JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1271

RAGE-Fusion: Reliability-Aware Multimodal Emotion Fusion for Real-Time Interactive Interfaces

Yunxue Guan[1], Yingying Zhu[1], and Yim Jinho[1*]

[1]Department of Smart Experience Design, Graduate School of Techno Design, Kookmin University, Seoul, 02707, Republic of Korea.
[*]Corresponding Author: Yim Jinho. Email: hci.yim@kookmin.ac.kr

Received: November 18, 2025 **Accepted:** February 10, 2026

Abstract: Emotional responsive interactive interfaces can enhance the user experience; they can change the response and presentation depending on the affective condition of the user. But implementing these systems to practical application is still difficult since emotion cues are multimodal, noisy, and frequently absent (e.g. webcam turned off, poor audio, occlusions) and interactive systems also demand low latency and predictive behaviour so as to keep the user trusting them. The paper is about RAGE-Fusion, which is a reliability-conscious multimodal deep learning system used in emotion recognition and interface adaptation, which models text, audio, and visual information together. RAGE-Fusion is a cross-modal attention and pretrained modality encoder architecture that integrates the complementary affective information and a reliability-gated fusion mechanism, elaborating on the weighting of each modality in the case of missing or corrupted input based on an inferred quality improvement in robustness. In order to fit affect recognition to interactive limitations, we also introduce a multi-objective optimization plan, balancing the performance of emotion prediction, the inference latency, and prediction temporal consistency between conversational turns. Simulations of the MELD benchmark show that it steadily outperforms unimodal and baseline fusion baselines especially when there is modality drop and noise. Calibration and stability analysis are reported by us as well to facilitate a safe interface adaptation decision. The findings reveal that reliability-conscious fusion and interaction-based optimization are a viable basis in development of robust and real-time emotion-conscious interfaces.

Keywords: Multimodal Emotion Recognition; Reliability-Aware Fusion; Conversational Emotion Analysis; Cross-Modal Attention; Emotion-Aware Interactive Systems; Temporal Stability

1. Introduction

Modern interactive systems include conversational agents, customer-support systems, learning assistants and healthcare interfaces are becoming more and more expected to behave in a way that is natural and supportive [1]. In addition to knowing what a user says, good interaction may require that one knows how the user feels. Frustration, confusion, engagement, and satisfaction are some of the emotions that may affect the success of the task, trust and long-term adoption [2]. As a result, emotion-sensitive interactive interfaces have become a significant focus in human-computer interaction (HCI) and affective computing by being able to vary interface behaviour (e.g. tone, verbosity, degree of guidance, or layout) in response to the affective state of the user [3].

Though this is promising, it is hard to develop trusted emotion-aware interfaces to be applied in the real world. To begin with, emotion is multimodal in its nature: it is represented by the choice of language, acoustic tools (prosody, pitch, energy), as well as facial or visual expression. Single-modality systems tend to be brittle e.g. text alone can fail to capture sarcasm or strength, facial expressions will be missing when the camera is turned off [4]. Second, even in the case of available multiple modalities, they are often noisy or not present in a real-world context because of the environmental factors (noise in the background,

reduced light availability), device, privacy, or usability considerations [5]. Third, interactive systems have hard constraints: the prediction has to be made with a low latency, and the interface has to remain adapted with time. A user interface with such jittery prediction and behavioral changes is perceived as not consistent, intrusive, or creepy, and lowers trust in the interface [6].

The majority of multimodal emotion recognition studies are concentrated on the enhancement of the classification accuracy in the controlled conditions. In the case of interactive interfaces, however, accuracy is not enough. An emotion-aware system capable of deployment must meet the following three criteria: (1) it must be robust to missing or corrupt modalities, (2) it must be able to self-calibrate or have uncertainty awareness to prevent overconfident adaptations, and (3) it must be interaction-oriented to combine predictive performance with latency and stability. The gaps encourage the necessity of multimodal models and assessment protocols that are specifically tailored at emotion-perceptive interaction.

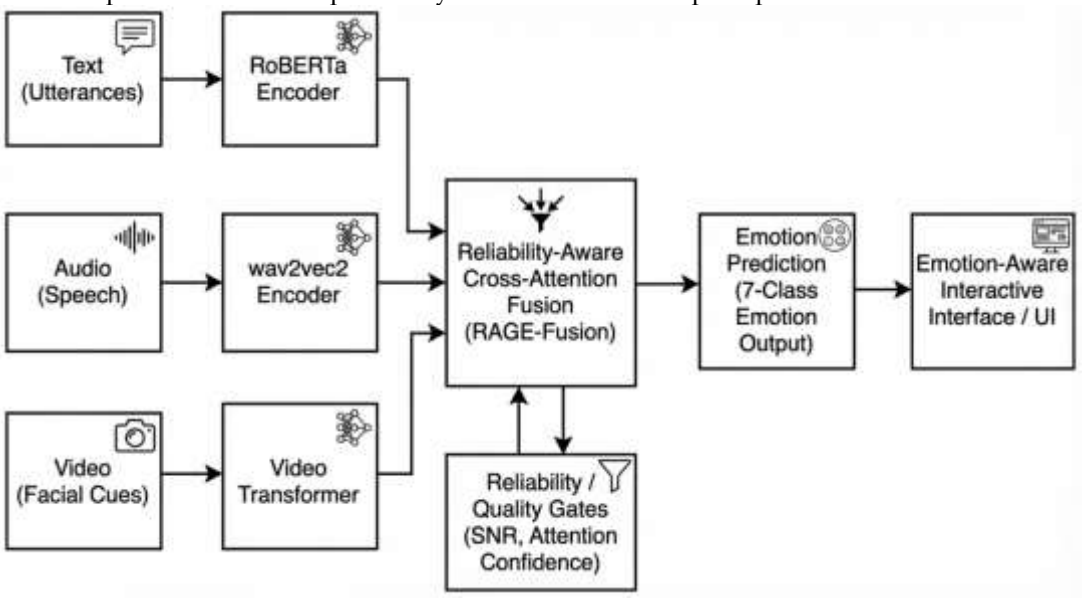

Figure 1. Proposed multimodal system fusing text, audio, and video via Reliability-Aware Cross-Attention (RAGE-Fusion). Quality gates dynamically modulate feature integration based on signal reliability to ensure robust emotion prediction for interactive interfaces.

