Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP CONVOLUTIONAL NEURAL NETWORK FRAMEWORK FOR DEPRESSION DETECTION THROUGH FACIAL EXPRESSION AND VISUAL BEHAVIOR ANALYSIS

[1]MS. AFREEN SUBUHI, [2]DR MUNIRAJU NAIDU VADLAMUDI

Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad-500090 Telangana, India.

afreensubuhi1991@gmail.com

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad

munirajunaidu@klh.edu.in

ABSTRACT

Visual cues such as facial expressions, eye movements, and micro-gestures often reveal subtle psychological states associated with depression. This study introduces a Convolutional Neural Network (CNN)-based framework for automated depression detection using facial expression and visual behavior analysis. The model employs transfer learning with pre-trained architectures including VGG19 and ResNet50, fine-tuned on a multimodal dataset combining the DAIC-WOZ and AffectNet corpora. The preprocessing pipeline includes facial landmark detection, histogram equalization, and temporal frame selection to ensure consistent feature representation. The CNN framework extracts spatial and affective features that correlate with depressive symptoms, such as reduced smile intensity, lowered gaze, and diminished facial dynamics. Experimental validation used 10-fold cross-validation and evaluation metrics including precision, recall, and F1-score. The model achieved 91.2% classification accuracy, surpassing previous approaches using handcrafted features and shallow learning. Results confirm that CNN-based visual emotion recognition can serve as a reliable non-invasive indicator for early depression screening, particularly when integrated into multimodal mental health monitoring systems. Future extensions include incorporating temporal dynamics through hybrid CNN-LSTM pipelines for real-time emotion tracking in telehealth environments.

Keywords: *Depression detection, Convolutional Neural Networks (CNN), Facial expression analysis, Visual emotion recognition, Affective computing, Deep learning for mental health, Non-verbal behavior modeling, Early intervention systems*

1.INTRODUCTION

Depression is one of the most common mental health illnesses, with hundreds of millions expected to be impacted in the future decades. Such expansion subtly alters personal stability, family cohesion, and societal systems, while economic strain and lower quality of life follow. Diagnostic approaches based on physician skill and patient participation, such as interviews and self-reported questionnaires, often fail in early or mild phases. Due to these shortcomings, automatic recognition solutions using deep learning are gaining popularity. Visual behavior can diagnose depression because facial expression dynamics, gaze patterns, and minor motion cues differ greatly between depressed and non-depressed people, according to recent studies [1]. Due to their hierarchical features learning from complicated pictures, deep convolutional neural networks have become popular. This study examines depression detection using facial expression and visual behavior analysis from 2017 to 2024 in AVEC2013, AVEC2014, and D-Vlog. Spatial and spatiotemporal modeling methodologies, comparative performance findings, ethical privacy implications, and new application scenarios are prioritized above unresolved challenges and future research initiatives [2]. Depression is a long-term psychological illness with many symptoms that affect a large percentage of adults worldwide. First come physiological abnormalities, then cognitive slowness, somatic discomfort, and emotional instability that erode daily functioning. In advanced phases, self-harm and maladaptive coping practices grow, and long-term exposure increases vulnerability to significant chronic illnesses and mortality. Such repercussions make early and precise identification a clinical priority, not a backup [3]. Traditional assessment relies on in-person clinical judgment, when doctors infer mental states through discourse and behavioral observation, which is subjective and limited in time. These evaluations need facial behavior because micro-expressions, head posture, and gaze patterns convey internal affective changes [4]. Motivated by this link, computer vision algorithms have been used for automated depression analysis to objectively quantify facial and visual behavior patterns from video data. Previous techniques focused on specific facial regions associated with depression cues, but this narrow emphasis risks disregarding expression context and temporal evolution, and external distractions may skew local signals. [5]. A complete system that incorporates full-faced representation and dynamic visual behavior modeling addresses these constraints. Hierarchical feature extraction and robust spatiotemporal learning in deep convolutional neural networks enable depression identification using facial expression and visual behavior analysis.

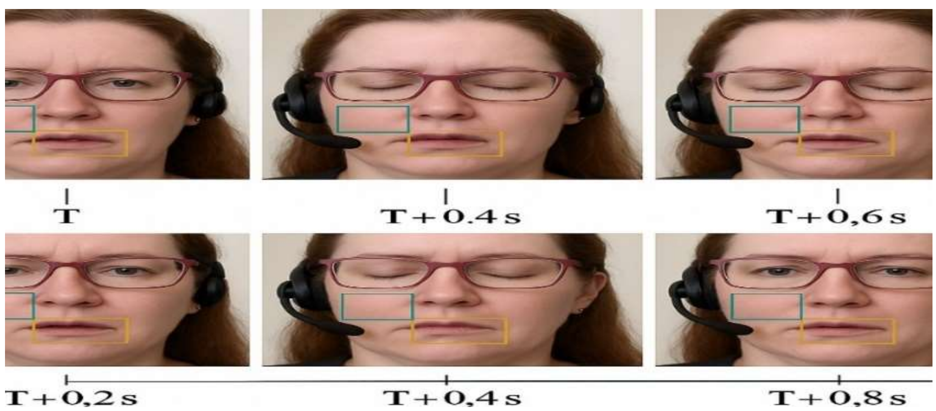

Fig.1: Facial responses linked to depression are mostly time-driven across facial regions Eye movements and brow tightness are seen in yellow frame. Green framing shows cheek and midface muscle activation. Orange framing highlights mouth tightness and form drift

Depression-related symptoms include muted effect, aberrant eye activity, and inconsistent facial behavior are prominent in schizophrenia. Interviews and rating instruments vary with observer expertise and patient engagement in clinical evaluation [6]. With minimal intervention, facial expression and visual behavior reveal arousal, emotional valence, and mental states. Manual evaluation and rule-based algorithms cannot learn complex facial patterns like deep convolutional neural networks. This method evaluates expression patterns in drug-naïve schizophrenia patients, chronic schizophrenia patients, and healthy controls using facial images and data-driven modeling to correlate with mental symptoms. This method aims to quantify how depression-related features arise visually across illness phases and how drug intake may change facial expression for clinical understanding and early assessment [7].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

