTM layer generate compact sequence for emotion classification. The extracted temporal features are processed using fully connected layer containing 128,64 and 32 neurons with the ReLU activation function. The Softmax function classifies the emotions into seven categories namely Angry, Disgust, fear, Happy, Neutral, sad and surprise. The proposed CNN-LSTM model improves the audio emotion recognition process by effectively capturing local spectral features and long- term temporal dependencies of audio.



	precision	recall	f1-score	support
angry	0.96	0.97	0.97	1484
disgust	0.97	0.95	0.96	1558
fear	0.96	0.97	0.96	1505
happy	0.96	0.95	0.96	1619
neutral	0.97	0.98	0.97	1558
sad	0.96	0.97	0.96	1478
surprise	0.98	0.97	0.97	528
accuracy			0.96	9730
macro avg	0.96	0.96	0.96	9730
weighted avg	0.96	0.96	0.96	9730

Figure 3. Confusion matrix for AudioEmoNet model

The Figure 3 presents the confusion matrix of The AudioEmoNet model. The Model achieves an accuracy of 97.47%, which indicates consistence performance of all emotion classes. The diagonal elements of the confusion matrix represent correctly classified samples.

4.4 Methodology for VideoEmoNet: Proposed video emotion recognition model

The Figure 4 illustrate the proposed VideoEmoNet model used for the video emotion recognition process. The model uses a deep convolutional neural network for classifying emotions in seven different categories. The model is trained using FER2013 dataset, which consists of grey scale images of seven 48 X 48 pixels with seven different emotion category.