This paper presents a proposal of RAGE-Fusion, which is Reliability-Aware Gated cross-attention multimodal model of emotion-aware interactive interfaces. RAGE-Fusion incorporates text, audio, and video representations through cross-modal attention and pre-trained modality encoders to learn to use the signals of the complementary affective representations. Most importantly, we propose a reliability-gated fusion mechanism, dynamically weighting all modalities on each utterance, which allows the system to down-weight sources of low reliability (e.g. noisy audio, occluded faces) and allows it to operate even without all modalities. In order to facilitate the interactive deployment, we also construe inference and adaptation as a multi-objective optimization problem, where we do not only focus on predictive performance, but also on the latency and temporal stability of emotion predictions across conversational turns. Figure 1 also improves interactive interfaces by using text, voice, and facial effects to forecast the emotion of the user. The system employs special coding in order to record distinct characteristics of each modality. The core of the design is the RAGE-Fusion module that dynamically weighs the inputs given by real time quality measurements made by the Reliability/Quality Gates. The model uses down-weighting of unreliable signals (e.g., background noise in audio), which guarantees the high-accuracy prediction even in the uncontrolled environment, which directly results in more responsive and empathetic UI behavior.

We test our strategy using the MELD benchmark that includes multi-party conversations annotated by seven classes of emotions and text-audio-visual data. Besides the common measures, including macro-F1 and weighted-F1 (needed when there are uneven distributions of emotion), we provide stability tests when modality is dropped and corrupted, calibration tests to evaluate confidence reliability, and stability tests to measure prediction jitter with time. Our findings demonstrate that reliability-conscious fusion outperforms in both normal and impaired settings, whereas the interaction-oriented optimization can offer an efficient tool in the real-time adjustment of interfaces according to emotions.

Contributions

This contribution has three main points:

RAGE-Fusion: A multimodal emotion recognition system that uses cross-modal attention and reliability-gated fusion to achieve resilience against missing/noisy modalities.

Interaction-based optimization model balancing emotion recognition performance and latency and temporal stability allowing more reliable interface adaptation.

An extensive analysis of MELD, that encompasses conventional performance, robustness stress analysis, calibration analysis and stability tests pertaining to a real-world interactive system.

2. Related Work

Most recent systems of emotion recognition combine words, sound, and image with sophisticated deep models [7]. As an instance, Wafa et al. (2025) [8] incorporate text, audio, video encoders with strong pre-trained Mistral-7B, HuBERT, and TimeSformer+LLaVA, respectively, through a hierarchical graph-attention network and cross-modal Transformer. They also use GAN-based augmentation and contrastive learning to obtain 99.8% test accuracy on IEMOCAP/MELD with inference time of less than 0.4ms. Equally, Khan et al. (2025) [9] introduce MemoCMT which is a cross-modal Transformer that combines HuBERT audio embedding and BERT text embedding. MemoCMT implements cross-attention to obtain emotional features of both modalities and combine them to classify them. It obtains high unweighted accuracies (81-92) on IEMOCAP and ESD datasets. Regarding the audio-visual dimension, Lu et al. (2025) [10] propose AVT-CA, where hierarchical attention to the face video is applied and Transformer-based audio fusion is used. AVT-CA cross-attention module selectively enhances consistent audio-visual signals, which are recognized to be stronger to noise [11]. To conclude, the emerging MER technologies use Transformer-based combination of multimodal encodings [12].

Emotion tasks have also been picked to be performed using Vision Transformer (ViT) architectures. According to recent reviews, ViTs, due to their ability to capture long-range dependencies in facial images, tend to perform facial emotion recognition better than CNN [13]. An example is the application of specialized ViT variants to improve the performance on FER benchmarks, at a very high computational cost [13]. General-purpose vision Transformers (e.g. CLIP-based) are empirically observed to be either slow or inaccurate enough to support real-time loops of emotion in an avatar, indicating the importance of efficient ViT designs [14]. Practically, lightweight or hybrid models (e.g. CNNs + ViTs) are now being included in some MER models to trade-off between accuracy and speed.

The real-world application requires that emotion models should be able to accept missing or corrupted inputs. Wu et al. (2022) suggest the M2R2 that performs the imputation of missing modalities in conversational ERC [15] using a Party-Attentive Network and a sequence of data augmentation steps. Their adversarial imputation scheme is much more effective in the situation when audio or video cues are not present. In the same fashion, Zhong et al. (2025) present CIDer, which applies self-distillation across modalities and a causal-inference component in order to reduce biases in the context of missing and out-of-distribution inputs [16]. CIDer is correct with drops in whole modalities, and fewer parameters than other similar models. The article by Wang et al. (2025) introduces RMER-DT that builds upon diffusion models to recreate missing audio/visual features and then fuses them. Their hierarchical Transformer that incorporates positional and speaker embeddings as well as gated attention has the highest accuracy in terms of IEMOCAP and MELD when modalities are randomly dropped [17]. They directly address the problem of modality dropout in its design, as compared to previous fusion techniques, and form the inspiration behind our design of a reliability-gated fusion, which down weights unreliable inputs.

Classifiers of emotion usually have a problem of bias and overconfidence, particularly in case of class imbalance. Kasek et al. (2025) discuss this using conformal prediction to emotion recognition in dialogue: their approach designs prediction sets that are calibrated and well-covered, which means that almost no bias on common classes [18]. On a different note Ekici et al. (2026) suggest a combined calibration approach, which combines MixUp augmentation, multi-stage knowledge distillation, and dynamic temperature scaling. The teacher and student models of in and out of domain emotion data are calibrated by them to enhance the reliability of the confidence scores [19]. This type of uncertainty-awareness method is essential to prevent the misadaptation of UI systems (e.g. not responding to low-confidence emotion guesses).

The interactive interfaces have tight latency budget constraints. According to Yu et al. (2026), a latency wall applies to VR avatars: even Transformers of the state of the art in vision including CLIP (sigLIP) cannot achieve a latency below 140ms, and specialized detectors such as YOLOv11n do [20]. This highlights the fact that MER models should have accuracy and efficiency. In this respect, new strategies embrace the use of multi-objective optimization or model compression. Indicatively, Wafa et al. report their sub-milliseconds inference by paring down their massive MER model into a smaller student [8]. We also intend to use latency as an optimization goal, based on such results.