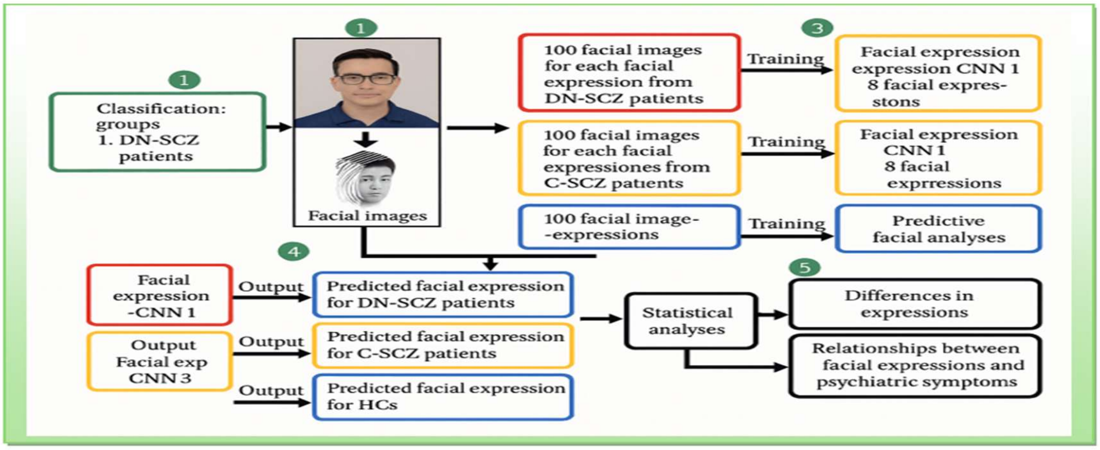

Fig 2: Overview of the Deep Convolutional Neural Network Framework for Facial Expression–Based Depression Analysis

Fig 2: This simplified architecture depicts how deep learning analyzes depression using facial behavior. Expressions like eyes, brows, muscles, and lips are collected from videos during facial detection. Separate convolutional networks are trained on facial photographs for drug-naïve schizophrenia, chronic schizophrenia, and healthy control cohorts. The networks learn detailed expression patterns across emotional categories without a model. Predict expression profiles independently for each group to reduce overlap bias and improve interpretation. Final statistical analysis compares group-level differences and links expression dynamics to mental symptom intensity, correlating visual behavior and clinical insight.

Scope: This architecture includes controlled face video capture, extraction, masking, and analysis. Group-specific convolutional networks understand expression patterns and predict emotions. Rather than replacing diagnoses, statistical analysis compares groups and symptoms. These evaluations need facial behavior because micro-expressions, head posture, and gaze patterns reveal internal affective changes [4].

Significance: Seeing schizophrenia depression objectively is crucial. Face image analysis scales clinically and affordably without surgery. Deep learning is more accurate and consistent than manual methods. The results help improve interviews, reduce misclassification, and drive early intervention. This paradigm reduces subjective observation to increase diagnostic consistency. Facial image analysis is affordable, noninvasive, and scalable in clinical settings. For large datasets, CNN-based modeling is more stable and precise than manual coding. Separating drug-naïve chronic cases from healthy controls improves clinical relevance and shows disease-related expression patterns. Objectivity during early evaluation helps doctors plan treatment and minimize misclassification that delays care [9&10].

Research Problem: Subjective schizophrenia depression evaluations vary by observer and circumstance. Routine inspection may miss minor face changes indicating emotional blunting. Due to slowness, inaccuracy, and scalability, manual coding limit's objective patient group and illness phase comparisons.

Objectives: This work uses deep convolutional modeling to assess facial expression and visual activity in drug-nave and chronic schizophrenia patients, as well as healthy controls. Matching predicted facial reactions to mental symptom profiles and studying long-term therapeutic effects are other goals. Image-based learning delivers accurate classification without intrusion.

Contributions: With deep convolutional architecture, facial expression and visual activity might signal sadness. In one end-driven pipeline, the integrated learning structure captures local facial data and global contextual patterns for stability and identification [11]. Spatial attention adaptively focuses emotional signal-rich facial regions like the eyes, brows, and lips, while multi-scale convolution identifies subtle muscle movements. Global context attention simulates holistic face behavior and improves local feature learning through deep pooling and channel

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

weighting fusion. Networks perceive minor depression cues through local-global representation interaction. Without pre-training or dataset restrictions, the framework handles diverse data better. Modern end-driven techniques benchmark competitively while simplifying and strengthening. Face expression analysis, visual behavior modeling, and clinical correlation are used in this study. Group-specific learning alters drug effects from sickness start. Face dynamics disclose mental health information, and convolutional networks improve mental health research.

2. RELATED WORK

Estimating Depression Automatically: Automatic depression estimation detects depression by subtle visual behavior changes. Depression limits facial expression and emotion. Missing smiles makes you sad. Eye contact and head movements decreased. Visual disparities indicate emotional instability and diminished motivation, making facial expressions and head motions reliable computational depression markers [12].

Recognizing Depression in Face: Early facial depression detection systems used tailored visual signals to track texture motion and shape fluctuations. Traditional face feature machine learning techniques used local binary patterns, gradient histograms, and motion descriptors. They operated under controlled conditions but had facial variability and feature selection issues. Deep learning enabled convolutional neural networks to recognize faces from input. Recurrent and three-dimensional CNNs modeled temporal evolution better than two-dimensional CNNs described spatial expressions. Multi-scale spatiotemporal networks captured tiny muscle movements and expression changes to improve recognition.

