Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.39369

Multimodal Assessment of Emotional Expression in Individuals with Depression and Emotional Alexithymia Through Embodied Performing Arts Intervention

Zelin Tang

School of Education, Communication & Society, King's College London, London, UK
wellington589125@gmail.com

Abstract. Alexithymic depression is marked by difficulties in identifying emotions, poor verbalization, and lack of nonverbal expression, while measurement of the latter with a scale has low utility for capturing the dynamics of intervention-related change. The effects of an eight-week performing arts program on behavior were analyzed by recording synchronously videoed facial expressions, speech, full-body poses, and heart rate variability in 60 adults assigned randomly to groups of intervention and controls. Visual, acoustic, postural, and physiological features extracted with MediaPipe, openSMILE, and RR-interval analysis were used for modeling with Elastic Net, XGBoost, and late multimodal fusion. Late fusion resulted in classification accuracy of 0.883, F1 of 0.879, and ROC-AUC of 0.918; regression models produced MAE of 2.12 (R2 of 0.62) for PHQ-9 and 3.47 (R2 of 0.58) for TAS-20. The intervention group exhibited gains of 0.63 in movement amplitude, 0.49 in postural openness, 0.46 in facial activity, 0.41 in F0 range, and 0.37 in RMSSD, which were superior to changes in the control group across all modalities. This combination of behavioral, vocal, and physiological changes suggests that performing arts could be evaluated not only by scales but also by objectively recorded indicators.

Keywords: Multimodal Assessment, Depression, Alexithymia, Embodied Performing Arts, Affective Computing

1. Introduction

It is common for depression and alexithymia to have similar difficulties in recognizing feelings, distinguishing emotions from physical sensations, and expressing one's internal states, resulting in the existence of a clinically significant disconnect between inner suffering and external manifestations thereof [1]. Additionally, facet-based research reveals that difficulty identifying feelings is significantly linked with depressive symptoms in both positive and negative emotional categories [2]. The current approach of assessment is limited to questionnaire-based methods and interviews; however, emotional changes can also be seen through facial expressions, prosodic speech, body movement, and autonomic regulation. Embodied performing arts offer numerous opportunities for engaging in the processes of movement, vocal production, imitative actions, and social interaction that result in the elicitation of expressive behavior. A multimodal affective

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.39369

computing system can translate such behavioral output into numerical variables and establish whether there is a link between behavioral change and clinical improvement. The analysis will include an eight-week intervention using facial videos, speech, body pose (33-points), heart rate variability, PHQ-9, and TAS-20 measures.

2. Literature review

2.1. Clinical emotional phenotypes

The clinical emotional phenotype forms the observable foundation for computer-based analysis. There is evidence of depressed patients having impairment in the recognition of negative emotions from faces, and alexithymia, the inability to describe one's feelings, correlates with poor recognition ability [3]. The vocal phenotype is complementary to the above as depression alters the pitch, loudness, timing, and lexical content of speech [4]. Different aspects of emotional functioning are captured by facial and vocal phenotypes, with the former capturing affective activation that can be seen and the latter motor control of the voice and prosody. Challenges posed by alexithymia include the coexistence of self-descriptive difficulty with objectively identifiable response. PHQ-9 and TAS-20 measures act as clinical markers and not substitutes for behavior measurement.

2.2. Embodied performance regulation

Emotional performance is centered on the connection between physical arousal and emotion expression. The somatosensory response in perception of facial emotions varies according to the level of alexithymia, thus justifying the importance of bodily processes in relation to problems in emotion recognition [5]. Frameworks of artistic emotion regulation show that a well-organized creative process increases emotional sensitivity, facilitates the use of regulation strategies, and offers multiple opportunities to monitor changes related to depression [6]. The dramatic approach includes improvisation, storytelling, and role play that have demonstrated their capacity to help relieve depressive symptoms [7]. Such processes can be used with mirror movements, rhythm imitation, vocal expression of emotions, and role improvisation since each of the tasks allows externalizing an internal state through observable action.