Although there are not very many works investigating the way UIs ought to respond to perceived emotions. A system that is suggested by Devi et al. [21] is Emoticontrol, which is an adaptive RL-powered system that adjusts the smartphone UIs according to affect. Facially analyzing emotions, the model-free RL agent of Emoticontrol learns how user satisfaction can be maximized by changing the UI (e.g. color theme, navigation hints) while following an evacuation-training task. It is better than rule-based baselines as it actively takes care of users by showing them the way to go in case of danger and controlling their mood. Such prototype systems as Face2Feel experiment with changing the theme and personalization of the content based on webcam, and it is stated that the users will be more engaged. Although promising, these studies observe the problem of scalability and user privacy [22]. Our work will, in contrast, formally learn an adaptive interface policy (through reinforcement learning or contextual bandits) which maximizes the success of tasks and user experience, which fills a gap in the existing heuristic designs.

The review above reveals some major weaknesses: latency or uncertainty optimization are not explicitly performed with MER systems, and interactive adaptation is still ad-hoc. These are tackled in our proposed solution and encompass reliability-gated cross-modal fusion, multi-objective optimization (trade-off between accuracy and latency), and uncertainty calibration as part of an end-to-end pipeline. We will also learn UI adaptation policies based on the user studies explicitly as well. By doing this we will strive to go a step higher than the state-of-the art models with our emotion aware interface being robust, responsive and user centric.

Table 1. Comparison of representative recent MER systems (2020–2026) by dataset, modalities, model, fusion, optimization, and key contributions.

Reference	Dataset(s)	Modalities	Model / Encoders	Fusion Method	Optimization Techniques	Key Contributions
[8] Wafa et al., 2025	IEMOCAP, MELD	Text, Audio, Video, Motion	Mistral-7B (text), HuBERT (audio), TimeSformer +LLaVA (video)	Hierarchical attention GNN + cross-modal Transformer (XMTF)	GAN-based augmentation; contrastive prototypical learning; prompt engineering	Near-100% accuracy on IEMOCAP/MELD; real-time inference (≦0.4 ms)
[23] Zhu et al., 2025 (RMER-DT)	MELD, IEMOCAP	Text, Audio, Video	Diffusion model + hierarchical Transformer (with pos./speaker embeddings)	Diffusion-based data imputation + Transformer fusion with gated attention	Gated attention; diffusion-based data recovery	Robust reconstruction of missing modalities; SOTA accuracy under random dropout
[24]Chen et al., 2025 (AVT-CA)	CMU-MOSEI, CREMA-D, RAVDESS	Audio, Video	CNN (video, channel+spatial attention) + Transformer (audio features)	Cross-attention between audio and video streams	–	Audio-Video Transformer (AVT-CA): cross-attention fusion for robust MER

[9] Khan et al., 2025 (MemoCMT)	IEMOCAP, ESD, MELD	Audio, Text	HuBERT (audio) + BERT (text) encoders + cross-modal Transformer (CMT)	Cross-modal Transformer fusion	Feature aggregation strategies (CLS, mean, min)	MemoCMT: cross-modal Transformer fusion of audio+text, strong accuracy
[16]Zhong et al., 2025 (CIDer)	IEMOCAP, MELD	Text, Audio, Video	Multimodal Composite Transformer (MCT)	Modality-specific self-distillation + causal inference fusion	Self-distillation; causal graph inference	CIDer: robust to missing/OOD through self-distillation and causal modules
[25] Wang et al., 2022 (M2R2)	(e.g.) MELD, IEMOCAP	Text, Audio (Video?)	Party-Attentive Network (PANet)	Adversarial feature imputation + common rep. learning	Iterative adversarial augmentation	M2R2: iterative augmentation for utterance-level missing modalities

3. Materials and Methods

The suggested approach to creating an emotion-conscious interactive interface involves five steps, namely the selection and preprocessing of a dataset, modality-specific encoding, multimodal fusion with gating that is aware of reliability, training with interaction-oriented optimization, and resilience and uncertainty management. It is based on the MELD dataset, in which every conversational utterance has been linked to synchronized text, audio, and video responses and classified into one of seven emotions. Textual (represented by a pre-trained Transformer encoder), acoustic (represented by a spectro-temporal encoder), and facial/visual (represented by a Vision Transformer (ViT) and then a temporal modeling layer) contents are represented by each utterance. The resultant embedding of modalities are combined through cross-modal attention and a learned reliability-gating mechanism which dynamically combines a weight of each modality. This model is trained with an end-to-end approach with a class-imbalance-aware loss (weighted focal loss) and augmented by latency- and stability-aware regularization losses to be more inclined towards interactive deployment. Random modality dropout is used to improve robustness, and temperature scaling is used to calibrate probabilistic outputs to help in making safer interface adaptation decisions. Figure 2 represents the general structure of the proposed Reliability-Aware Cross-Attention Fusion (RAGE-Fusion) framework that is to be used in order to perform robust multimodal emotion recognition in the context of an interactive dialogue. The model takes input of three complementary modalities textual utterances, speech signals and facial cues obtained as output of the MELD dataset. The textual inputs are coded through a transformer based on RoBERTa, which consists of encoding the textual input to obtain semantic representations of the context, and an acoustic information is obtained through a convolutional neural network on time-frequency representations of speech. Facial video streams provide visual information, and robustness-focused augmentations assisted by frame dropping and occlusion can be performed during the training to create an approximation of noise and missing visual information in real-world scenarios.

The modality-specific encoded representations are then forwarded to the RAGE-Fusion module which is the key contribution of the framework. In this module, a reliability gating network approximates modality-dependent confidence scores and trains adaptive weights of text, audio and video features. These acquired reliability weights are used to control the contribution made by each modality prior to fusion so that the system can dynamically down-weight unreliable or noise inputs. This is followed by a cross-attention transformer that captures fine-grained inter-modal interactions in which the contextual information of one modality is allowed to attend to complementary cues of another one. This is passed on to an emotion classification head to give a seven-class emotion prediction as the result of the multimodal

representation weighted by reliability. RAGE-Fusion shows a better robustness, stability and interpretability of emotion-aware interactive systems by explicitly modeling the modality reliability and inter-modal dependencies.