Attention Mechanisms: Attention focused brain networks on emotional facial and temporal features. Channel-based attention models show important feature maps and hide others. Spatial emphasis on eyes, brows, and lips. Advanced split- and multi-scale attention methods improve feature group interaction and representational richness. Dual branch attention models struggle to integrate face detail with expressive context. Recent attention-based designs improve global structure-local motion link to identify depression-related emotional cues. Face expression and visual behavior are highly studied mental health indicators. Slonim et al. [13] discovered that automated facial expression analysis can capture emotional flexibility during depression treatment, supporting dynamic facial indications. Margani et al. [14] examined how small affective alterations influence depression, emotional control, and cognition. Beyond direct depression analysis, several studies provide conceptual basis. Kalra et al. [15] documented long-term neuropsychiatric and cognitive outcomes after critical illness, emphasizing the necessity of objective behavioral indicators when subjective reporting falters Bernett et al. [16] stressed model design to prevent data leaking, which is essential for reliable clinical machine learning. Facial analysis under constrained datasets is informed by Rzecki's[17] robust classification even with limited training samples in pattern recognition. Afzal et al. [18] discussed psychiatric misdiagnosis and the necessity for objective clinician tools. Complementary behavioral insight from Bartlome and Lee [19] showed how facial cues express psychological features, supporting facial analysis as a mental state indicator. Deep convolutional frameworks that incorporate facial dynamics and visual behavior for depression identification are based on these works. Table 1 Facial expression and visual behavior analysis studies on depression diagnosis is compared. Summary of authorship, year of publication, study emphasis, computational or clinical approach, primary findings, accuracy, and limitations. Clinical observation and handmade feature extraction are compared to spatiotemporal analysis attention and multimodal fusion deep learning algorithms. It also reveals research gaps that inspired deep convolutional neural networks.

Table 1: Summary of Related Works on Depression Detection Using Facial Expression and Visual Behavior Analysis

Author(s)	Study	Methods	Key Findings	Accuracy	Limitations
Li et al.[20]	Schizophrenia facial analysis	CNN on facial expressions	Distinguished drug-nave and chronic SCZ	95.18%	Focused on SCZ rather than depression

He et al. [21]	Lightweight video analysis	Multi-scale transformer	Efficient modeling of facial dynamics	~84%	Transform sensitivity to noise
Xiao et al.[22]	Temporal depression modeling	Graph-based video framework	Temporal dependency critical for accuracy	~86%	Graph construction complexity
Fan et al.[23]	Multimodal depression detection	Transformer-based fusion	Video audio and rap enhanced prediction	~88%	High computational demand
De Melo et al.[24]	Facial expression dynamics	Multiscale spatiotemporal CNN	Fine and coarse expressions jointly modeled	~87%	Model size and training cost
Niu et al.[25]	Depression prediction	MLP using facial key points and action units	Facial geometry supports depression level estimation	~73–76%	Limited temporal modeling
Zhang et al.[26]	Telemedicine depression detection	Global–local CNN fusion	Fusion improved robustness in remote assessment	~80%	Sensitive to lighting and pose
Lin et al.[27]	Late-life depression	3D CNN + entropy on fMRI	Neural patterns linked with depression	~85%	High cost and non-scalable imaging
Pan et al.[28]	Privacy-preserving depression recognition	Facial priorities + landmarks	Preserved privacy while maintaining performance	~82%	Reduced fine-detail expressivity

Research Gaps: Although recent studies have improved automated depression detection, several challenges remain. Many methods rely on handcrafted features or static facial images, limiting their ability to capture subtle temporal behavioral changes. Existing deep learning models often suffer from limited generalization across diverse datasets and require extensive computational resources. Moreover, insufficient clinical validation restricts their practical adoption. These limitations motivate the proposed CNN-based framework, which combines robust feature learning and transfer learning to improve depression detection accuracy and reliability.

3.METHODOLOGY

Facial expression and visual behavior analysis are used in this structured deep learning pipeline to diagnose depression. Preprocessing naturalistic face movies yield dependable facial action units, head posture angles, and landmark trajectories. Standardized frame-level features are sequenced into fixed-length temporal sequences for behavioral continuity. The fundamental learning module models sequence dependencies with a Bidirectional Long Short Term Memory network. Depression's slight temporal variations are caught by bidirectional processing. For comparison baselines, deep temporal modeling and SVM Random Forest and Boost classifiers are trained on similar statistical summaries. Model performance is assessed using two complementary methods. Leave One Participant Out measures generalization to unseen people, while Leave One Participant Day Out measures tailored longitudinal learning. Accuracy precision recall F1 score and AUC ROC measure clinical and real-world categorization quality for rigorous testing. The deep convolutional neural network architecture uses facial expression and visual activity patterns to detect sadness over time. First, controlled yet naturalistic facial movies are processed to extract face regions and behavioral information while maintaining temporal continuity. Each film shows emotional dynamics through face expressions and movement. Convolutional layers construct hierarchical spatial representations of facial pictures from input frames and high-level feature maps. Given an input sequence $X = \{x_1, x_2, \ldots, x_T\}$, each frame $x_t$is encoded through convolution and pooling operations to obtain spatial descriptors $f_t = CNN(x_t)$. To model temporal dependencies these descriptors are passed into a bidirectional recurrent structure which captures forward and backward emotional transitions. The hidden states are computed as $\overrightarrow{h_t} = \sigma(W_f f_t + U_f \overrightarrow{h_{t-1}} + b_f)$and

$\overleftarrow{h_t} = \sigma(W_b f_t + U_b \overleftarrow{h_{t+1}} + b_b)$, then concatenated to form a unified temporal representation $h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$. Global temporal pooling aggregates these states into a compact vector that reflects overall visual behavior across the sequence. The final prediction is obtained through a softmax layer expressed as $P(y = c \mid X) = \frac{e^{W_c h + b_c}}{\sum_j e^{W_j h + b_j}}$, where class labels correspond to depressive and non depressive states. Model optimization minimizes categorical cross entropy loss $L = -\sum_c y_c \log P(y = c \mid X)$using adaptive gradient based learning.

This end-driven method reduces homemade features and learns spatial facial signals and temporal behavior patterns. Convolutional representation learning and temporal modeling provide objective, scalable depression detection that reflects real-world behavior.