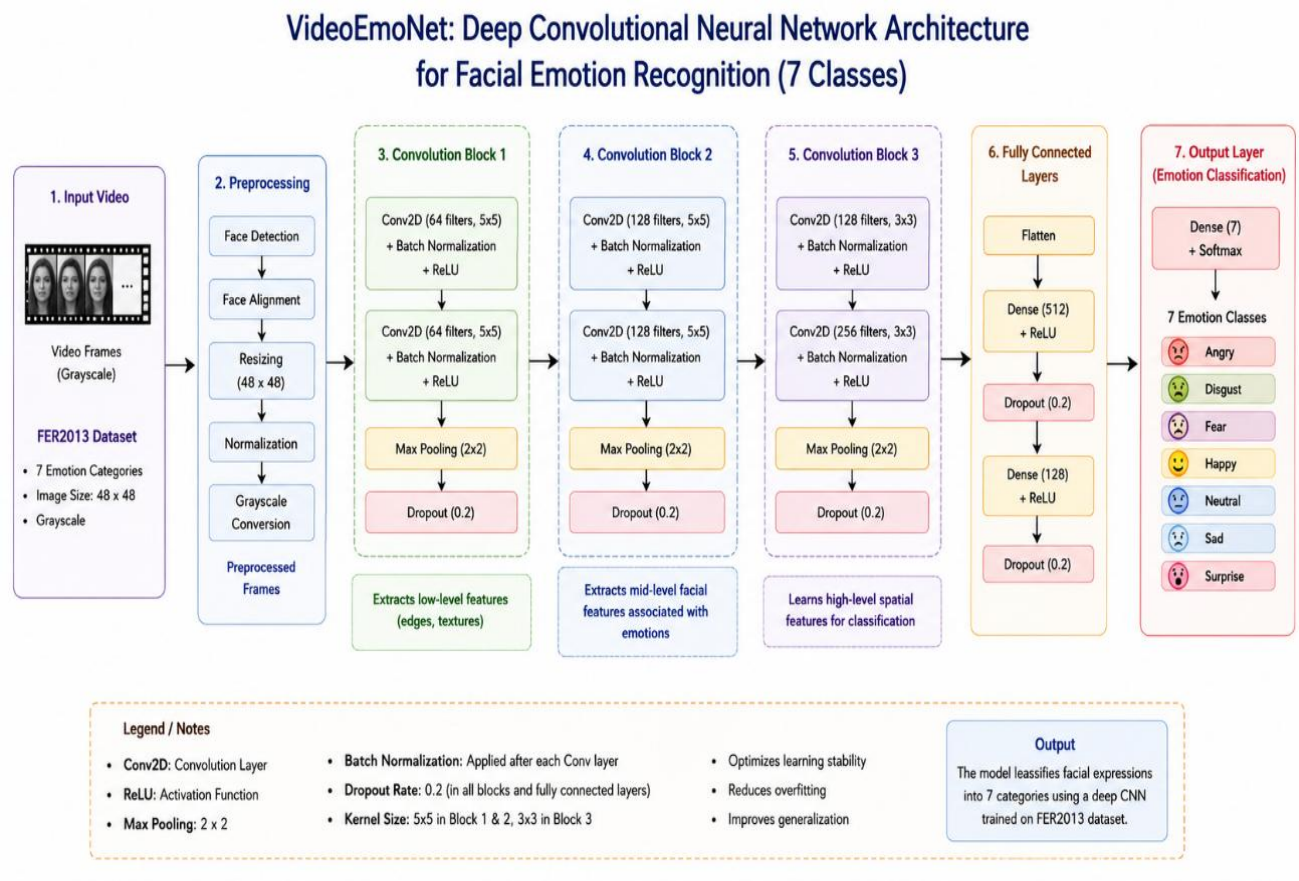
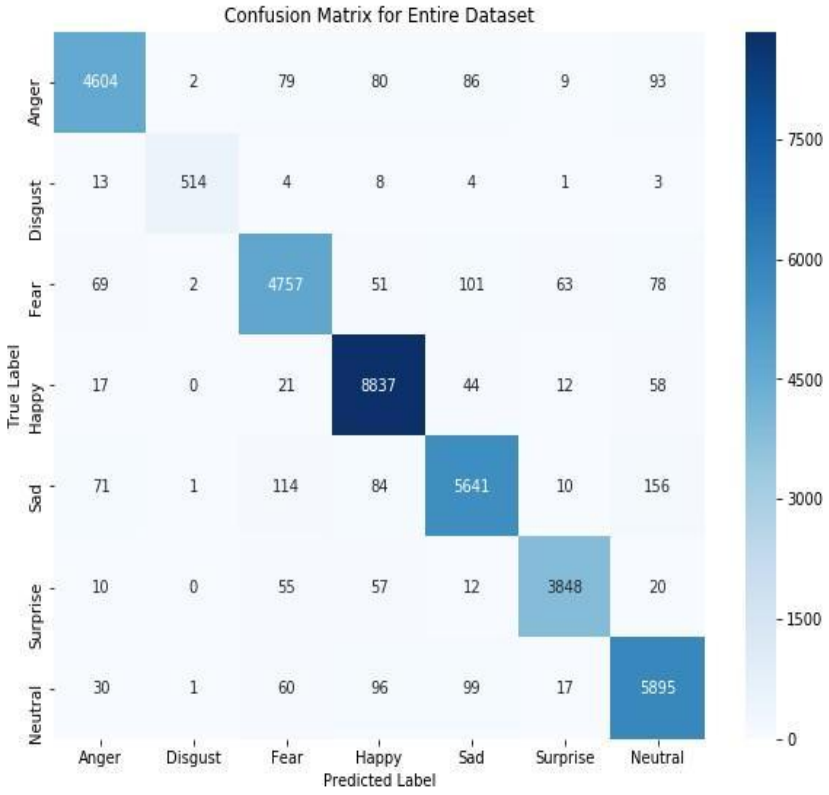


Figure 4. Methodology for VideoEmoNet model

The preprocessed images are given as input to the proposed DCNN architecture for feature extraction and emotion classification. The proposed architecture consists of three convolution blocks. The first block consists of two convolution layers with 64 filters of size 5 X 5, followed by 2 X 2 max pooling layer. The second layer consists of two convolutional layers with 128 filters of size 5 X 5, followed by batch normalisation and max pooling. This layer helps to extract facial features associated with different emotions. The third layer consists of two convolutional layers with 256 filters of size 3X 3 kernels followed by batch normalization and max pooling. This layer learns the spatial features, which helps for classification of emotions.

