Three-Stream Region-Aware Residual Attention Network for Facial Depression Recognition*

Muhammad Turyalai Khan[1,*], Faisal Shafait[2] and Pradorn Sureephong[3]

[1]Guangdong CAS Cogniser, Information Technology, Co. Ltd., Guangzhou, China
[2]National University of Sciences and Technology, Islamabad, Pakistan
[3]College of Art Media and Technology, Chiang Mai University, Thailand

Abstract

Facial expressions provide critical cues for understanding emotional and mental states, including depression. However, existing deep learning approaches often lack mechanisms to emphasize subtle yet informative features from distinct facial regions. This work introduces TS-RAN (Three-Stream Region-Aware Residual Attention Network), a novel architecture designed for facial depression recognition. TS-RAN extracts and fuses global and local features from the face, eyes, and mouth using three parallel customized residual branches, each integrated with a coordinate attention mechanism to enhance spatial feature learning. The fused representation enables a comprehensive and discriminative understanding of depressive facial cues. Experiments are conducted on the AVEC2014 (Audio-Visual Emotion Challenge 2014) and self-collected CZ2024 (Changzhou No. 2 People's Hospital) datasets. TS-RAN achieves MAE/RMSE (mean absolute error/root mean square error) of 8.04/9.65 and 6.84/8.77 on the respective datasets, demonstrating competitive performance compared to existing methods. These results highlight its potential in medical and affective computing applications.

Keywords

Depression recognition, facial expression recognition, three-stream region-aware residual attention network

1. Introduction

Depression is a prevalent and debilitating mental disorder, affecting over 300 million people globally and projected to become the second leading cause of disability by 2030 [1]. Timely and accurate detection is essential for effective treatment, yet clinical diagnosis remains largely dependent on subjective assessments, often leading to delays or inconsistencies [2]. With the increasing availability of visual data and advances in artificial intelligence, facial expression analysis has emerged as a promising tool for supporting objective mental health evaluation [3, 4].

Recent advances in machine learning and deep learning have enabled robust modeling across domains such as speech analysis, biomedical signal processing, and computer vision [5, 6, 7]. In the context of depression recognition, diverse physiological and behavioral modalities have been explored, including electro-encephalography [8], spectroscopy [9], brain imaging [10], and eye movement tracking [11]. Among these, facial expressions are particularly advantageous due to their non-intrusive acquisition and strong correlation with emotional and cognitive states associated with depression [12, 13].

Deep learning-based facial expression analysis has emerged as a promising approach for automated depression recognition. Early models applied pre-trained convolutional neural networks (CNNs), such as visual geometry group (VGG) and residual network (ResNet), to extract global facial features, sometimes incorporating temporal dynamics using optical flow or frame differencing [14, 15]. More recent approaches emphasize regional analysis, dividing the face into patches or focusing on areas like the eyes and mouth, combined with attention mechanisms to highlight diagnostically relevant regions [16, 17, 18]. Multistream and spatiotemporal networks have also been proposed to integrate

4DMR@IJCAI25: International IJCAI Workshop on 1st Challenge and Workshop for 4D Micro-Expression Recognition for Mind Reading, August 29, 2025, Guangzhou, China.

*

*Corresponding author.
khan@cogniser.cn (M. T. Khan); faisal.shafait@seecs.edu.pk (F. Shafait); pradorn.s@cmu.ac.th (P. Sureephong)
0000-0002-1858-2710 (M. T. Khan); 0000-0002-0922-0566 (F. Shafait)

local and global cues [19, 20]. Despite these advancements, many methods lack the ability to model inter-region dependencies and capturing fine-grained cues that are critical for recognizing subtle depressive expressions. Additionally, traditional architectures rarely leverage the complementary nature of features derived from different facial regions, and often overlook the importance of robust feature fusion strategies[21].

To address these limitations, this work introduces the Three-Stream Region-Aware Residual Attention Network (TS-RAN) for facial depression recognition. TS-RAN consists of three parallel residual branches, each customized to extract features from the face, eyes, and mouth, respectively. Each stream incorporates a coordinate attention mechanism to enhance spatial feature learning and selectively emphasize important cues [22]. The resulting global and local features are fused into a unified representation for comprehensive and discriminative depression analysis.