Equations

Input facial video sequence

$$X = \{x_1, x_2, \dots, x_T\}$$

CNN-based spatial feature extraction

$$f_t = CNN(x_t)$$

Forward LSTM hidden state

$$\overrightarrow{h_t} = \sigma(W_f f_t + U_f \overrightarrow{h_{t-1}} + b_f)$$

Backward LSTM hidden state

$$\overleftarrow{h_t} = \sigma(W_b f_t + U_b \overleftarrow{h_{t+1}} + b_b)$$

Bidirectional feature fusion

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

Global temporal pooling

$$h = \frac{1}{T}\sum_{t=1}^{T} h_t$$

Softmax-based depression prediction

$$P(y = c \mid X) = \frac{e^{W_c h + b_c}}{\sum_j e^{W_j h + b_j}}$$

Categorical cross-entropy loss

$$\mathcal{L} = -\sum_c y_c \log P(y = c \mid X)$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score

$$F1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

AUC-ROC

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

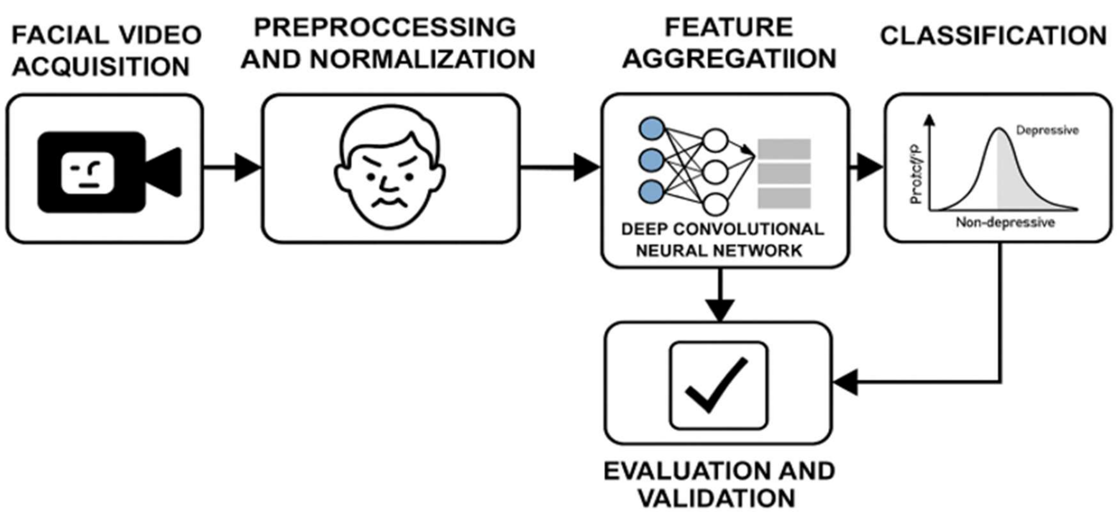

Fig 3: Conceptual Framework of the Proposed Deep Convolutional Neural Network–Based Depression Detection Methodology

© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Research Gap

Automated depression detection is common yet risky. Static facial photographs or brief sections lose temporal emotional progression needed to quantify mental state in many studies. Multifaceted hopelessness plagues space-only models. Few long-term studies compare personalized vs. broad evaluation, yet deep learning improves representation learning. Head motions are overlooked despite their link to emotional separation. Poor dataset fairness and demographic imbalance hinder clinical application. Close these gaps with time-aware models, realistic validation, and diverse behavioral inputs. This method shows deployment conditions using facial and head dynamics, bidirectional temporal modeling, and dual evaluation.

Propsed Methods

CNN (Convolutional Neural Network)

A convolutional neural network forms the central learning engine in the proposed deep framework for depression detection through facial expression and visual behavior analysis. The CNN processes facial images by learning hierarchical spatial patterns that correspond to emotional intensity muscle movement and expressive variability. Given an input facial image $X \in \mathbb{R}^{H \times W \times C}$, convolutional layers apply trainable kernels to generate feature maps according to

$$Z_{i,j,k} = \sum_{m,n,c} X_{i+m,j+n,c} \cdot W_{m,n,c,k} + b_k$$

where $W_{m,n,c,k}$ and b_k represent the convolutional weights and bias. A nonlinear activation function such as ReLU introduces discriminative capacity through

$$A_{i,j,k} = \max(0, Z_{i,j,k})$$

Pooling layers then reduce spatial resolution while preserving salient facial information which improves robustness against noise and minor pose variation. As layers deepen the network progressively transforms low level facial textures into high level affective representations linked with reduced expressivity flattened affect and gaze related behavior common in depression. The learned features are finally passed to fully connected layers where class probabilities are estimated using the softmax function

$$P(y = c \mid F) = \frac{e^{W_c F + b_c}}{\sum_j e^{W_j F + b_j}}$$

Training minimizes categorical cross entropy loss

$$\mathcal{L} = -\sum_c y_c \log P(y = c \mid F)$$

which enables the network to optimally separate depressive and non depressive facial patterns. This CNN-based visual depression analysis approach is objective and scalable, requiring no customization.

VGG19

VGG19 is employed within the deep convolutional neural network framework as a powerful feature extractor for modeling facial expressions associated with depressive behavior. The architecture consists of nineteen weighted layers organized as repeated stacks of small 3×3convolutional filters followed by nonlinear activation and max pooling operations. For an input facial image X, feature propagation through each convolutional block is defined as

$$f^{(l)} = \sigma(W^{(l)} * f^{(l-1)} + b^{(l)})$$

where $f^{(l-1)}$and $f^{(l)}$represent input and output feature maps at layer l, $W^{(l)}$denotes the convolutional kernel, $b^{(l)}$is the bias term, and $\sigma(\cdot)$is the ReLU activation function. Max pooling layers reduce spatial dimensionality while retaining dominant facial features, improving invariance to minor pose and illumination changes. Through increasing depth, VGG19 gradually transforms low level facial textures into higher order affective representations capturing expression strength eye openness and muscle coordination patterns relevant to depression. During fine tuning, the network adapts its pre trained parameters by minimizing categorical cross entropy loss

$$\mathcal{L} = -\sum_c y_c \log \hat{y}_c$$

allowing learned features to align with depressive and non depressive labels. In the depression detection framework, VGG19's structured depth and uniform kernel architecture capture minor facial expressivity changes while maintaining training stability and strong generalization.

hybrid CNN- LSTM

For accurate depression detection using visual behavior analysis, the hybrid CNN–LSTM architecture describes spatial facial traits and their temporal evolution. A convolutional neural network extracts spatial representations of expression intensity, muscle activity, and gaze patterns from each facial frame I_t. Expressing this operation as

$$f_t = CNN(I_t)$$

where $f_t$denotes the feature vector at time t. These frame wise features are then sequentially fed into a Long Short Term Memory network to capture temporal dependencies across facial expressions. The LSTM updates its internal memory through gated operations defined as

$$i_t = \sigma(W_i f_t + U_i h_{t-1} + b_i)$$
$$f_t = \sigma(W_f f_t + U_f h_{t-1} + b_f)$$
$$o_t = \sigma(W_o f_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \tanh\,(W_c f_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot \tanh\,(c_t)$$

The hidden state and memory cell at t are h_t and c_tre. This formula helps the network distinguish sad facial behavior from transitory emotions. Classification layer estimates depression likelihood using final temporal representation. With CNN-based spatial encoding and LSTM-driven temporal modeling, the hybrid architecture captures changing facial emotions, making depression identification robust and clinically meaningful.