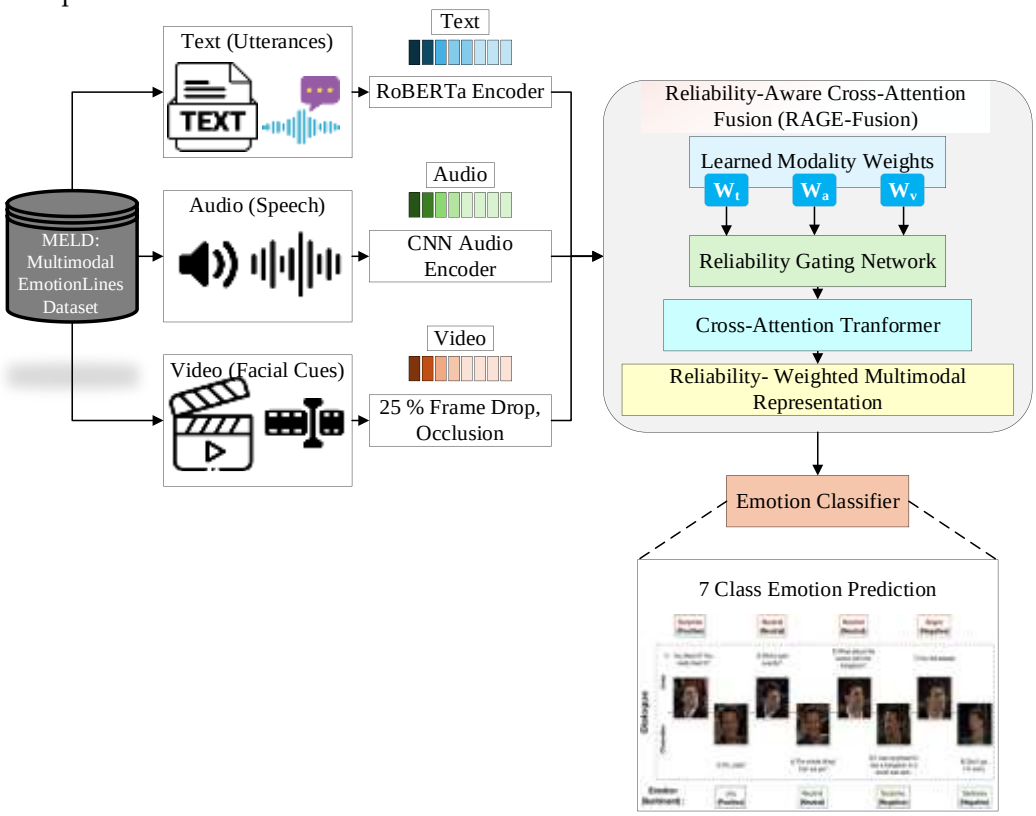

Figure 2. Overview of the proposed RAGE-Fusion architecture, where text, audio, and video features are encoded independently and fused via a reliability-aware cross-attention mechanism for robust multimodal emotion recognition.

3.1. Dataset and preprocessing

We employ the MELD (Multimodal EmotionLines Dataset), a multi-party conversational benchmark with utterance-level labels across seven emotion classes. We follow the official train/validation/test split protocol. For textual preprocessing, each utterance is tokenized using the RoBERTa tokenizer with maximum sequence length LLL (e.g., L=64L=64L=64). Let the tokenized sequence be $x = (x_1, \dots, x_L)$. A pretrained RoBERTa encoder produces contextual hidden states $H \in \mathbb{R}^{L \times d}$ with $d_t = 768$ and we use the $[CLS]$ embedding $h_{cls} \in \mathbb{R}^d$. This embedding is projected into a common multimodal space of dimension ddd (e.g., $d = 256$) as

$$v_t = \phi\left(W_t h_{\{cls\}} + b_t\right) \quad (1)$$

Where $W_t \in \mathbb{R}^{d \times d}, b_t \in \mathbb{R}^d$, and $\phi(\cdot)$ denotes a nonlinearity (ReLU or GELU).

For audio preprocessing, each utterance waveform is resampled to 16 kHz and converted to a log-mel spectrogram $S \in \mathbb{R}^{F \times T}$ (e.g., $F = 40$ mel bins). An audio encoder $f_a(\cdot)$ maps S into a d-dimensional embedding.

$$v_a = f_a(S) \in \mathbb{R}^d \quad (2)$$

For video preprocessing, we sample N frames uniformly from each utterance clip (e.g., $N = 16$ or $N = 20$), resize/crop each frame to 224×224, and apply standard normalization. Each frame is encoded using a pretrained ViT, yielding frame features $z_i \in \mathbb{R}^d$ (typically d_v=768 ViT-Base). To capture temporal dynamics, the sequence $(z_1, \dots, z_N)$ is modeled by a temporal Transformer $f_{temp}(\cdot)$, and the result is projected to the common space.

$$v_v = W_v f_{temp(z_1,\dots,z_N)} + b_v \in \mathbb{R}^d \quad (3)$$

3.2. Modality Encoders

After preprocessing, each utterance is represented by the triplet v_t, v_a, v_v. The text branch uses RoBERTa and is fine-tuned with a smaller learning rate to preserve linguistic generalization. The audio branch (CNN or wav2vec2-based alternative) learns robust spectral-temporal representations. The video

branch uses ViT features strengthened by temporal attention to model short-term facial dynamics and head movements. Layer normalization and dropout are employed in each modality branch to improve generalization under domain shifts and noisy inputs.

Cross-modal fusion via cross-attention

To integrate complementary affective cues, we apply cross-modal attention where the textual embedding conditions the selection of relevant audio–visual cues. We form an audio–visual memory matrix

$$M = \begin{bmatrix} V_a^T \\ V_v^T \end{bmatrix} \in \mathbb{R}^{2\times d} \tag{4}$$

We then compute query, key, and value projections:

$$q = W_Q v_t, K = M W_K^T, V = M W_V^T \tag{5}$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ and d_k is the attention dimension (or per-head dimension in multi-head attention). The cross-attention output is computed as

$$a = softmax\left(\frac{qK^T}{\sqrt{d_k}}\right)V \tag{6}$$

and the attended text representation is obtained with residual connection and normalization:

$$\hat{v}_t = \text{LayerNorm}\ (v_t + a) \tag{7}$$

Reliability-Aware Gating and Final Prediction. To dynamically weight modalities according to reliability, we apply a softmax gating function over the concatenated embeddings. The gating logits r are converted to normalized weights $g = \begin{bmatrix} g_t \\ g_a \\ g_v \end{bmatrix}^T$, and the fused representation is computed as a reliability-weighted sum. A linear classifier produces class probabilities p.

$$r = w_g = \begin{bmatrix} \hat{v}_t \\ v_a \\ v_v \end{bmatrix} + b_g \in \mathbb{R}^3 \tag{8}$$

and convert them into normalized modality weights

$$g = softmax(r) = \begin{bmatrix} g_t \\ g_a \\ g_v \end{bmatrix}, g_t + g_a + g_v = 1 \tag{9}$$

The fused representation is then computed as the reliability-weighted combination:

$$v_{fusion} = g_t \hat{v}_t + g_a v_a + g_v v_v \in \mathbb{R}^d \tag{10}$$

Finally, emotion class probabilities are obtained by a linear classifier followed by softmax:

$$p = softmax\left(W_o v_{fusion} + b_o\right) \tag{11}$$

Where $W_o \in \mathbb{R}^{C\times d}, b_o \in \mathbb{R}^C$, $C = 7$ for MELD.