The proposed method is evaluated on two benchmark datasets: Audio-Visual Emotion Challenge 2014 (AVEC2014) [23], a public dataset for affective computing, and Changzhou No. 2 People's Hospital 2024 (CZ2024), a clinical dataset collected from Changzhou No. 2 People's Hospital. Experimental results show that TS-RAN achieves competitive performance in terms of mean absolute error (MAE) and root mean square error (RMSE), outperforming several state-of-the-art methods. These findings underscore the potential of region-aware modeling in advancing automated mental health diagnostics.

The main contributions of this work are summarized as follows:

- A novel three-stream region-aware attention network (TS-RAN) is introduced to extract and fuse facial features from the face, eyes, and mouth regions for depression recognition.
- Each residual stream is enhanced with coordinate attention to strengthen spatial encoding and highlight subtle but informative facial cues.
- Comprehensive experiments are conducted on both a public dataset (AVEC2014) and a newly collected clinical dataset (CZ2024), demonstrating the model's robustness across controlled and real-world conditions.

The remainder of the paper is organized as follows. Section 2 presents the proposed methodology, including preprocessing, feature extraction, and fusion. Section 3 describes the datasets, evaluation metrics, and training setup. Section 4 reports experimental results, ablation analysis, and comparative evaluations. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Methodology

The overall architecture of the proposed Three-Stream Region-Aware Residual Attention Network (TS-RAN) is illustrated in Figure 1. Given an input video, frames are sampled at regular intervals and passed through a facial landmark detection module to localize key facial regions. Three separate branches are constructed to process the full face, eyes, and mouth independently. Each branch uses a customized residual network (ResNet) backbone integrated with a coordinate attention (CA) mechanism to extract region-specific features. These features are subsequently concatenated and passed through a fully connected layer to generate a continuous depression score.

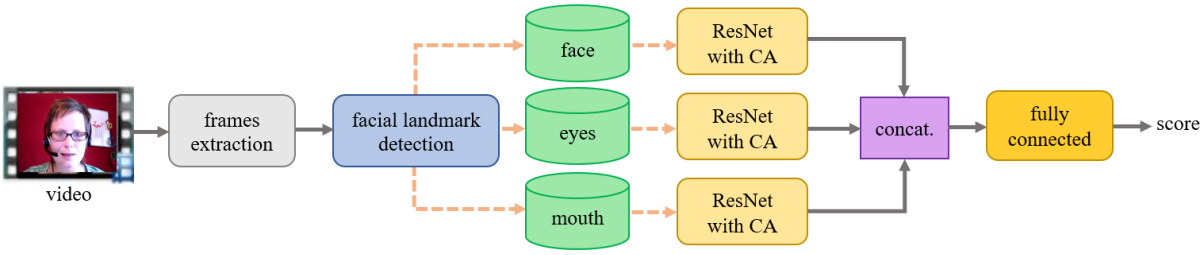


Figure 1: Block diagram of the proposed three-stream region-aware residual attention network (TS-RAN). Abbreviations: residual network - ResNet, coordinate attention - CA, and concatenation - concat.

2.1. Preprocessing

Video frames are extracted at regular intervals of 5 seconds using the open-source computer vision library (OpenCV) to reduce temporal redundancy and computational load. Each sampled frame is processed by a multi-task cascaded convolutional network (MTCNN) to detect facial landmarks, including key points around the eyes and mouth. Based on the landmark positions, three specific facial regions are cropped using predefined bounding boxes: the full face, the eyes region, and the mouth region. Each cropped region is extracted with fixed spatial dimensions and three RGB channels, resulting in face, eyes, and mouth crops of size 3×256×256, 3×96×192, and 3×96×128, respectively, before being passed into their corresponding feature extraction branches.

2.2. Region-Aware Feature Extraction

Each of the three facial regions is processed by a dedicated ResNet-50 architecture, modified to act as a pure feature extractor. The classification head is removed and replaced with an identity mapping, allowing the network to output high-dimensional feature embeddings without any class-specific bias. Each ResNet is augmented with a coordinate attention module, enhancing its ability to capture spatial dependencies across both horizontal and vertical dimensions.

An overview of the modified feature extraction module is shown in Figure 2. This architecture extends the standard ResNet-50 by integrating coordinate attention into each residual block, specifically after the final convolutional layer. This modification enables the network to capture both spatial structure and contextual dependencies across the height and width axes, thereby enhancing the representation of depression-relevant facial cues in each region.