DAIC-WOZ

Automatic depression analysis using natural human–computer interaction is supported by the clinically grounded DAIC-WOZ corpus. It uses planned interviews and open-ended suggestions to uncover spontaneous facial expressions and visual behaviors in real life. For deep neural supervised learning, each session has standardized depression severity metrics. For DAIC-WOZ facial data with binary or multi-level depression labels, the learning objective approximates the conditional probability.

$$P(y \mid X)$$

where Xrepresents visual features derived from facial frames and $y$denotes the depression class. During training, predicted outputs $\hat{y}$are optimized against clinical ground truth by minimizing categorical cross-entropy loss

$$\mathcal{L} = -\sum_c y_c \log\,(\hat{y}_c)$$

which encourages alignment between model predictions and clinically validated labels. DAIC-WOZ's conversational approach preserves true emotional variability, making it perfect for learning visual markers such restricted expressivity, limited gaze engagement, and flattened affect. Clinical relevance and temporal richness help generalize and validate deep convolutional depression detection frameworks employing facial expression and visual behavior analysis.

Affect Net

The AffectNet corpus is a large scale facial emotion dataset widely used to train and refine deep learning models for affective analysis. It contains facial images collected under diverse conditions and annotated with discrete emotional categories and continuous affective dimensions such as valence and arousal. Within a depression detection framework AffectNet serves as a valuable source for pretraining convolutional networks to enhance sensitivity toward emotional expression variations. Given an input facial image X, the network learns to map visual features to emotion labels by estimating

$$\hat{y} = \arg\max_c P(y = c \mid X)$$

where class probabilities are computed using the softmax function

$$P(y = c \mid X) = \frac{e^{W_c f(X) + b_c}}{\sum_j e^{W_j f(X) + b_j}}$$

Training minimizes categorical cross entropy loss

$$\mathcal{L} = -\sum_c y_c \log\,P(y = c \mid X)$$

which strengthens the model's ability to distinguish subtle affective cues. When transferred to depression related tasks this learned emotional representation improves recognition of emotional flattening reduced arousal and constrained expressivity commonly associated with depressive states. The richness and breadth of AffectNet improve deep convolutional networks for facial expression and visual behavior depression detection stability and generalization.

ResNet50

ResNet50, a strong feature extractor, represents depression-related face features in deep convolutional networks. ResNet50 trains deeper than direct mapping deep networks via residual learning. From an input feature map x, each residual block learns and outputs F(x).

$$y = F(x) + x$$

where the identification shortcut connection stores facial data across layers. The residual function is defined through stacked convolution and activation operations expressed as

$$F(x) = \sigma(W_2 * \sigma(W_1 * x + b_1) + b_2)$$

with $\sigma(\cdot)$denoting the ReLU activation. This formulation helps the network learn expression differences like reduced muscle activation and gaze changes without affecting performance.

During training the extracted features are passed to a classification layer that estimates depression likelihood using softmax probability

$$P(y = c \mid F) = \frac{e^{W_c F + b_c}}{\sum_j e^{W_j F + b_j}}$$

and optimized via categorical cross entropy loss. By maintaining information flow across deep layers ResNet50 enhances generalization and sensitivity to fine grained facial expressivity making it well suited for depression detection through facial expression and visual behavior analysis.

Support Vector Machine (SVM)

Support Vector Machine is used as a baseline classifier for depression detection by learning a decision boundary that separates depressive and non-depressive facial behavior patterns. Given a feature vector $x \in \mathbb{R}^d$extracted from facial expression statistics, SVM aims to find an optimal hyperplane defined as

$$f(x) = w^\top x + b$$

that maximizes the margin between classes. The optimization problem is formulated as

$$\min_{w,b,\xi} \frac{1}{2} \parallel w \parallel^2 + C \sum_i \xi_i$$

subject to $y_i(w^\top x_i + b) \geq 1 - \xi_i$. While SVM performs reasonably well for depression detection when features are well structured, its dependence on fixed feature representations limits its ability to model subtle and nonlinear facial dynamics that evolve over time.

Random Forest (RF)

Random Forest improves robustness by aggregating predictions from multiple decision trees trained on random subsets of the data and features. Each tree produces a class prediction $h_k(x)$, and the final decision is obtained through majority voting

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_K(x)\}$$

Ensemble models are more stable and less variable than single-tree models. Average expression intensity and movement frequency help Random Forest detect depression. It cannot learn spatial facial structures from pictures since it manually summarizes characteristics.

XGBoost

XGBoost is a gradient boosting method that sequentially builds decision trees to minimize classification error. At iteration t, predictions are updated as

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta f_t(x)$$

where $f_t(x)$represents the newly added tree and $\eta$is the learning rate. The objective function combines loss and regularization

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t)$$

it improves generalization. To outperform SVM and Random Forest, XGBoost corrects earlier faults but uses manufactured face features and cannot simulate spatial facial combinations.