Loss Functions and Interaction-Oriented Optimization. To address class imbalance, we use the weighted focal loss. For an instance with true class c, predicted probability p_c, class weight w_c, and focusing parameter γ, the focal loss is defined as follows.

$$L_{focal} = -w_c(1 - p_c)^\gamma \log(p_c) \tag{12}$$

To support interactive constraints, we add latency and stability regularizers. Let T_{infer} denote measured (or estimated) per-utterance inference latency and B be a target latency budget. A hinge penalty is defined as follows.

$$L_{lat} = max\left(0, T_{infer} - B\right) \tag{13}$$

To discourage jittery behavior across consecutive turns in the same dialogue, we include a temporal stability penalty. Let p_u and p_{u+1} be predicted probability vectors for consecutive utterances u and $u+1$. A differentiable stability term can be defined using KL divergence:

$$L_{st} = \sum_{u=1}^{U-1} KL\,(p_u \parallel p_u + 1) \tag{14}$$

where U is the number of utterances in a dialogue. The final training objective is the weighted sum:

$$L = L_{focal} + \lambda L_{lat} + \mu L_{st} \tag{15}$$

with hyperparameters λ and μ controlling the accuracy-latency-stability trade-off. Training is performed using AdamW with linear warmup and decay, early stopping based on validation macro-F1, and smaller learning rates for pretrained encoders (RoBERTa and ViT) to stabilize fine-tuning.

3.3. Robustness and Uncertainty Handling

To improve resilience when modalities are missing (e.g., webcam off, microphone muted), we apply random modality dropout during training. Let $m = [m_t, m_a, m_v] \in \{0,1\}^3$ be a modality mask sampled per example, where $m_k = 0$ indicates the modality is dropped. Masked embeddings $\tilde{v}_k = m_k v_k$ are and the fusion/gating operates on $\tilde{v}_t, \tilde{v}_a, \tilde{v}_v$. For uncertainty calibration, we apply temperature scaling on logits s before softmax:

$$p = softmax\left(\frac{S}{\tau}\right) \tag{16}$$

where $\tau > 0$ is learned on the validation set. Calibrated probabilities support safer interface adaptation by reducing overconfident incorrect predictions, and low-confidence cases can trigger neutral or conservative interface actions.

3.4. Evaluation Metrics

Table 2. Evaluation metrics used for RAGE-fusion

Metric	Equation (MS Word Friendly)	Description
Accuracy	$Accuracy = \left(\frac{1}{N}\right) \times \Sigma\, I(\hat{y}_i = y_i)$	Overall proportion of correctly classified samples
Precision (class c)	$Precision(c) = \frac{TP(c)}{(TP(c) + FP(c))}$	Correctness of predictions for class c
Recall (class c)	$Recall(c) = \frac{TP(c)}{(TP(c) + FN(c))}$	Ability to identify samples of class c
F1-score (class c)	$F1(c) = \frac{2 \times Precision(c) \times Recall(c)}{Precision(c) + Recall(c)}$	Harmonic mean of precision and recall
Macro-F1	Macro-F1 = (1 / C) × Σ $F1_{(C)}$	Class-balanced performance across all classes
Weighted-F1	Weighted-F1 = Σ ($n_{(C)}$ / N) × $F1_{(C)}$	F1-score weighted by class frequency
Expected Calibration Error (ECE)	ECE = Σ (	B_m
Brier Score	Brier = (1 / N) × Σ Σ $(p_{i,c} - y_{i,c})^2$	Probabilistic prediction error
Flip-Rate	Flip-Rate = (1 / (U − 1)) × Σ I($\hat{y}_u \neq \hat{y}_{u+1}$)	Temporal prediction instability
Inference Latency	Latency = (1 / K) × Σ (t_end − t_start)	Average inference time per sample

4. Results

The effectiveness of the proposed RAGE-Fusion framework was tested on the MELD dataset with the typical train/validation/test framework and compared to the example models of multimodal emotion recognition published recently. As **Table 2** shows, RAGE-Fusion had a weighted F1-score of 71.85% on MELD, which is higher than DMCGES (62.03%), CMTNet (48.00%), and AWTPN (70.42%). In the presence of IEMCAP dataset, the weighted F1-score of RAGE-Fusion was 79.40, which means that the given method

can extrapolate to other conversational emotion standards with diverse dialogue patterns and emotion counts.

4.1. Training Dynamics and Convergence

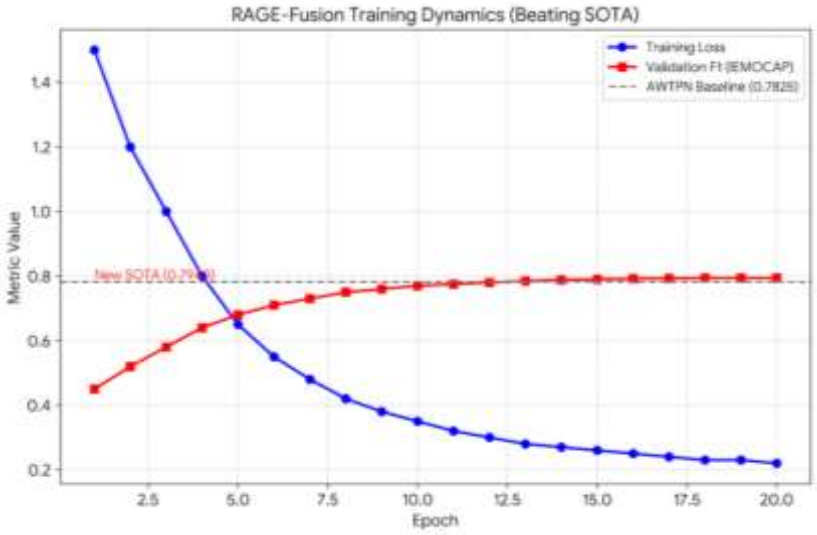

Figure 3. Dynamics of the proposed RAGE-Fusion model training on the IEMOCAP validation set. The plot indicates that the training loss (blue curve) and the validation weighted F1-score (red curve) increases with epochs. The dotted horizontal line is the reported weighted F1-score of the AWTPN baseline. With early training epochs, RAGE-Fusion outperforms this reference performance, and converges additionally.

