| Volume 02 | Issue 04 | September 2026| ISSN: 3139-1737

International Journal of Advances and Applications in Convergent Studies & Transdisciplinary (IJ-AACST)

Bridging Disciplines, Bridging Knowledge

Aligned with the Vision of NEP 2020 and UGC Guidelines

Website: https://aacstij.org || Email: journaleditor@aacstij.org

XMER: An Explainable Multimodal Emotion Recognition Framework for Trustworthy Human–AI Collaboration

DOI: https://doi.org/10.5281/zenodo.22337947

Sneha P. Shirke[1]

[1]Assistant Professor Faculty of Science & Technology.

New Model Degree College Hingoli-431513, State Maharashtra, India

snehasaw1988@gmail.com

ABSTRACT

The new domain of human-agent collaboration relies more and more on emotion recognition systems that leverage speech, facial expression and text to infer affective states. The majority of the existing multimodal emotion recognition (MER) systems are predominantly black-box systems that might limit their application in critical areas like health care, education and robotic assistance. This paper proposes the combination of cross-modal attention and explainability module based on Grad-CAM, SHAP (SHapley Additive exPlanations), and attention-based modality attribution for multimodal emotion recognition in an explainable MER framework denoted XMER. Not only does XMER predict the emotion, it also provides attribution data for the input modalities and features, which enables understanding of the important factors influencing the prediction. The training goal is a combination of classification loss function and explanation-fidelity regularization. Experimental results on IEMOCAP, CMU-MOSEI and RAVDESS show that they achieve reported accuracies of 81.7% to 84.9% with higher human trust ratings. The framework offers a potential route towards more transparent and human-centered affective computing systems.

KEYWORDS: Multimodal emotion recognition; explainable artificial intelligence; XAI; SHAP; Grad-CAM; attention mechanism; trustworthy AI; human–AI collaboration; affective computing; cross-modal fusion.

1. INTRODUCTION

Affective computing is a technology that seeks to have machines perceive and respond to human affect; it ranges from virtual tutors to clinical decision support and intelligent driver-monitoring systems. The research on computational recognition and interpretation of affective states [1] was the formal beginning of affective computing as a research area. However, no single modality is consistently able to capture the complete range of human affect, and contemporary systems are increasingly designed using a multi-modal approach, which integrates audio, visual and textual information. This complementary information can be learned from multiple channels, and a number of multimodal datasets, like IEMOCAP [3] and CMU-MOSEI [4] have been used to develop these models. Furthermore, the analysis of multimodal sentiment and emotion studies shows the need for the proper design of fusion strategies [5].

However, the depth and complexity of modern multimodal architectures can reduce transparency. Users may receive only a final label such as angry, sad, or neutral without knowing which modality or cue contributed to that decision. Explainability is therefore an important consideration when AI-based affect recognition contributes to decisions affecting people [6]. Reviews of affective computing have also identified interpretability and deployment as continuing challenges [7]. Transparency and explainability are recognized as important principles for trustworthy AI in the European Commission's Ethics Guidelines for Trustworthy AI [8]. The broader debate concerning explanations for automated decisions further emphasizes the importance of understanding the basis of consequential algorithmic decisions [9].

The present paper addresses this gap through XMER, an Explainable Multimodal Emotion Recognition framework for trustworthy human–AI collaboration. XMER combines modality-specific encoders, cross-modal attention fusion, instance-adaptive gated fusion, and a parallel explainability module. The explainability module provides modality-level attribution, visual saliency through Grad-CAM, and acoustic/lexical feature attribution through SHAP.

The major contributions of this work are:

(1) A unified formulation for explainable multimodal fusion.

(2) Model-native modality-wise explanations.

(3) Complementary visual and feature-level attribution.

(4) An explanation-fidelity-aware training objective.

(5) Evaluation on three affective datasets with a human trust assessment.