The batch normalization, ReLU activation and dropout rate of 0.2 are used to improve learning stability and reduce overfitting.



	precision	recall	f1-score	support
Anger	0.96	0.93	0.94	4953
Disgust	0.99	0.94	0.96	547
Fear	0.93	0.93	0.93	5121
Happy	0.96	0.98	0.97	8989
Sad	0.94	0.93	0.94	6077
Surprise	0.97	0.96	0.97	4002
Neutral	0.94	0.95	0.94	6198
avg / total	0.95	0.95	0.95	35887

Figure 5. Confusion matrix for VideoEmoNet model

The Figure 5 presents the confusion matrix of the VideoEmoNet model classification performance of VideoEmoNet model. The model achieves accuracy of 95.01% which, indicates strong and consistent performance of the model for seven different emotion classes. The diagonal elements of confusion matrix indicate correctly classified samples.

4.4 Training Configuration

To maintain fair comparison, both VideoEmoNet and AudioEmoNet model used same pre-processing and evaluation techniques. The training and implementation parameters are summarized in Table 4.

Table 4. Training and implementation parameter

Parameters	**Audio Model**	**Video Model**
Model Name and Architecture	AudioEmoNet	VideoEmoNet
Dataset	RAVDESS,TESS,SAVEE,CREMA-D	FER2013
Input datatypes	Acoustic features	Grayscale facial images
Deep Learning Framework	Tensor Flow/Keras	Tensor Flow/Keras
Input Dimension	MFCC,ZCR,RMSE	48 X 48 Grayscale facial image
Output dimension / classes	7 emotion classes	7 emotion classs
Convolutions filter	1024, 512, 256	64,128,256
Convolution Kernel Size	5X5,5X5,5X5,5X5,5X5,5X5	5X5,5X5,3X3,3X3,3X3,3X3
Pooling	Maxpooling	Maxpooling
Dense Layer	Final 7 class layer	7 class layer
Activation	ReLU, Softmax	ReLU, Softmax
Training algorithm /Optimizer	Adam	Adam
Learning Rate	0.001	0.001
Loss Function	Categorical Cross-entropy	Categorical Cross-entropy
Batch size	64	64
Maximum Epochs	100	100
Early stopping	yes	yes

5. Result and discussion

This section gives experimental findings for proposed framework. The comparative analysis of unimodal and fusion model is provided to demonstrate the performance improvement achieved through the fusion model.

5.1 Comparative analysis of existing models and proposed models

The Table 5 to Table 7 gives a comparative analysis of result of unimodal and fusion models in terms of accuracy, precision, recall and f1-score. visual emotion recognition and audio emotion recognition are the unimodal models for mental health assessment. The multimodal fusion model integrates visual and audio emotion recognition results to provide better accuracy in mental health analysis process.

Table 5 Comparison of audio emotion recognition models

Method	**Audio**	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-score**
1D CNN[14]	✓	SAVEE	93.00	92.00	91.00	91.00
CNN + BiLSTM[19]	✓	RAVDESS, TESS, EMO-DB	90.12	90.00	89.00	89.00
CNN+LSTM	✓	RAVDESS, TESS,	97.47	97.48	97.47	97.47

(AudioEmoNet model)		SAVEE, CREMA -D				

Table 6 Comparison of visual emotion recognition models

Method	**Visual**	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-score**
DCNN pre trained and Harr classifier[4]	✓	FER2013	82.56	65	62	63
DCNN[12]	✓	FER2013	82	80	79	79.5
DCNN (VideoEmoNet model)	✓	FER2013	95.01%	95.54%	94.60%	95.05%