2.3. Multimodal computational recognition

The multimodal approach to computational recognition deals with the fact that the expressions indicating depression are scattered in channels but not in one biomarker alone. Facial and body expressions have been seen to have a synergistic effect in recognizing depression with temporal fusion performing better than individual expressions [8]. Recent studies on emotion in relation to stimulus show that facial and body expressions combined at the level of decision yield superior recognition performance to that of individual modes while keeping distinct behavioral information [9]. Handcrafted temporal statistics and regularization methods are feasible alternatives for a modest clinical group where end-to-end models require too many parameters. Independent validation is needed since the windows from the same subject have consistent identity and behavioral features.

DOI: 10.54254/2753-7064/2026.39369

3. Experimental methods

3.1. Randomized intervention design

Enrollment includes 60 adults between the ages of 18 and 45 with depressive disorders as well as TAS-20 screeners' cut-off scores, after which they are randomized to intervention and control groups at a ratio of 1:1. Pre-intervention assessment involves taking PHQ-9, TAS-20 scores, facial videos, speech, full body poses, and RR interval measurements during standardized picture description and improvisational movement tasks. Intervention includes performing-arts sessions twice a week for 8 weeks with a total of 16 sessions focusing on movement mimicking, rhythmical mimicry, emotional vocalization, and improvisational role-playing tasks. Control group members will continue undergoing their standard treatment with no performing-arts sessions during this period of time. Post-intervention assessment will involve the same tasks with the same sensor measures. Every observation will be identified by the ID number, group, session number, task name, and pre/post indicator. Behavioral and physiological data will have synchronized timestamps. Randomization will take place prior to the intervention assignment while the modeling process will use participant identifiers to avoid having windows from one individual appearing in both training and test datasets.

3.2. Multimodal feature extraction

Processing for facial and posture stream is performed using MediaPipe where facial landmarks and 33 body key points are extracted and openSMILE is used to extract F0, Energy, Speaking Rate, Jitter, and Shimmer from speech stream. Chest sensor signal RR intervals are filtered to detect any ectopic or missing beats before calculating heart rate variability [10]. All streams are aligned by timestamps, corrupted segments are removed, and valid sequences are divided into non-overlapping five-second windows. Mean keypoint velocity quantifies movement intensity as shown in Eq. (1):

$$v_t = \frac{1}{K} \sum_{k=1}^{K} \frac{\| p_{k,t} - p_{k,t-1} \|_2}{\Delta t} \tag{1}$$

Where K is the number of tracked keypoints, $p_{k,t}$ is the position of keypoint k at time t, and Δt is the frame interval. Cardiac variability is represented by RMSSD as shown in Eq. (2):

$$RMSSD = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (RR_{i+1} - RR_i)^2} \tag{2}$$

Where RR_i denotes the i-th normal beat interval and N is the number of valid intervals. Features are finally normalized within participants before aggregation.

3.3. Multimodal fusion modeling

The five-second window features are described by mean, standard deviation, and the rate of change with time at the level of participant. The Elastic Net is used for the regularization of a linear model for correlated multimodal variables, whereas XGBoost identifies the non-linear relationships between multimodal features of face, voice, posture, and physiological data [11]. Elastic Net estimates coefficients using the objective in Eq. (3):

DOI: 10.54254/2753-7064/2026.39369

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\{\frac{1}{2n}\|y-X\beta\|_2^2+\lambda[\alpha\|\beta\|_1+\frac{1-\alpha}{2}\|\beta\|_2^2]\} \quad (3)$$

Where X is the feature matrix, y is the target, λ controls regularization, and α balances L1 and L2 penalties. Modality-specific probabilities are then combined through late fusion as shown in Eq. (4):

$$P(c|x)=\sum_{m=1}^{M} w_m P_m(c|x_m),\ \sum_{m=1}^{M} w_m=1 \quad (4)$$

Where P_m is the probability from modality m and w_m is its validation-derived weight. Classification predicts pre/post intervention state, while regression predicts PHQ-9 and TAS-20 change. Participant-level five-fold cross-validation, permutation testing, modality ablation, and SHAP interpretation evaluate robustness and feature contribution.