2. RELATED WORK

2.1 Multimodal Emotion Recognition

Early multimodal emotion-recognition systems commonly processed modalities independently and combined decisions or features. With the development of deep learning, learned multimodal fusion became more prominent. Convolutional architectures have been widely used for facial affect recognition, while CNN–LSTM approaches have been applied to speech emotion recognition, as summarized in recent surveys [10]. The

RAVDESS database has also supported audiovisual and speech emotion-recognition research [11]. Quaternion convolutional approaches have been investigated for speech emotion recognition [12].

Attention-based fusion subsequently became an important approach for multimodal representation learning. Tensor Fusion Network explicitly models multimodal interactions [13], while the Multimodal Transformer uses directional cross-modal attention for unaligned multimodal sequences [14]. MISA separates modality-invariant and modality-specific representations for multimodal sentiment analysis [15]. These studies demonstrate that fusion strategy can strongly influence multimodal performance.

2.2 Explainable AI for Perception Models

Explainable AI techniques can be generally categorized into intrinsic and post-hoc methods. LIME provides explanations of individual predictions through "local perturbations" [16]. The technique of Shapley-value-based attribution has been adapted to estimate feature contributions in SHAP [17] and the visual explanation technique of Grad-CAM has been adapted to convolutional neural networks [18]. These approaches have been used for perception models such as facial emotion recognition. However, post-hoc explanations should not be taken as a given to be faithful. It has been demonstrated that LIME and SHAP explanations can be adversarially attacked [19]. This incentive is a reason why explanation-fidelity aspects are incorporated into the design and validation of XMER instead of being a separate layer of visualization.

2.3 Trust in Human–AI Collaboration

Human–AI studies indicate that explanations can influence user understanding, task performance, and trust. Research on Human–AI team performance demonstrates that explanations can have an impact on complementary team performance [20]. The human–AI collaboration research also highlights the significance of transparency, explainability, and situation awareness in the process of collaboration [21]. The results inspire a study of the human trust in addition to the predictive performance in XMER.

3. RESEARCH GAP

Although there has been improvement in multilabel emotional recognition and explainable Artificial Intelligence, there are still three gaps. First, there are systems that are efficient that can attend to themselves without revealing useful modality level information to the user. Second, most XAI methods have been tested in a unimodal visual, tabular or textual setting, and there is relatively little XAI work that addresses the audio-visual-textual setting. Third, there are fewer studies directly assessing the impact of explanations on human trust in multimodal emotion recognition. XMER fills these gaps by using integrated multimodal fusion, explanation-fidelity regularization and human trust evaluation.

4. PROPOSED XMER FRAMEWORK

XMER consists of four principal components: modality-specific encoders, a cross-modal attention fusion layer, an emotion classification head, and a parallel explainability module that reads intermediate representations without changing the forward inference path.

The facial branch uses a CNN on aligned face crops. The speech branch uses a CNN on log-Mel spectrograms followed by a bidirectional recurrent layer. The text branch uses a transformer encoder on transcripts or ASR output. Each branch produces a fixed-dimensional embedding that is projected into a common latent space.

The fusion layer uses multi-head cross-modal attention to update each modality using information from the other modalities. The architecture is motivated by transformer-based multimodal attention methods for unaligned multimodal sequences [14]. Gated fusion then learns per-instance reliability weights, allowing the system to reduce the contribution of noisy or unavailable modalities.

The same fusion gates and intermediate representations are used for explanation generation. This design links the explanation mechanism to the actual computations of the model rather than constructing a separate surrogate explanation model.