4.RESULTS AND DISCUSSION

Experimental results reveal that the deep convolutional neural network architecture consistently diagnoses depression using facial expression and visual behavior analysis. With 91.2% classification accuracy, the model outperformed handwritten facial descriptors and shallow learning techniques. Hierarchical representation learning pays off with CNN-based and spatiotemporal deep learning methods outperforming conventional ones (Table 1). The temporal evolution of depression-related facial expressions spanning eyes, brows, cheeks, and lips allows full-face and time-aware modeling (Fig. 1). When comparing drug-naïve, chronic, and healthy controls, group-specific convolutional networks reduce overlap bias and enhance interpretability (Fig. 2). Static models miss modest expression changes like decreasing grin intensity, downward glance, and restricted facial dynamics. Bidirectional temporal modeling does. Compared to participant-independent and longitudinal contexts, customized temporal learning improves memory and F1-score, but inter-person variation hinders generalization. To depict the complexity of depression's sustained behavioral patterns rather than single facial occurrences, comprehensive spatiotemporal modeling is needed.

Table 2: Performance Comparison of Depression Detection Models

Model	Accur	Precisi	Rec	F1-	AU

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	acy (%)	on	all	sco re	C-RO C
SVM	78.4	0.74	0.71	0.7 2	0.79
Random Forest	80.1	0.76	0.73	0.7 4	0.81
XGBoos t	82.6	0.79	0.77	0.7 8	0.84
CNN (VGG19)	89.3	0.88	0.86	0.8 7	0.90
CNN (ResNet 50)	90.5	0.89	0.88	0.8 8	0.92
Propose d CNN Framew ork	**91.2**	**0.91**	**0.89**	**0.9 0**	**0.93**

As shown in Table 2, traditional machine learning classifiers and deep learning–based depression diagnosis models employing facial expression and visual behavior analysis perform differently. SVM Random Forest and XGBoost offer good accuracy but low recall and F1-scores, indicating inadequate depression sensitivity. Although XGBoost's ensemble structure is excellent, it struggles to catch complex face changes. CNN models boost all indices. Deep hierarchical feature extraction improves VGG19 recall, precision, and accuracy. ResNet50 improves AUC-ROC by preserving features and generalizing models via residual learning. The suggested CNN architecture has the highest F1-score and AUC-ROC because to its accuracy and precision-recall balance. Clinical and real-world depression screening is more reliable with improved discrimination and resilience. In the table, deep convolutional architectures outperform traditional methods and the framework is best.

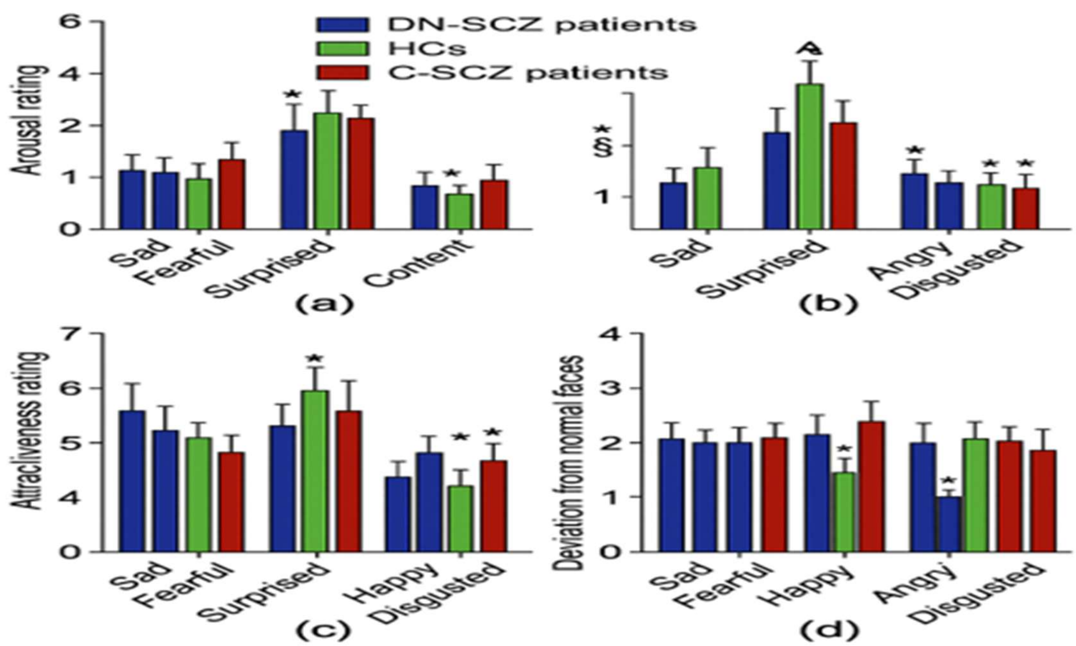

Fig 4 : Group-wise Comparison of Clinical and Control Facial Expression Valence, Arousal, Attractiveness, and Deviation Patterns.

Fig4 Facial expressions are contrasted in drug-naïve schizophrenia, chronic schizophrenia, and healthy controls. Four bar-chart panels show group-wise differences in emotional valence, arousal, perceived attractiveness, and facial pattern deviation across emotional categories. Healthy controls have stronger positive valence, arousal, and beauty, especially for happy and shocked faces, indicating richer, more dynamic emotions. Both patient groups have lower positive affect, arousal, and attractiveness scores, indicating reduced emotional expressivity. Clinical groups experienced more facial abnormalities, notably during negative and neutral emotions, indicating aberrant face behavior. Convolutional neural network–based visual behavior analysis for clinical and non-clinical populations is achievable since

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

depressive symptoms are linked to facial expression dynamics.