Table 7 Comparison of audio-visual models

Method	**Visual**	**Audio**	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1-score**
LSTM based multimodal[23]	✓	✓	**DAIC-WOZ**	**93.86**	**93**	**93.35**	**93.70**
Dynamic Gating+ Multi-head attention + EAI[30]	✓	✓	DAIC-WOZ	93.00	93.00	92.45	92.65
Audio-visual fusion model	✓	✓	DAIC-WOZ	98.20%	98.24%	98.22%	98.20%

The experimental results show that each model contributes separately for final mental health analysis process. The audio model provides highest accuracy The audio-visual fusion model gives better results than unimodal models.

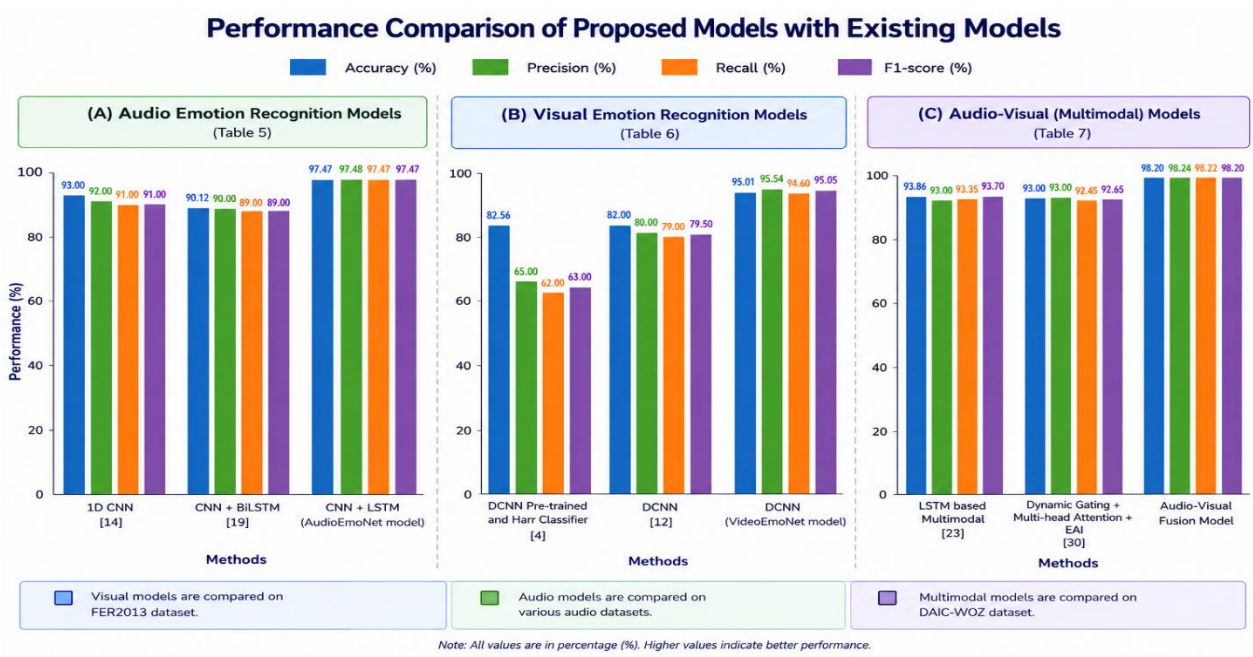


Figure 6. Performance comparison of proposed models with existing models

The graphical performance comparison of existing and proposed model in terms of accuracy, precision, recall and F1-score.are given in Figure 6. These results indicate that combining audio and visual behavioural information through the proposed fusion approach provides improved and consistent performance compared with the existing multimodal methods shown in the comparison.

6. Conclusion

This paper presented a fusion framework for mental health assessment using audio-visual emotion recognition. This framework used behavioral cues obtained from facial expressions and audio signals. The deep convolutional networks are designed to recognize facial expressions from video using the FER2013 dataset. The convolutional neural network with LSTM is designed to classify audio emotions using acoustic features extracted from the TESS, RAVDESS, SAVEE and CREMA-D datasets. The output generated by audio and visual modalities is integrated through a fusion model to provide more robust behavioral representation.