4. Results

4.1. Model recognition performance

As seen from Table 1, multimodal integration outperformed all unimodal approaches under the participant-level five-fold cross-validation. The late-fusion approach demonstrated the best classification results in terms of accuracy equal to 0.883, F1 equal to 0.879, and ROC-AUC equal to 0.918 in comparison with facial ROC-AUC equal to 0.812, acoustic ROC-AUC equal to 0.798, postural ROC-AUC equal to 0.784, and physiological ROC-AUC equal to 0.754. Elastic Net and XGBoost multimodal approaches showed ROC-AUC equal to 0.865 and 0.891 correspondingly, reflecting the increasing performance due to the introduction of non-linear interactions and decision fusion. Regression analysis demonstrated the similar trend. The use of late fusion resulted in PHQ-9 MAE equal to 2.12 and R2 equal to 0.62, while TAS-20 prediction achieved MAE equal to 3.47 and R2 equal to 0.58. The ablation study showed the biggest ROC-AUC drop when facial features were removed, followed by acoustic and postural features.

Table 1. Classification and regression performance of different models

Model	Accuracy	F1	ROC-AUC	PHQ-9 MAE	PHQ-9 R2	TAS-20 MAE	TAS-20 R2
Facial	0.767	0.758	0.812	3.18	0.31	5.42	0.28
Acoustic	0.750	0.741	0.798	3.26	0.28	5.61	0.24
Postural	0.733	0.721	0.784	3.41	0.24	5.84	0.21
Physiological	0.700	0.688	0.754	3.63	0.18	6.12	0.15
Elastic Net multimodal	0.817	0.811	0.865	2.61	0.48	4.38	0.44
XGBoost multimodal	0.850	0.846	0.891	2.34	0.56	3.86	0.52
Late fusion	0.883	0.879	0.918	2.12	0.62	3.47	0.58

DOI: 10.54254/2753-7064/2026.39369

4.2. Emotional expression patterns

In terms of reduction in PHQ-9 from 15.8 to 9.9 and TAS-20 from 65.8 to 56.1, the intervention group had a better result than the control group, where PHQ-9 decreased from 15.6 to 13.7 and TAS-20 decreased from 65.4 to 62.9. As can be seen from Figure 1 below, in all aspects of the experiment mentioned above the gain was greater for the experimental group than for the control group. The mean gain for movement amplitude, postural openness, facial activity, F0 range, and RMSSD was 0.63, 0.49, 0.46, 0.41, and 0.37 standard-deviation units, respectively, and 0.11, 0.07, 0.09, 0.04, and 0.06 for the control group, correspondingly. The group-by-time interaction was significant for all measures except RMSSD at $p < 0.01$, and the RMSSD was significant at $p = 0.021$. The highest divergence between groups was achieved in the measure of movement amplitude, as expected given the nature of the intervention.

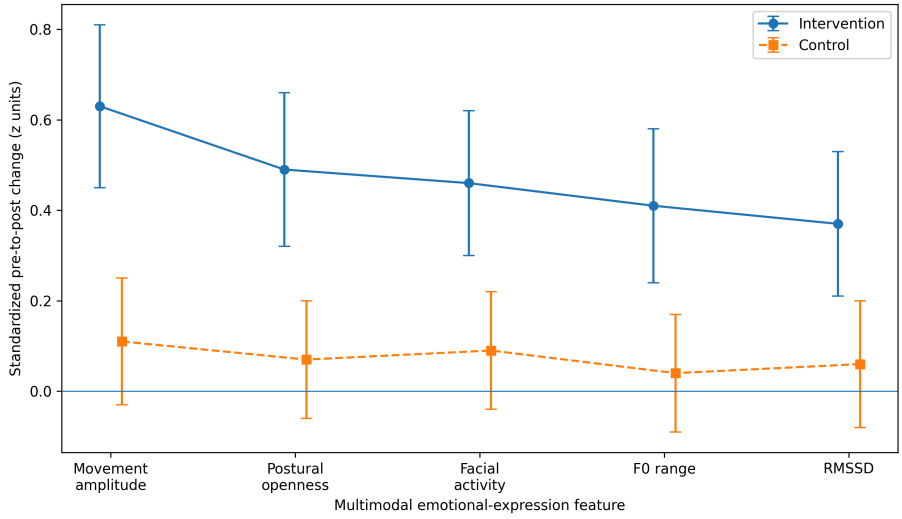

Figure 1. Standardized pre-to-post changes in multimodal emotional expression features