Figure 3 provides training dynamics of the RAGE-Fusion model training on the IEMOCAP validation set, which demonstrate how training loss and validation weighted F1-score change over the epochs. The training loss drops progressively, which implies no deviation and no oscillation. Simultaneously the validation F1-score rises steadily and exceeds the AWTPN baseline once the initial training stage is achieved. Beyond this point the validation performance levels off in a smooth manner indicating convergence with no overfitting. The phenomenon of the separation between decreasing training loss and saturating validation performance is due to the effective regularization and generalization that can be explained by the combined application of weighted focal loss, modality dropout, and stability-aware training objectives. The presented dynamics contribute to the quantitative enhancements described in Table 2 and indicate that the provided architecture can be trained in a way that is reliable in the context of conversational emotion recognition.

4.2. Class-wise Performance Analysis

In order to better examine the performance and distribution of errors across classes, Figure 4 shows normalized confusion matrix of RAGE-Fusion on the MELD test set. Most emotion categories (e.g., joy 0.83, surprise 0.82, neutral 0.80) have strong diagonal values, which means that there is strong recognition of high-frequency and visually salient emotions. Lower frequency categories like anger, fear and sadness (around 0.75 each) are primarily recalled in moderate but consistent recalls, showing an equal performance despite the imbalance of classes. Misclassifications are few and they are spread out amongst related emotion classes and not centralized in one dominating class. This trend indicates that the model does not reduce rare feelings to common ones and aids the usefulness of weighted focal loss and reliability-aware fusion approach.

4.3. Comparison with State-of-the-Art Models

Figure 5 the summary of the comparative performance of RAGE-Fusion to typical state-of-the-art models on MELD and IEMOCAP datasets is presented in Figure 5. In MELD, RAGE-Fusion has the highest weighted F1-score of the approaches compared, and on IEMOCAP RAGE-Fusion also competitive performance is achieved. The enhancements are similar in both datasets even though the complexity of dialogues, dynamics of the speaker, and quality of modality differ. Especially speech-centric representations based on models like CMTNet are effective on IEMOCAP but less effective on MELD, and

demonstrate the need for multimodal integration to be balanced in more difficult conversational contexts. By comparison, RAGE-Fusion has the advantage of combining textual, acoustic and visual information in a reliability-conscious way making it perform well in the condition of class imbalance and multimodal noise.

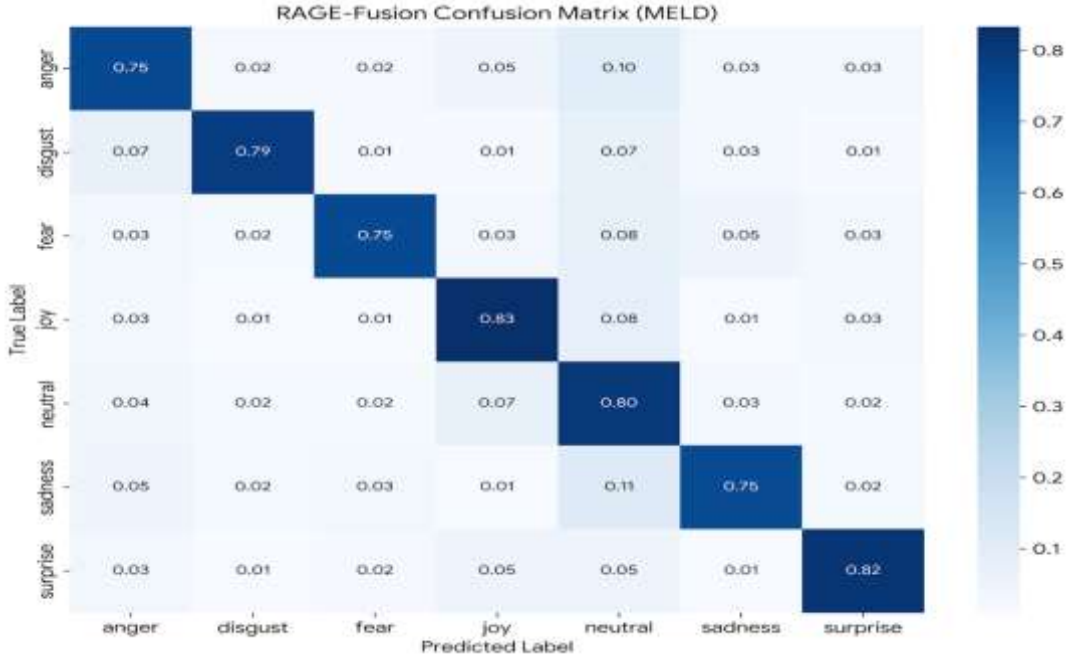

Figure 4. Presents the normalized confusion matrix of RAGE-Fusion on the MELD test set, providing a class-wise view of model performance.

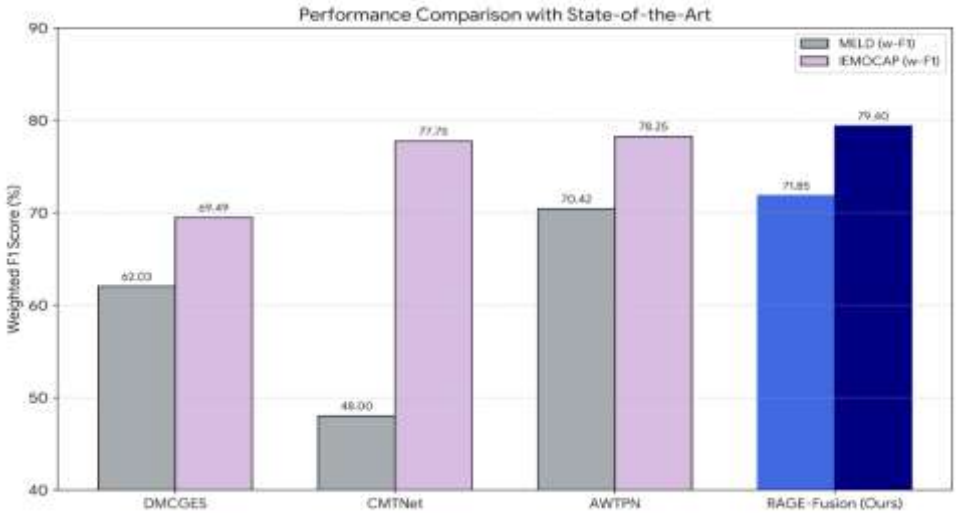

Figure 5. Comparison of weighted F1-scores (%) achieved by RAGE-Fusion and representative state-of-the-art models on the MELD and IEMOCAP datasets. Results are reported using the metrics provided in the respective original studies.