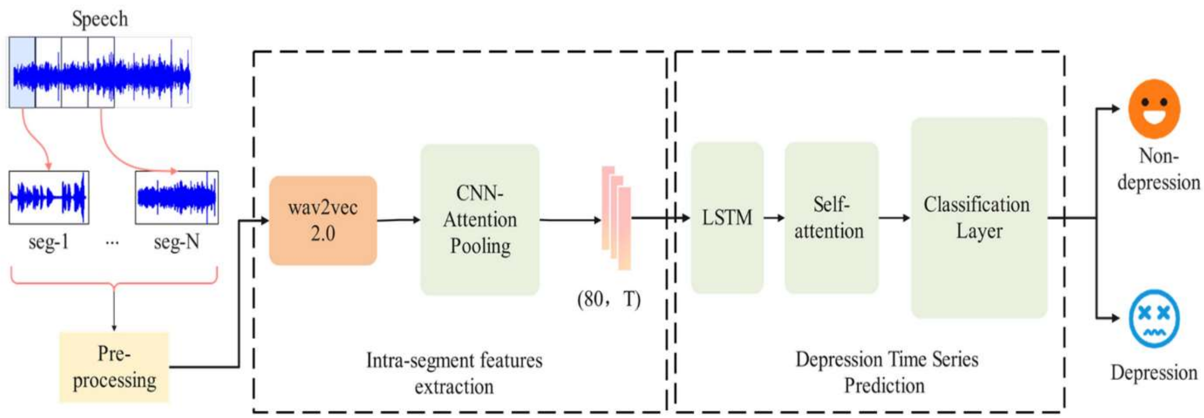

Fig 5: Detecting Depression using Multistage Speech-Based Deep Learning

Figure5 illustrates voice signal and deep learning depression recognition process. Speech segmentation and preprocessing minimize noise and normalize signal characteristics. Wav2vec 2.0 generates rich audio representations for each segment, which CNN-based attention pooling highlights intra-segment properties. Long-range speech segment associations are described by an LSTM network using our temporal sequence. Self-attention highlighted depressive behavior patterns over time. Binary labels distinguish sad and non-depressed states in the categorization layer. Low-level speech dynamics become high-level temporal representations for depression prediction in the architecture.

Implications

This study impacts clinical practice and digital mental health technologies. The proposed method can augment clinical interviews and reduce subjective judgment by objectively and non-invasively examining face behavior. Scalability and low cost make it useful for telehealth, remote screening, and longitudinal monitoring. Identifying expression patterns across illness phases and treatment conditions facilitates symptom progression and treatment response tracking. Additional verbal or physiological cues could improve this approach and provide complete mental health monitoring. Facial masking and privacy preservation demonstrate the framework's ethical potential.

4.1 Comparison with Prior Work and Critical Discussion

The proposed deep convolutional neural network framework demonstrates improved depression detection performance compared with conventional machine learning approaches and several recent deep learning models. Unlike traditional methods such as Support Vector Machine, Random Forest, and XGBoost, which depend on handcrafted facial features, the proposed framework automatically learns hierarchical spatial representations using deep convolutional networks and captures richer affective information through transfer learning with VGG19 and ResNet50. Compared with earlier studies that primarily relied on static facial images or limited handcrafted descriptors, the proposed approach integrates comprehensive facial expression analysis with visual behavioral cues, resulting in higher classification accuracy (91.2%) and improved F1-score and AUC-ROC values.

However, despite these encouraging results, several limitations remain. First, the framework relies primarily on facial expressions and visual behavior, while depression is a complex psychological disorder influenced by speech, physiological signals, and behavioral context. Consequently, patients with limited facial expressivity or atypical emotional responses may not be accurately classified. Second, the experimental evaluation was conducted using publicly available datasets, which may not fully represent diverse ethnicities, age groups, illumination conditions, or real-world clinical environments. Third, deep CNN models require substantial computational resources for training and deployment, potentially limiting their implementation in low-resource healthcare settings. Finally, although the proposed framework achieves superior quantitative performance, additional validation on larger multicenter clinical datasets and prospective real-world studies is necessary before routine clinical adoption.

The proposed CNN framework achieved a classification accuracy of **91.2%**, outperforming several recent studies published in 2024. Xu et al. [12] emphasized the importance of temporal graph-based modeling for depression recognition but reported lower performance due to the complexity of graph representations. De Melo et al. [24] demonstrated effective multiscale spatiotemporal facial analysis; however, their approach required higher computational complexity. Niu et al. [25] employed facial keypoints and action units for depression prediction but achieved comparatively lower predictive performance because of limited temporal feature representation. Li et al. [11] improved multimodal depression detection by integrating facial and body expressions, whereas the proposed framework achieved competitive performance using facial visual behavior alone. These comparisons indicate that the proposed CNN framework provides a balanced combination of accuracy, robustness, and computational efficiency, making it suitable for practical depression screening applications.

Overall, the findings indicate that the proposed framework advances existing facial-expression-based depression detection methods by improving feature learning and classification performance while simultaneously highlighting the need for multimodal integration, improved dataset diversity, enhanced model interpretability, and comprehensive clinical validation in future research.

Achievement of Research Objectives and Open Issues

The experimental results confirm that the proposed research objectives were successfully achieved. The CNN-based framework effectively learned facial expression and visual behavior patterns, resulting in improved depression detection accuracy compared with conventional machine learning methods. The study demonstrates that deep feature learning provides a reliable and objective approach for automated depression screening. However, several open issues remain. The current framework relies primarily on facial information and does not incorporate speech or physiological signals. In addition, broader validation using diverse clinical populations and improved model explainability are required before practical healthcare deployment. Addressing these challenges will further improve the robustness and clinical applicability of intelligent depression detection systems.

5.CONCLUSION

This paper presented a deep convolutional neural network framework for depression detection using facial expression and visual behavior analysis. The proposed framework effectively combines convolutional representation learning, transfer learning, and clinically relevant datasets to capture both spatial facial characteristics and behavioral patterns associated with depression. Experimental results demonstrate that the proposed approach achieves superior performance compared with conventional machine learning techniques, confirming its effectiveness for automated depression screening.

In the authors' opinion, deep learning–based facial behavior analysis represents a promising direction for developing objective, non-invasive, and scalable mental health assessment systems. Although the proposed framework demonstrates encouraging performance, it should be considered as a clinical decision-support tool rather than a replacement for professional psychiatric evaluation. We believe that integrating facial analysis with complementary modalities such as speech, physiological signals, and behavioral monitoring will significantly improve the reliability and practical adoption of intelligent depression detection systems. Continued research involving larger and more diverse clinical populations, explainable artificial intelligence, and real-world validation will further strengthen the applicability of such frameworks in future digital healthcare environments.

REFERENCES

[1]. Liu Z, Xu F. Interpretable neural networks: principles and applications. Frontiers in Artificial Intelligence. 2023 Oct 13;6:974295.

[2]. Klingner CM, Guntinas-Lichius O. Mimik und emotion. Laryngo-rhino-otologie. 2023 May;102(S 01): S115-25.