5. Discussion

Results indicate that performing arts may be embodied through an interplay of expressive modes instead of a single indicator of behavior. Larger differences in motoric activity and posturing make sense, since mirroring and improvising constantly force individuals to translate affective states into behavior, while voice tasks put flattened prosody into question. Facial and acoustic modalities were of greater importance for classification than HRV, which means that intervention-state discrimination was based mainly on the former, while autonomic processes yielded additional cues. The significance of the smaller difference in RMSSD values has its place since autonomic flexibility can be learned through different processes than those responsible for performance improvement. Correlation of decreased PHQ-9 and TAS-20 values with increased non-verbal expressiveness proves that the positive effects of alexithymic symptoms' alleviation cannot be evaluated only via questionnaires.

6. Conclusion

A performing arts intervention was assessed with synchronized facial expression, acoustical, postural, physiological and clinical features. An interpretable pipeline consisting of MediaPipe, openSMILE, HRV analysis, Elastic Net, XGBoost, and late fusion was developed for linking

DOI: 10.54254/2753-7064/2026.39369

behavioral intervention with PHQ-9 and TAS-20 results. A multimodal approach demonstrated higher classification and regression performance compared to unimodal solutions and positive effects of an intervention were identified in all aspects except for RMSSD. Major drawback consists in the moderate sample size and two-point testing that limits generalization in time. Future studies should include large multisite samples and additional tests on several points to assess whether fusion with personal weights can increase the model sensitivity.

References

[1] Kieraité, M., Bättig, J. J., Novoselac, A., Noboa, V., Seifritz, E., Rufer, M., Egger, S. T., & Weidt, S. (2024). "Our similarities are different": The relationship between alexithymia and depression. *Psychiatry Research, 340*, 116099.

[2] Preece, D. A., Mehta, A., Petrova, K., Sikka, P., Pemberton, E., & Gross, J. J. (2024). Alexithymia profiles and depression, anxiety, and stress. *Journal of Affective Disorders, 357*, 116-125.

[3] Gul, H., Tas Torun, Y., Cakmak, F. H., & Gul, A. (2023). Facial emotion recognition in adolescent depression: The role of childhood traumas, emotion regulation difficulties, alexithymia and empathy. *Indian Journal of Psychiatry, 65*(4), 443-452.

[4] Menne, F., Dörr, F., Schräder, J., Tröger, J., Habel, U., König, A., & Wagels, L. (2024). The voice of depression: Speech features as biomarkers for major depressive disorder. *BMC Psychiatry, 24*, 794.

[5] Arslanova, I., Meletaki, V., Calvo-Merino, B., & Forster, B. (2023). Perception of facial expressions involves emotion specific somatosensory cortex activations which are shaped by alexithymia. *Cortex, 167*, 223-234.

[6] Lee, M., & Choi, H. (2023). Art-based emotion regulation in major depression: Framework for intervention. *The Arts in Psychotherapy, 83*, 102018.

[7] Sharkiya, S. H., Malak, M. Z., & Shehadeh, A. (2024). The effect of drama therapy on depressive symptoms and quality of life among older adults in residential care facilities: A systematic review. *The Arts in Psychotherapy, 91*, 102222.

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

[9] Liu, Y., Li, X., Wang, M., Bi, J., Lin, S., Wang, Q., Yu, Y., Ye, J., & Zheng, Y. (2025). Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli. *Journal of Affective Disorders, 381*, 44-54.

[10] Bylsma, L. M., DeMarree, K. G., McMahon, T. P., Park, J., Biehler, K. M., & Naragon-Gainey, K. (2024). Resting vagally-mediated heart rate variability in the laboratory is associated with momentary negative affect and emotion regulation in daily life. *Psychophysiology, 61*(12), e14668.

[11] Fu, C., Qian, F., Su, Y., Su, K., Song, S., Niu, M., Shi, J., Liu, Z., Liu, C., Ishi, C. T., & Ishiguro, H. (2025). Facial action units guided graph representation learning for multimodal depression detection. *Neurocomputing, 619*, 129106.