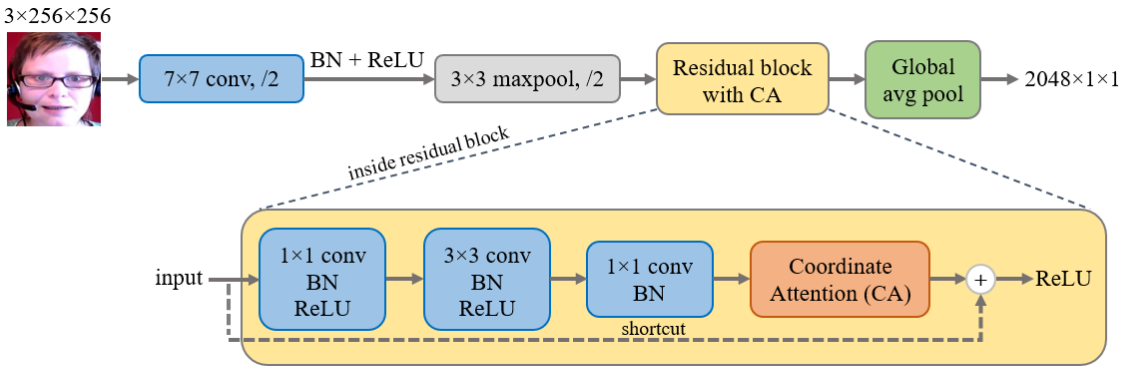


Figure 2: Modified ResNet-50 architecture used as the feature extraction module, where coordinate attention (CA) is integrated after the final convolution within each residual block to enhance spatial and contextual representation of facial regions.

Coordinate attention improves upon traditional channel-only attention mechanisms by preserving positional information [22]. Instead of applying global average pooling over spatial dimensions, it encodes directional context separately along the height and width axes. Let $F \in \mathbb{R}^{C\times H\times W}$ denote the input feature map, where C, H, and W are the number of channels, height, and width, respectively. Coordinate attention computes:

$$F_h = \frac{1}{W}\sum_{i=1}^{W} F(:,:,i), \quad F_w = \frac{1}{H}\sum_{j=1}^{H} F(:,j,:) \tag{1}$$

These aggregated features F_h and F_w are passed through shared transformations to produce attention maps, which are then applied back to the original feature map. This allows the network to focus on salient spatial locations relevant to depression cues.

Each stream outputs a 2048-dimensional feature vector corresponding to its respective region. The use of coordinate attention ensures that both global semantics and local details are preserved, improving the discriminative capacity of the extracted features.

2.3. Feature Fusion and Regression

The high-dimensional feature embeddings generated by the face, eyes, and mouth streams, each of size 2048, are concatenated to form a unified 6144-dimensional feature vector. This fused representation integrates complementary spatial information from global and local facial regions, capturing both holistic appearance and subtle region-specific variations associated with depressive states. By combining features from different facial regions, the model benefits from enhanced context awareness and improved discriminative power.

The concatenated feature vector is passed through a fully connected regression head, which maps the fused representation to a scalar depression severity score. This final output is designed to reflect the continuous nature of depression levels, aligning with clinical assessment standards. The regression head consists of dense layers with non-linear activation functions, followed by a final linear layer that outputs a single value.

The model is trained to minimize the mean squared error (MSE) between the predicted and ground truth depression scores, promoting accurate regression. Formally, the loss is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

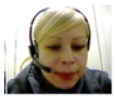
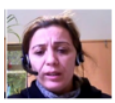
where y_i is the ground truth score, $\hat{y}_i$ is the predicted score for the i-th sample, and N is the total number of training samples. This loss encourages the network to produce continuous estimates that closely match clinical annotations.