The experimental results show that the proposed AudioEmoNet model achieves 97.47% accuracy, while VideoEmoNet model achieves 95.01% accuracy. The fusion of audio and visual modality improves the reliability and robustness. of behavioral emotion analysis as compared to unimodal models. The proposed framework Provide an objective and automated approach for mental health state conditions. This framework shows how Audio-visual fusion can be used for intelligent decision-making for early mental health evaluation.

7. Future Scope

To improve accuracy and robustness of mental health assessment, the proposed model can be extended by adding additional modality like electroencephalography (EEG), electrocardiography (ECG) and wearable sensor data. The proposed objective behavioral assessment can be extended by adding subjective methods like questionnaires data collected from individuals. The subjective and objective assessment together can help to increase accuracy of the proposed framework.

The proposed framework should be validated using large and real-time multimodal dataset collected from different age groups to increase the generalizability. Explainable Artificial intelligence (EAI) techniques can also be incorporated, which helps clinicians for analysis of the mental health state. The framework can be used for continuous monitoring of mental health state by integrating it with mobile applications.

Competing interests: The Author declare no competing of interest.

Funding: This research received no external funding.

Author Contribution: All authors contributed to the conception and design of the study. Data collection, analysis, and interpretation were performed by the authors. The manuscript was written and revised collaboratively by all authors. All authors read and approved the final manuscript.

Ethical Statement: The data collection involving human participants was conducted in accordance with the applicable institutional ethical guidelines. Participants were informed about the purpose of the study, the data collection procedure, and the intended use of the recordings. Informed consent was obtained from all participants prior to data collection. The collected audio, video data were handled confidentially and were used only for research purposes.

Acknowledgements: The authors would like to acknowledge the support and resources provided by their affiliated institution during the course of this research work. The authors also gratefully acknowledge the valuable guidance and support provided by Counselling Psychologist Ms. Roopa Kanojjia, particularly regarding the interpretation of mental health assessment and emotional well-being aspects considered in this study.

Data availability Statement: The datasets used in this study are publically available at the sources given below. The FER2013 dataset was used for facial emotion recognition, while the RAVDESS, TESS, SAVEE and CREMA-D were used for audio emotion recognition. DAIC-WOZ dataset is provided by USC university of southern California on request.

- FER2013: https://www.kaggle.com/datasets/msambare/fer2013
- RAVDESS: https://zenodo.org/record/1188976
- TESS: https://tspace.library.utoronto.ca/handle/1807/24487
- SAVEE: http://kahlan.eps.surrey.ac.uk/savee/
- CREMA-D : Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) | YorVoice Catalogue