5. MATHEMATICAL MODEL

Let the three modality encoders produce embeddings for the facial (f), speech (s), and text (t) channels:

$$h_m = E_m(x_m), \quad m \in \{f, s, t\}, \quad h_m \in \mathbb{R}^d \tag{1}$$

For query modality m and key/value modality m′, scaled dot-product attention is defined as:

$$A(h_m, h_{m'}) = \text{softmax}(Q_m K_{m'}^T / \sqrt{d_k})\, V_{m'} \tag{2}$$

where $Q_m = h_m W_Q$, $K_{m'} = h_{m'} W_K$, and $V_{m'} = h_{m'} W_V$ are learned projections [22].
The refined embedding is:

$$\hat{h}_m = h_m + \Sigma_{m' \neq m}\, A(h_m, h_{m'}) \tag{3}$$

Gated fusion combines the refined embeddings using instance-adaptive reliability weights:

$$z = \Sigma_{m \in \{f,s,t\}}\, g_m \hat{h}_m \tag{4}$$

$$[g_f, g_s, g_t] = \text{softmax}(G(\hat{h}_f, \hat{h}_s, \hat{h}_t)) \tag{5}$$

The fused vector is passed to the classification head:

$$\hat{y} = \text{softmax}(W_c\, z + b_c) \tag{6}$$

Training minimizes a composite objective combining cross-entropy classification loss with explanation-fidelity regularization:

$$L = L_{CE}(y, \hat{y}) + \lambda_1 \|g^{(k)} - g^{(k-1)}\|_2^2 + \lambda_2(1 - \text{Consistency}(\text{SHAP}, \text{Grad-CAM})) \tag{7}$$

where $\lambda_1 = 0.01$ and $\lambda_2 = 0.05$ in the reported experiments. The second term penalizes instability in modality gates, while the third term penalizes disagreement between complementary attribution mechanisms.

6. EXPLAINABLE AI MODULE

6.1 Modality-Level Attribution

The learned gate weights g_f, g_s, and g_t are directly exposed as normalized modality contributions. For example, a prediction may be represented as 62% facial, 25% vocal, and 13% textual contribution. This provides a model-native answer to which modality influenced a prediction.

6.2 Spatial Attribution Using Grad-CAM

For the facial branch, Grad-CAM is used to identify spatial regions associated with the target emotion [18]. The gradients of the target class with respect to convolutional feature maps are globally averaged to obtain channel weights, which are combined and passed through a ReLU to highlight positive contribution regions.

6.3 Feature-Level Attribution Using SHAP

SHAP values are calculated for interpretable speech and text features, including pitch, energy, speaking rate, and token-level contributions [17]. Kernel SHAP is used to balance attribution stability and latency. A background sample of 100 instances per class is used in the reported implementation. Explanation consistency is monitored during validation to identify potentially unreliable attribution patterns.

7. EXPERIMENTAL DESIGN

Three benchmark datasets were used. IEMOCAP provides synchronized speech, video, and transcripts and is evaluated using a leave-one-session-out protocol [3]. CMU-MOSEI contains more than 23,000 video segments annotated for sentiment and emotion-related analysis [4]. RAVDESS is used as an audiovisual speech and face benchmark [11].

The baselines are a unimodal facial CNN, a unimodal CNN-LSTM speech model, a black-box late-fusion model, and a cross-modal transformer without explainability. All models use the same modality encoders and the reported training budget of 50 epochs, Adam optimization, learning rate 10^{-4}, and batch size 32.

Classification performance is evaluated using accuracy and weighted F1-score. Human trust is measured using a five-point Likert scale. Forty postgraduate students and faculty reviewers evaluated label-only, confidence-score, and complete-explanation conditions in the reported study.

8. RESULTS AND DISCUSSION

Model	Dataset	Accuracy (%)	Weighted F1 (%)	Trust (1–5)
Unimodal CNN (facial only)	IEMOCAP	68.4	64.9	2.9
CNN-LSTM (speech only)	IEMOCAP	70.1	66.3	2.8
Black-box late fusion	IEMOCAP	74.2	70.8	3.1
Cross-modal Transformer (no XAI)	CMU-MOSEI	82.6	79.4	3.3
XMER (proposed)	IEMOCAP	81.7	78.9	4.5
XMER (proposed)	CMU-MOSEI	84.9	81.6	4.6
XMER (proposed)	RAVDESS	83.1	80.2	4.4

Table I. Accuracy, F1-score, and human trust ratings across models and datasets.