3. Experiments

3.1. Datasets

The proposed TS-RAN architecture is evaluated on two datasets: the publicly available AVEC2014 and the clinical CZ2024 dataset collected from Changzhou No. 2 People's Hospital. The AVEC2014 was released as part of the Audio-Visual Emotion Challenge, and contains 300 videos divided equally across two tasks: Freeform and Northwind [23]. In the Freeform task, participants respond to emotionally driven prompts such as recalling personal experiences, while the Northwind task involves reading a neutral passage aloud. Each task contributes 150 videos, with 50 allocated to each of the training, validation, and test sets. Video durations range from 6 to 248 seconds, recorded at 30 frames per second. Ground truth depression scores are provided based on the Beck Depression Inventory-II (BDI-II), which classifies severity into minimal (0–13), mild (14–19), moderate (20–28), and severe (29–63) [24]. Dataset statistics are summarized in Table 1.

Table 1
Examples of different severity levels from the AVEC2014 dataset.

	BDI-II	Severity level
	0	minimum
	12	mild
	24	moderate
	37	severe

CZ2024 is a clinically collected dataset comprising 327 videos of patients undergoing depression assessment at Changzhou No. 2 People's Hospital in Jiangsu, China. Subjects span a diverse age range of 14 to 73 years and include both male and female participants. Depression severity is annotated according to the Hamilton Depression Rating Scale (HAMD), which classifies depression into minimal (0–7), mild (8–19), moderate (20–34), and severe (35). The dataset is divided into 193 training, 95 validation, and 39 test videos. All recordings were captured in natural indoor settings, reflecting variations in lighting, pose, and expression. Key characteristics of CZ2024 are presented in Table 2.

Table 2
Examples of different severity levels from the CZ2024 dataset.

	HAMD	Severity level
	5	minimal
	16	mild
	25	moderate
	37	severe

3.2. Evaluation Metrics

To assess the regression performance of the proposed TS-RAN model in predicting depression severity, two commonly used metrics are employed: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) [12]. These metrics are widely adopted in affective computing and regression tasks due to their ability to capture both prediction accuracy and consistency.

RMSE evaluates the average squared difference between predicted and actual scores, placing greater emphasis on larger errors. MAE, in contrast, measures the average absolute difference and provides a more interpretable measure of typical prediction error. Together, they offer a robust evaluation of the model's precision in estimating continuous depression severity scores. Mathematically, RMSE and MAE are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4}$$

where N is the total number of samples, y_i is the ground truth score, and $\hat{y}_i$ is the predicted score.

3.3. Training Details

The training of TS-RAN was performed using the Python programming language, and its configuration is summarized in Table 3. The model was optimized using the RMSprop algorithm with a smoothing factor (alpha) of 0.9 to stabilize gradient updates. A learning rate of 0.0005 was selected to ensure a balance between convergence speed and stability [25, 26], while mini-batches of size 24 were used to maintain efficiency without compromising model generalization.

To mitigate overfitting, a weight decay of 0.05 was applied as regularization. Additionally, a stepwise learning rate scheduler was used, reducing the learning rate by a factor of 0.5 every 4 epochs. This

gradual adjustment strategy facilitates more refined convergence during later training stages, especially on limited data.

Table 3
Training hyperparameters used for optimizing the TS-RAN model.

Hyperparameter	Value
Optimizer	RMSprop
Alpha	0.9
Learning Rate	0.0005
Batch Size	24
Weight Decay	0.05
Learning Rate Decay	Every 4 epochs
Gamma	0.5
Loss	Mean squared error

4. Results and Discussion

This section presents the experimental results of the proposed TS-RAN model for facial depression recognition. Quantitative evaluations are reported on both the AVEC2014 and CZ2024 datasets using standard regression metrics. Ablation studies are conducted to examine the contribution of the coordinate attention mechanism and the impact of multi-region feature fusion in enhancing model performance. In addition, a comparative analysis with existing state-of-the-art methods is provided to demonstrate the effectiveness of the proposed architecture.

4.1. Quantitative Results

The performance of the proposed TS-RAN model was quantitatively evaluated on both the AVEC2014 and CZ2024 datasets using mean absolute error (MAE) and root mean square error (RMSE) as evaluation metrics. The results are summarized in Figure 3.

On the AVEC2014 dataset, TS-RAN achieved an MAE of 8.04 and an RMSE of 9.65, reflecting reliable prediction accuracy across diverse video samples and depression severity levels. On the CZ2024 clinical dataset, the model attained even lower error rates, with an MAE of 6.84 and an RMSE of 8.77, indicating improved generalization to real-world, clinically acquired facial recordings.