REFERENCES

1. Xiao Xu, Keyin Zhou, Yan Zhang, Yang Wang, Fei Wang and Xizhe Zhang, "Faces of the Mind: Unveiling Mental Health States Through Facial Expressions in 11,427 Adolescents" IEEE transaction May 2024
2. Xiaofeng Lu, "Deep Learning Based Emotion Recognition and Visualization of Figural Representation" Psychol., 06 January 2022 Sec. Human-Media Interaction
3. Gabrielle Simcock, Larisa T. McLoughlin, Tamara De Regt, Kathryn M. Broadhouse ,"Associations between Facial Emotion Recognition and Mental Health in Early Adolescence" International Journal of Environmental Research and Public Health 2020
4. Ala Saleh Alluhaidan , Oumaima Saidani , Rashid Jahangir , Muhammad Asif Nauman and Omnia Saidani Neffati,"Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network" Appl. Sci. 2023, 13, 4750. https://doi.org/10.3390/app13084750
5. Tae-Wan Kim andKeun-Chang KwakSpeech ,"Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques" Appl. Sci. 2024, 14, 1553. https://doi.org/ 10.3390/app14041553
6. Konstantinos Mountzouris, Isidoros Perikos , and Ioannis Hatzilygeroudis ,"Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism" Neural Networks with Attention Mechanism. Electronics 2023, 12, 4376. https://doi.org/10.3390/ electronics12204376
7. Jagjeet Singh, Lakshmi Babu Saheer and Oliver Faust ,"Speech Emotion Recognition Using Attention Model"Int. J. Environ. Res. Public Health 2023, 20, 5140. https://doi.org/10.3390/ ijerph20065140
8. Swami Mishra, Nehal Bhatnagar, Prakasam P,Sureshkumar T. R,"Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model", Multimedia Tools and Applications (2024) 83:37603–37620 https://doi.org/10.1007/s11042-023-16849-x
9. Chawki Barhoumi,Yassine BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation"Artificial Intelligence Review (2025) 58:49 https://doi.org/10.1007/s10462-024-11065-x
10. Rolinson Begazo , Ana Aguilera , Irvin Dongo and Yudith Cardinale ,"A Combined CNN Architecture for Speech Emotion Recognition" Sensors 2024, 24, 5797.https://doi.org/10.3390/s24175797
11. Rizwan Ullah , Muhammad Asif , Wahab Ali Shah , Fakhar Anjam , Ibrar Ullah , Tahir Khurshaid Lunchakorn Wuttisittikulkij ,Shashi Shah , Syed Mansoor Ali and Mohammad Alibakhshikenari,"Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer" Transformer. Sensors 2023, 23, 6212. https://doi.org/10.3390/s23136212
12. Loan Trinh Van , Thuy Dao Thi Le , Thanh Le Xuan 1 and Eric Castelli ,"Emotional Speech Recognition Using Deep Neural Networks" Sensors 2022, 22,1414. https://doi.org/10.3390/s22041414
13. Dhvanil Bhagata, Abhi Vakilb, Rajeev Kumar Guptac, Abhijit Kumard,"Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)"
14. Emmanuel Gbenga Dada ,David Opeoluwa Oyewola , Stephen Bassi Joseph , Onyeka Emebo , and Olugbenga Oluseun Oluwagbemi ,"Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10)" Hindawi Applied Computational Intelligence and Computing Volume 2023, Article ID 2457898, 19 pages https://doi.org/10.1155/2023/2457898
15. Vasundhara Rathod,Abhishek Chohan,Sakshi Nema, Adhney Nawghare,Prateeksha Devika,"Improved remote mental health illness assessment and detection using facial emotion detection and speech emotion detection" International Journal of Health Sciences, 6(S2),2022, 9577–9590. https://doi.org/10.53730/ijhs.v6nS2.7508
16. Erlangga SatrioAgung, Achmad Pratama Rifai & Titis Wijayanto,"Image based facial emotion recognition using convolutional neural network on emognition dataset", www.nature.com/scientificreports
17. Ozioma Collins Oguine , Kaleab Alamayehu Kinfu , Kanyifeechukwu Jane Oguine , Hashim Ibrahim Bisallah , Daniel Ofuani," Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction"