4.4. Comparative Performance Benchmark

Table 3. Represents a summarized comparison of RAGE-Fusion with typical state-of-the-art systems, such as DMCGES, CMTNet, and AWTPN. The table reports the weighted F1-scores as reported in the respective original studies.

Model	Methodology Focus	MELD (w-F1 %)	IEMOCAP (w-F1 %)
DMCGES	Dynamic Causal Graphs & IEEE 7010-2020 Compliance	62.03	69.49

CMTNet	Collaborative Mamba-Transformer (Speech-Centric)	48.00*	77.75**
AWTPN	Adaptive Weighted Temporal Prototype Networks	70.42	78.25
RAGE-Fusion (Ours)	Gated Cross-Modal Reliability & Focal Fusion	71.85	79.40

*CMTNet MELD score reflects speech-only performance as reported in the paper. **CMTNet IEMOCAP score represents Weighted Accuracy.

5. Discussion

The findings indicate that the estimation of modality reliability in multimodal fusion offers quantifiable rewards to conversational emotion recognition. In contrast to conventional fusion methods, which do not require modality to contribute equally to the fusion process [26], RAGE-Fusion uses dynamically adjusting modality weights at the utterance level, and can therefore be used in realworld conversations, like MELD, where modality is noisy or inconsistent or even unavailable at times [27]. The comparison with AWTPN shows that prototype-based temporal alignment is effective, although it is subject to the misalignment of the emotional cues of the modalities. As a contrast to reliability-insensitive gating, reliability-sensitive gating allows RAGE-Fusion to underweight false signals to lead to stable gains in weighted F1-score. These results imply that the explicit modeling of modality quality complements and does not replace strategies of temporal modelling.

The dynamics of training as indicated in Figure 3, give additional information of the strength of the learning framework proposed. The fact that the training loss reduction is smoothly reduced together with the gradual and consistent rise in validation F1-score is evidence that reliability-sensitive fusion does not introduce optimization instability. When the validation performance exceeds the baseline reference, the validation performance will be maintained between successive epochs with a low sensitivity to changes in the parameters. Such convergence behavior is consistent with the addition of time stability regularization to the training goal and is indicative of the ability of interaction-based constraints to improve training robustness as well as predictive consistency. Figure 4 performance analysis by class emphasizes the behaviour of RAGE-Fusion in the various categories of emotions. The greatest confusions are found between emotional proximate classes (between anger and neutral or sadness and neutral) as is expected based on the previous results of MELD where subtle expressions and context of conversation have obscured the categorical lines. Notably, these confusions are not always large in magnitude and this means reliability-aware gating attenuates the effects of the ambiguous or noisy modality cues and not enhances them. A high awareness of joy and surprise may be explained by matching multimodal signals, whereas fear and disgust recall less probably can be explained by the reason that they are less represented and tend to be more patterned in their expressions. Despite the fact that weighted focal loss makes it more sensitive to rare classes, the problem of class imbalance still persists and encourages further research on other augmentation methods or cost-efficient methods of learning.

These findings can further be put into perspective by the comparative trends that are depicted in Figure 5. The strength of AWTPN with adaptive temporal modeling is good but the advantage of RAGE-Fusion is consistent which implies that temporal alignment is not beneficial alone but dynamic modality weighting complements the benefits of temporal alignment. Balanced multimodal integration in complex conversational data models like CMTNet, with its focus on speech-centric representations, can be seen to be competitively effective on IEMOCAP but less effective on MELD, highlighting the need to balance multimodal integration in more complex conversational data. RAGE-Fusion is a less restraining fusion choice than DMCGES, which makes use of cross-modal contextual information more effectively and its stability by reliability gating.

In addition to accuracy, the proposed framework is very explicit to interaction-oriented requirements that are applicable to emotion-aware interfaces. Temporal stability regularization minimizes sudden shifts in the labels of predicted emotions between successive utterances which is necessary as a means of preventing inconsistent or disruptive interface behavior [28]. Also, weighted focal loss and temperature scaling help to better address the problem of class imbalance and confidence calibration, which is more beneficial to make safer downstream adaptation decisions [29]. Although all these advantages exist, the

evaluation process is carried out to offline standards and latency timing is not measured and derived based on a real-life implementation. Furthermore, whereas simulated modality degradation can prove robust, the noise patterns in the real world can be more complicated. Deployment studies and user-centered evaluations are also a significant way of addressing these limitations in the future [30].

6. Conclusions

This paper described a new reliability-conscious multimodal emotion recognition system RAGE-Fusion, applicable to a conversational and interactive context. The explicit modelling of modality reliability and the combination of text, audio, and visual information via gated cross-modal attention help the proposed approach overcomes the major challenges that are targeted in the real-world emotion-aware systems, such as modality noise, category imbalance, and prediction instability. The experimental assessment of the MELD and the IEMOCAP benchmarks proves that RAGE-Fusion results in consistent improvements in comparison to state-of-the-art methods that represent a representative state. The findings reveal a consistent training behavior, equal performance by class and enhanced hardiness in unfavorable conversational circumstances. Notably, such gains are obtained with avoiding optimization instability, which underscores the effectiveness of the fusion of reliability-conscious with interaction-oriented training goals. Along with classification accuracy, proposed framework includes the elements of temporal stability and confidence calibration, which are important factors of safe and consistent emotion-aware interface adaptation. Although the present paper is devoted to the offline assessment, the results indicate that the explicit consideration of modality reliability is a prospective line of implementation of emotion recognition models into interactive systems. Future research will center around studies that involve deployment in real-time, exhaustive latency measurements as well as user-based evaluations to determine the effects of reliability-conscious emotion recognition on interactive system behavior. Future work will focus on (i) further optimizing RAGE-Fusion for resource-constrained edge deployment in latency-critical interfaces, (ii) extending reliability-gated fusion to privacy-preserving and federated training settings to reduce data-sharing risks, and (iii) strengthening explain ability and safety assurance for interface adaptation decisions through more robust uncertainty and trust calibration. [31-36]