These results demonstrate the model's robustness in handling both benchmark and clinical data, highlighting its capacity to extract discriminative features from global and local facial regions for accurate depression severity estimation.

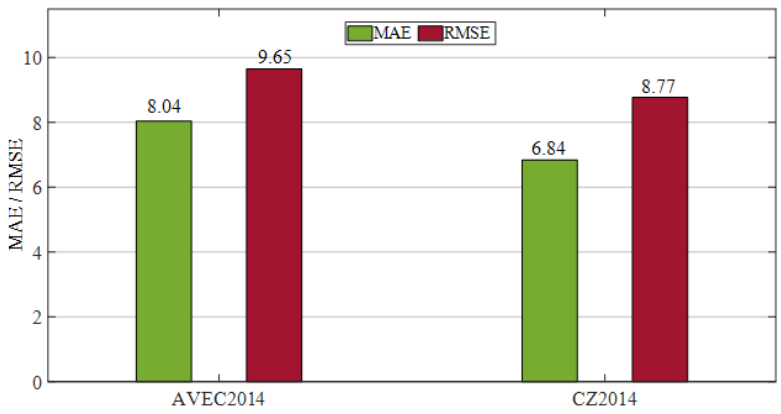


Figure 3: Regression performance of the TS-RAN on the AVEC2014 and CZ2024 datasets.

4.2. Ablation Study

To evaluate the contribution of individual components within the TS-RAN architecture, ablation studies were conducted on both the AVEC2014 and CZ2024 datasets. The analysis focused on two critical aspects: the role of coordinate attention and the impact of local facial features (eyes and mouth). The results are presented in Table 4.

When the coordinate attention module was removed, the performance deteriorated on both datasets, confirming its importance in enhancing spatial feature representation. Specifically, the MAE increased from 8.04 to 8.23 on AVEC2014 and from 6.84 to 7.61 on CZ2024, indicating reduced accuracy in the absence of attention-guided feature enhancement. The exclusion of local features (eyes or mouth) also led to performance drops, though less severe than removing attention. Notably, removing both eyes and mouth features while retaining only the global face stream resulted in higher MAE and RMSE on both datasets, suggesting that local region cues provide complementary information critical for accurate depression prediction.

The full TS-RAN model, incorporating coordinate attention and all three facial regions, consistently outperformed all ablated variants, confirming the effectiveness of both spatial attention and multi-region feature fusion in facial depression recognition.

Table 4
Ablation study on the AVEC2014 and CZ2024 datasets showing the impact of removing key components from TS-RAN.

Removed component(s)	AVEC2014		CZ2024	
	MAE	RMSE	MAE	RMSE
Coordinate attention	8.23	9.94	7.61	9.55
Eyes features	8.24	9.77	6.98	8.88
Mouth features	8.28	9.99	7.03	8.96
Eyes and mouth features	8.22	9.80	7.03	9.04
None (full model – TS-RAN)	**8.04**	**9.65**	**6.84**	**8.77**

4.3. Comparison with State-of-the-Art

To assess the effectiveness of the proposed TS-RAN model, performance comparisons were made against several existing state-of-the-art methods reported on the AVEC2014 and CZ2024 datasets. The results, shown in Table 5 and Table 6, include models that utilize various combinations of handcrafted features and deep learning architectures.

On the AVEC2014 dataset, TS-RAN achieves competitive results with an MAE of 8.04 and an RMSE of 9.65, outperforming earlier methods such as [23] and [27]. It also performed better than recent Transformer-based method [28]. Furthermore, the performance of TS-RAN is close to other deep models including [14, 29, 12]. Although these methods report slightly lower errors, TS-RAN maintains robustness through its multi-region structure without requiring heavy temporal modeling.

On the CZ2024 dataset, TS-RAN outperforms a hybrid Vision Transformed-based method [30] by a noticeable margin, achieving 6.84 MAE and 8.77 RMSE compared to 7.46 MAE and 9.15 RMSE, indicating its superior ability to generalize across clinical settings with varied facial presentations.

5. Conclusion and Future Work

This paper presented TS-RAN, a three-stream region-aware residual attention network designed for facial depression recognition. The model extracts and fuses global and local facial features from the full face, eyes, and mouth using customized residual networks enhanced with coordinate attention. Experimental results on the AVEC2014 and CZ2024 datasets demonstrated that TS-RAN delivers competitive performance, outperforming baseline approaches and showing strong generalization to

Table 5
Comparison with state-of-the-art methods on the AVEC2014 dataset.