The reported experiments show that XMER achieves 81.7% accuracy on IEMOCAP, 84.9% on CMU-MOSEI, and 83.1% on RAVDESS. The reported weighted F1 scores are 78.9%, 81.6%, and 80.2%, respectively. The results indicate that multimodal fusion can provide complementary information compared with the unimodal baselines.

The reported accuracy loss associated with the explanation-fidelity regularization was within 1–2 percentage points of the strongest black-box comparator. The human trust ratings for the complete XMER explanation condition were 4.5, 4.6, and 4.4 on IEMOCAP, CMU-MOSEI, and RAVDESS, respectively. The manuscript reports that these ratings were higher than those obtained under black-box conditions.

Qualitative feedback indicated that Grad-CAM heatmaps were particularly useful for facial emotion predictions, while SHAP feature rankings were useful for ambiguous cases. The reported SHAP computation time was approximately 180 ms per prediction, which may be a limitation for real-time applications.

9. CONCLUSION

This paper introduced XMER, an explainable multimodal emotion recognition framework that integrates cross-modal attention fusion with a model-native explainability module combining modality-gate weights, Grad-CAM spatial attribution, and SHAP feature attribution, trained using an explanation-fidelity-aware objective function. When evaluated on IEMOCAP, CMU-MOSEI, and RAVDESS, XMER achieved accuracy comparable to black-box baselines while yielding significantly higher human-rated trust in its predictions. The results support the view that explainability and predictive performance are not mutually exclusive in affective computing, and that in emotion-sensitive applications explainability should be built into the system rather than added as an afterthought, in order to achieve trustworthy human–AI interaction.

10. FUTURE WORK

XMER will be extended in three directions in future work. First, the Shapley attribution values will be distilled into a lightweight surrogate network to reduce SHAP inference latency, making the framework more feasible for interactive real-time applications such as driver-monitoring dashboards or tutoring systems. Second, an explanation-fidelity loss will be incorporated into the objective function, in the manner used for fairness-auditing techniques, to make XMER more suitable for deployment across diverse user populations. Third, a longitudinal study will examine whether trust built from XMER's explanations persists or fades over time, and whether trust in explanations leads to over-reliance on attributions that are not always accurate — providing a more comprehensive understanding of the role of explainability in sustained, in-the-wild human–AI collaboration.

11. REFERENCES

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[3] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[4] A. Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 2236–2246, doi: 10.18653/v1/P18-1208.

[5] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intell. Syst., vol. 33, no. 6, pp. 17–25, 2018.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing research based on function-component-context framework," IEEE Access, vol. 6, pp. 22681–22700, 2018.

[8] European Commission, High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI. Brussels, Belgium: European Commission, 2019.

[9] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," Int. Data Privacy Law, vol. 7, no. 2, pp. 76–99, 2017.

[10] S. Latif, R. Rana, S. Younis, J. Qadir, and J. E. Qadir, "Speech emotion recognition: An overview of recent research," IEEE Access, 2022.

[11] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, Art. no. e0196391, 2018, doi: 10.1371/journal.pone.0196391.

[12] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," arXiv:2111.00404, 2021.

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. EMNLP, 2017, pp. 1103–1114, doi: 10.18653/v1/D17-1115.

[14] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 6558–6569, doi: 10.18653/v1/P19-1656.

[15] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 1122–1131, doi: 10.1145/3394171.3413678.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[19] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, Soc., 2020, pp. 180–186, doi: 10.1145/3375627.3375830.

[20] G. Bansal et al., "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in Proc. CHI Conf. Human Factors in Computing Systems, 2021, Art. no. 81, pp. 1–16, doi: 10.1145/3411764.3445717.

[21] M. R. Endsley, "Supporting human-AI teams: Transparency, explainability, and situation awareness," Comput. Human Behav., vol. 140, Art. no. 107574, 2023, doi: 10.1016/j.chb.2022.107574.

[22] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017