18. Joseph Aina , Oluwatunmise Akinniyi , Md. Mahmudur Rahman , Valerie Odero-Marah , and Fahmi Khalifa , “A Hybrid Learning-Architecture For Mental Disorder Detection Using Emotion Recognition”
19. Darius TURCIANa and Vasile STOICU-TIVADARa,”Real-Time Detection of Emotions Based on Facial Expression for Mental Health” Telehealth Ecosystems in Practice M. Giacomini et al. (Eds.) 2023 European Federation for Medical Informatics (EFMI) and IOS Press.
20. Adel Aref Ali Al-zanam a, Omer Hussein Abdou Elsayed Hussein Alhomery a, Choo Peng Tan, “Mental Health State Classification Using Facial Emotion Recognition and Detection” International Journal on advances science Engineering Information Technology
21. Shereen A. Hussein, Abd El Rahman S. Bayoumi and Ayat M. Soliman, “Automated detection of human mental disorder” Journal of Electrical Systems and Inf Technol 2023
22. Haibo Xu,Xiang Wu ,Xin Liu,”A measurement method for mental health based on dynamic multimodal feature recognisation”, PubMed December 2022
23. L. Zhang, Z. Liu, Y. Wan, Y. Fan, D. Chen, Q. Wang, K. Zhang, and Y. Zheng, “DepITCM: An audio-visual method for detecting depression,” Frontiers in Psychiatry, vol. 15, Art. no. 1466507, Jan. 2025, doi: 10.3389/fpsyt.2024.1466507.
24. S M Jishanul Islam, Sahid Hossain Mustakim, Musfirat Hossain , Mysun Mashira , Nur Islam Shourav , Md Rayhan Ahmed , Salekul Islam c, A.K.M. Muzahidul Islam , Swakkhar Shatabda ,“An audio-video based multi-modal fusion approach for emotion recognition” Knowledge-Based Systems 341 (2026) 115762
25. Yichun Li , Yi Li , Rajesh Nair , Syed Mohsen Naqvi , “A novel audio-visual model with real-time gradient modulation for mental disorder detection” Biomedical Signal Processing and Control 120 (2026) 110164
26. Gyanendra Tiwari Shivani Chauhan,,“Multimodal Depression Detection Using Audio Visual Cues” 2023 International Conference on Computer Science and Emerging Technologies (CSET)
27. Yichun Li, Shuanglin Li, Syed Mohsen Naqvi ,“A Novel Audio-Visual Information Fusion System for Mental Disorders Detection” arXiv:2409.02243v1 [cs.CV] 3 Sep 2024
28. Xinyi Zhang , Liang Zhao , Hao Sun ,“Multimodal Data Fusion Techniques for Accurate Elderly Mental State Evaluation”2024
29. Liyuan Zhang, Shuai Zhang , Xv Zhang and Yafeng Zhao,“A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals” Electronics 2025, 14, 1464
30. Lal Khan , Mohammad Zubair Khan , and Ibrahim Aljubayri ,A Comprehensive Framework for Multi-Modal Depression Detection: Integrating Adaptive Fusion, Fairness Regularization, and Explainable AI, Mathematics 2026, 14, 711
31. Wanqing Xiea,b,c,1, Chen Wangd,1, Zhixiong Line, Xudong Luoe, Wenqian Chend, Manzhu Xuf, Lizhong Liange,g, Xiaofeng Liuc,h, Yanzhong Wangi, Hui Luoj, Mingmei Chenga,b,” Multimodal fusion of depression and anxiety based on CNN-LSTM model” https://doi.org/10.1016/j.compmedimag.2022.102128
32. A. F. K. Basha, R. M., D. Abishek, and D. B. Haripriya, “Real Time Emotion and Stress Recognition for Mental Health Monitoring Using Speech and Facial Recognition,” International Journal of Drug Delivery Technology, vol. 16, no. 47S, Jun. 2026, doi: 10.25258/IJDDT.16.47S.118
33. Filipe Fontinele de Almeida, Kelson Rˆ omulo Teixeira Aires, Andr´ e Castelo Branco Soares, Laurindo de Sousa Britto Neto, Rodrigo de Melo Souza Veras,” A Multimodal Fusion Model for Depression Detection Assisted by Stacking DNNs” March 11, 2024
34. https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en
35. https://www.kaggle.com/datasets/msambare/fer2013
36. Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, “The Distress Analysis Interview Corpus of human and computer interviews”, Proceedings of Language Resources and Evaluation Conference (LREC), 2014
37. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). “SimSensei kiosk: A virtual human interviewer for healthcare decision support”. In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS’14), Paris
38. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S., COVAREP - A collaborative voice analysis repository for speech technologies. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), pages 960-964, 2014.
39. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders. 2009 Apr 30;114(1):163-73.