Reference	MAE	RMSE
Baseline [23]	10.03	12.56
FDHH [27]	8.07	10.28
Deep CNN [14]	7.47	9.55
CNN-LSTM [29]	6.86	8.78
3D-CNN [19]	6.78	**8.42**
Handcrafted and deep features [12]	**6.57**	8.49
Transformer [28]	8.54	10.61
TS-RAN (ours)	8.04	9.65

Table 6
Comparison with a baseline method on the CZ2024 dataset.

Reference	MAE	RMSE
Hybrid Vision Transformer [30]	7.46	9.15
TS-RAN (ours)	**6.84**	**8.77**

clinical data. Ablation studies validated the contribution of coordinate attention and multi region feature fusion in improving recognition accuracy.

Future work will explore the integration of temporal modeling using three dimensional convolution or transformer based encoders to better capture subtle changes and micro expressions. In addition, combining visual features with other modalities such as audio or physiological signals may further enhance the robustness of depression estimation in real-world applications.

6. Acknowledgments

The authors would like to thank the Guangdong CAS Cogniser, NUST, and Chiang Mai University for providing computational resources for this research. This work was supported by the HORIZON-MSCA-SE-2022 project ACMod (grant 101130271). All authors declare that there is no conflict of interest. The CZ2024 dataset is confidential and cannot be shared due to ethical restrictions.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030, PLOS Meicine 3 (2006) e442. doi:10.1371/journal.pmed.0030442.

[2] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, IEEE Transactions on Affective Computing 11 (2018). doi:10.1109/TAFFC.2018.2828819.

[3] R. Wang, J. Huang, J. Zhang, X. Liu, X. Zhang, Z. Liu, P. Zhao, S. Chen, X. Sun, Facialpulse: An efficient rnn-based depression detection via temporal facial landmarks, in: MM '24: Proceedings of the 32nd ACM International Conference on Multimedia, ACM, 2024, pp. 311–320. doi:10.1145/3664647.3681546.

[4] C. Shi, C. Tan, L. Wang, A facial expression recognition method based on a multibranch cross-connection convolutional neural network, IEEE Access 9 (2021). doi:10.1109/ACCESS.2021.3063493.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Communication 71 (2015). doi:10.1016/j.specom.2015.03.004.

[6] M. T. Khan, A. Z. Sha'ameri, M. M. A. Zabidi, Classification of fhss signals in a multi-signal environment by artificial neural network, International Journal of Computing and Digital Systems 11 (2022). doi:10.12785/ijcds/110163.

[7] M. T. Khan, A. Z. Sha'ameri, M. M. A. Zabidi, C. C. Chia, Fhss signals classification by linear discriminant in a multi-signal environment, in: F. Thakkar, G. Saha, C. Shahnaz, Y.-C. Hu (Eds.), Proceedings of the International e-Conference on Intelligent Systems and Signal Processing, volume 1370 of *Advances in Intelligent Systems and Computing*, Springer, Singapore, 2022, pp. 143–155. doi:10.1007/978-981-16-2123-9_11.

[8] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, D. P. Subha, Automated eeg-based screening of depression using deep convolutional neural network, Computer Methods and Programs in Biomedicine 161 (2018). doi:10.1016/j.cmpb.2018.04.012.

[9] S. F. Husaina, T.-B. Tang, R. Yu, W. W. Tam, B. Tran, T. T. Quek, S.-H. Hwang, C. W. Chang, C. S. Ho, R. C. Ho, Cortical haemodynamic response measured by functional near infrared spectroscopy during a verbal fluency task in patients with major depression and borderline personality disorder, EBioMedicine 51 (2020). doi:10.1016/j.ebiom.2019.11.047.

[10] K. M. Han, D. D. Berardis, M. Fornaro, Y. K. Kim, Differentiating between bipolar and unipolar depression in functional and structural mri studies, Progress in Neuro-Psychopharmacology and Biological Psychiatry 91 (2019). doi:10.1016/j.pnpbp.2018.03.022.

[11] Y. Lin, H. Ma, Z. Pan, R. Wang, Depression detection by combining eye movement with image semantics, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, USA, 2021, pp. 269–273. doi:10.1109/ICIP42928.2021.9506702.