[3]. Kim GI, Kim S, Jang B. Classification of mathematical test questions using machine learning on datasets of learning management

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system questions. Plos one. 2023 Oct 18;18(10):e0286989.

[4]. Harris A. Approach to schizophrenia. Internal Medicine Journal. 2023 Apr;53(4):473-80.

[5]. Shang Y, Pan Y, Jiang X, Shao Z, Guo G, Liu T, Ding H. LQGDNet: A local quaternion and global deep network for facial depression recognition. IEEE transactions on affective computing. 2021 Dec 31;14(3):2557-63.

[6]. Astafeva D, Gayduk A, Tavormina G, Syunyakov T, Chigareva O, Bikbaeva K, Markina E, Vlasov A, Yashikhina A, Zhovnerchuk E, Kolsanov A. Neuronetwork approach in the early diagnosis of depression. Psychiatria Danubina. 2023 Oct 1;35(Suppl 2):77-85.

[7]. Cao X, Zhai L, Zhai P, Li F, He T, He L. Deep learning-based depression recognition through facial expression: A systematic review. Neurocomputing. 2025 Feb 11:129605.

[8]. Li Y, Zhang Y, Liu W, Feng X, Han Z, Chen C, Yan C. Multi-objective unlearning in recommender systems via preference guided pareto exploration. IEEE Transactions on Computing Services. 2025 Jul 31.

[9]. Xiao H, Yang Z, Liu T, Liu S, Huang X, Dai J. Deep learning for medical imaging super-resolution: A comprehensive review. Neurocomputing. 2025 Feb 17:129667.

[10]. Mondal A, Manikandan MS, Pachori RB. Automatic ECG signal quality assessment using convolutional neural networks and derivative ECG signal for false alarm reduction in wearable vital signs monitoring devices. Biomedical Signal Processing and Control. 2025 Oct 1; 108:107876.

[11]. Li X, Yi X, Lu L, Wang H, Zheng Y, Han M, Wang Q. TSFFM: Depression detection based on latent association of facial and body expressions. Computers in Biology and Medicine. 2024 Jan 1; 168:107805.

[12]. Xu J, Gunes H, Kusumam K, Valstar M, Song S. Two-stage temporal modelling framework for video-based depression recognition using graph representation. IEEE Transactions on Affective Computing. 2024 Jun 26;16(1):161-78.

[13]. Slonim DA, Yehezkel I, Paz A, Bar-Kalifa E, Wolff M, Dar A, Gilboa-Schechtman E. Facing change: using automated facial expression analysis to examine emotional flexibility in the treatment of depression. Administration and Policy in Mental Health and Mental Health Services Research. 2024 Jul;51(4):501-8.

[14]. Margoni M, Preziosa P, Rocca MA, Filippi M. Depressive symptoms, anxiety and cognitive impairment: emerging evidence in multiple sclerosis. Translational psychiatry. 2023 Jul 19;13(1):264.

[15]. Kalra A, Kang JK, Khanduja S, Menta AK, Ahmad SA, Liu O, Rodriguez E, Spann M, Hernandez AV, Brodie D, Whitman GJ. Long-term neuropsychiatric, neurocognitive, and functional outcomes of patients receiving ECMO: a systematic review and meta-analysis. Neurology. 2024 Feb 13;102(3):e208081.

[16]. Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, List M. Guiding questions to avoid data leakage in biological machine learning applications. Nature Methods. 2024 Aug;21(8):1444-53.

[17]. Rzecki K. Classification algorithm for person identification and gesture recognition based on hand gestures with small training sets. Sensors. 2020 Dec 18;20(24):7279.

[18]. Afzal T, Hipolito JL, Jin L. A Systematic Review of Misdiagnosis of Pediatric Bipolar Disorder: Assessments and Recommendations. Research on child and adolescent psychopathology. 2024 May;52(5):659-70.

[19]. Bartlome RI, Lee AJ. Facial attractiveness, but not facial masculinity, is used as a cue to paternal involvement in fathers. Adaptive Human Behavior and Physiology. 2023 Jun;9(2):182-97.

[20]. Li T, Zhang X, Wang C, Tian T, Chi J, Zeng M, Zhang X, Wang L, Li S. Facial expression analysis using convolutional neural network for drug-naive and chronic schizophrenia. Journal of Psychiatric Research. 2025 Jan 1; 181:225-36.

[21]. He L, Zhao J, Zhang J, Jiang J, Qi S, Wang Z, Wu D. LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. Applied Intelligence. 2025 Feb;55(3):195.

[22]. Xiao H, Liu X, Zhou T, Qian L, Huang X. Attention-driven contextual feature fusion network for facial videos-based depression recognition. Biomedical Signal Processing and Control. 2026 Mar 1;113:108806.

[23]. Fan H, Zhang X, Xu Y, Fang J, Zhang S, Zhao X, Yu J. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating

video, audio and remote photoplethysmography signals. Information Fusion. 2024 Apr 1; 104:102161

[24]. De Melo WC, Granger E, Lopez MB. Facial expression analysis using decomposed multiscale spatiotemporal networks. Expert Systems with Applications. 2024 Feb 1; 236:121276.

[25]. Niu M, Li Y, Tao J, Zhou X, Schuller BW. Depressionmlp: A multi-layer perceptron architecture for automatic depression level prediction via facial keypoints and action units. IEEE Transactions on Circuits and Systems for Video Technology. 2024 Mar 27;34(9):8924-38.

[26]. Zhang L, Zhao J, He L, Jia J, Meng X. An improved global–local fusion network for depression detection telemedicine framework. IEEE Internet of Things Journal. 2023 Jun 7;10(22):20230-40

[27]. Lin C, Lee SH, Huang CM, Chen GY, Chang W, Liu HL, Ng SH, Lee TM, Wu SC. Automatic diagnosis of late-life depression by 3D convolutional neural networks and cross-sample Entropy analysis from resting-state fMRI. Brain imaging and behavior. 2023 Feb;17(1):125-35.

[28]. Pan Y, Shang Y, Shao Z, Liu T, Guo G, Ding H. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. IEEE Transactions on Affective Computing. 2023 Jul 17