[12] C. Álvarez Casado, M. L. Cañellas, M. B. López, Depression recognition using remote photoplethysmography from facial videos, IEEE Transactions on Affective Computing 14 (2023). doi:10.1109/TAFFC.2023.3238641.

[13] X. Yuan, Z. Liu, Q. Chen, G. Li, Z. Ding, Z. Shangguan, B. Hu, Combining informative regions and clips for detecting depression from facial expressions, Cognitive Computation 15 (2023). doi:10.1007/s12559-023-10157-0.

[14] Y. Zhu, Y. Shang, Z. Shao, G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Transactions on Affective Computing 9 (2018). doi:10.1109/TAFFC.2017.2650899.

[15] W. C. de Melo, E. Granger, M. B. Lopez, Encoding temporal information for automatic depression recognition from facial analysis, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1080–1084. doi:10.1109/ICASSP40776.2020.9054375.

[16] L. He, J. C.-W. Chan, Z. Wang, Automatic depression recognition using cnn with attention mechanism from videos, Neurocomputing 422 (2021). doi:10.1016/j.neucom.2020.10.015.

[17] M. Niu, Z. Zhao, J. Tao, Y. Li, B. W. Schuller, Dual attention and element recalibration networks for automatic depression level prediction, IEEE Transactions on Affective Computing 14 (2023). doi:10.1109/TAFFC.2022.3177737.

[18] M. T. Khan, M. Imran, M. Kanwal, Mcnn: Multi-channel neural network with channel-wise attention for facial expression-based depression recognition, Multimedia Tools and Applications (2025). doi:10.1007/s11042-025-20962-4.

[19] L. He, C. Guo, P. Tiwari, H. M. Pandey, W. Dang, Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence, International Journal of Intelligent Systems 37 (2022). doi:10.1002/int.22426.

[20] W. C. de Melo, E. Granger, A. Hadid, Combining global and local convolutional 3d networks

for detecting depression from facial expressions, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8. doi:10.1109/FG.2019.8756568.

[21] M. T. Khan, U. U. Sheikh, A hybrid convolutional neural network with fusion of handcrafted and deep features for fhss signals classification, Expert Systems with Applications 225 (2023). doi:10.1016/j.eswa.2023.120153.

[22] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2021, pp. 13713–13722. doi:10.1109/CVPR46437.2021.01350.

[23] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, Avec 2014: 3d dimensional affect and depression recognition challenge, in: Proceedings of the 4th international workshop on audio/visual emotion challenge, 2014, ACM, 2014, pp. 3–10. doi:10.1145/2661806.2661807.

[24] B. Hajduska-Dér, G. Kiss, D. Sztahó, K. Vicsi, L. Simon, The applicability of the beck depression inventory and hamilton depression scale in the automatic recognition of depression based on speech signal processing, Frontiers in Psychiatry 13 (2022). doi:10.3389/fpsyt.2022.879896.

[25] M. T. Khan, A modified convolutional neural network with rectangular filters for frequency-hopping spread spectrum signals, Applied Soft Computing Journal 150 (2024). doi:10.1016/j.asoc.2023.111036.

[26] M. T. Khan, Machine Learning Classification of Frequency-Hopping Spread Spectrum Signals in a Multi-Signal Environment, Ph.D. thesis, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, 2023.

[27] A. Jan, H. Meng, Y. F. B. A. Gaus, F. Zhang, Artificial intelligent system for automatic depression level analysis through visual and vocal expressions, IEEE Transactions on Cognitive and Developmental Systems 10 (2017). doi:10.1109/TCDS.2017.2721552.

[28] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, J. Yu, Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals, Information Fusion 104 (2024). doi:10.1016/j.inffus.2023.102161.

[29] M. A. Uddin, J. B. Joolee, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm, IEEE Transactions on Affective Computing 13 (2022). doi:10.1109/TAFFC.2020.2970418.

[30] Z. Jiang, X. Gao, Y. Cao, Y. Zhang, G. Dong, Y. Chen, X. Zhu, Q. Zhang, R. Bi, K. Wang, Dnet: A depression recognition network combining residual network and vision transformer (2024). doi:10.21203/rs.3.rs-4465101/v1.