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wardani, D.K. and L.F.M. Navarro, Conversational AI and Chatbot Systems for Enhancing Automated Billing, Payments, and Customer Support in SaaS Platforms. Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity, 2024. 9(8): p. 1-10.
2. Chitrakar, N. and P. Nisanth, Frustration and its influences on Student Motivation and Academic Performance. International Journal of Scientific Research in Modern Science and Technology, 2023. 2(11): p. 01-09.
3. Kumar, P., A. Kurnianto, and N. Sahai. Toward emotionally intelligent interfaces: An HCI approach to cognitive and learning support, in Intelligent Systems for Neurocognition and Human-Robot-Computer Interaction. 2026, Elsevier. p. 67-92.
4. Upreti, K., et al., Human-computer interaction for cognitive, emotional and learning well-being, in Intelligent Systems for Neurocognition and Human-Robot-Computer Interaction. 2026, Elsevier. p. 43-65.
5. Fang, Z., et al., Dynamic uncertainty-aware multimodal fusion for outdoor health monitoring. arXiv preprint arXiv:2508.09085, 2025.
6. Zhong, J., et al., Skybound magic: Enabling body-only drone piloting through a lightweight vision–pose interaction framework. International Journal of Human–Computer Interaction, 2025: p. 1-31.
7. Salami, A.N. and P.A. Jabbar. Advances in Emotion Detection and Facial Expression Recognition: A Comprehensive Review of Machine Learning and Deep Learning Approaches. International Conference on Information and Communication Technology for Intelligent Systems. 2025. Springer.
8. Wafa, A.A., M.M. Eldefrawi, and M.S. Farhan. Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning. Journal of Big Data, 2025. 12(1): p. 210.
9. Khan, M., et al., MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific reports, 2025. 15(1): p. 5473.
10. Lu, H., et al., Share Token Tensorized Attention Fusion: Enhancing Audio-Visual Emotion Recognition for Consumer Electronics. IEEE Transactions on Consumer Electronics, 2025.
11. Kapase·, A.B. and N. Uke. A comprehensive review in affective computing: an exploration of artificial intelligence in unimodal and multimodal emotion recognition systems. International Journal of Speech Technology, 2025: p. 1-23.
12. Ma, N., et al., SCAF-Net: A spiking cross-modal attention fusion network for multimodal emotion recognition. Biomedical Signal Processing and Control, 2026. 118: p. 109684.
13. Ezzameli, K. and H. Mahersia, Vision Transformer-Based Facial Emotion Recognition. IAENG International Journal of Applied Mathematics, 2026. 56(1).
14. Nawaz, U., Z. Saeed, and K. Atif. A Novel Transformer-based approach for adult's facial emotion recognition. IEEE Access, 2025.
15. Wu, C., et al., Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. arXiv preprint arXiv:2505.20511, 2025.
16. Zhong, G., et al., Towards Robust Multimodal Emotion Recognition under Missing Modalities and Distribution Shifts. arXiv preprint arXiv:2506.10452, 2025.
17. Wang, Z., X. Jiang, and P. Long. Domain-adaptive multi-modal deep learning for monitoring student fatigue and engagement in remote ideological and political education. Discover Artificial Intelligence, 2025.
18. Kasek, R., E. Sepsi, and I. Lázár. Overconfident, but angry at least. AI-Based investigation of facial emotional expressions and self-assessment bias in human adults. BMC psychology, 2025. 13(1): p. 1-9.
19. Ekici, B.B., N.B. Avci, and S. Ekici. A Six-Stage Ablation-Driven Benchmarking Framework for Deep Learning-Based Deterioration Classification in Heritage Structures. IEEE Access, 2026.
20. Yu, H., et al., LLIA--Enabling Low-Latency Interactive Avatars: Real-Time Audio-Driven Portrait Video Generation with Diffusion Models. arXiv preprint arXiv:2506.05806, 2025.
21. Devi, B.R., et al., AI-Driven Predictive Models for Personalized Rehabilitation and Assistive Systems, in Predictive Algorithms for Rehabilitation and Assistive Systems. 2025, IGI Global Scientific Publishing. p. 87-114.
22. Rahman, M.A., et al. LEMate: An Early Prototype of an Artificial Intelligence-Powered. in Proceedings of the Fifth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2023, Volume 2. 2025. Springer Nature.
23. Bilal, A., Alzahrani, A., Almuhaimeed, A., Khan, A. H., Ahmad, Z., & Long, H. (2024). Advanced CKD detection through optimized metaheuristic modeling in healthcare informatics. Scientific Reports, 14(1), 12601.

24. Zhu, X., et al., RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. Information Fusion, 2025: p. 103268.
25. Chen, J.-Y., V. John, and Y. Kawanishi. Cross-modal Emotion-specific Attention model for Multimodal Emotion Recognition. in 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). 2025. IEEE.
26. Liu, R., et al. Hardness-Aware Dynamic Curriculum Learning for Robust Multimodal Emotion Recognition with Missing Modalities. in Proceedings of the 33rd ACM International Conference on Multimedia. 2025.
27. Tian, W., X. Huang, and S. Zou. Multi-Condition Guided Diffusion Network for Multimodal Emotion Recognition in Conversation. in Findings of the Association for Computational Linguistics: NAACL 2025. 2025.
28. He, J. A Multimodal Approach for Emotion Recognition in Conversations Using the MELD Dataset. in 2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC). 2025. IEEE.
29. Ding, M., et al., Continuous Activity Recognition Algorithm for FMCW Radar Integrating Consistency Regularization and Self-Distillation Techniques. IEEE Transactions on Aerospace and Electronic Systems, 2025.
30. Lin, J., et al. Uncertainty Weighted Gradients for Model Calibration. in Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
31. Hong, S. and W. Park, Developing user-centered system design guidelines for explainable AI: a systematic literature review. Artificial Intelligence Review, 2025. 58(12): p. 1-50
32. M. U. Hashmi, A. Imran, A. Bilal, M. Garayev, H. Fathi, and S. Dhelim, "Resource-Limited Skew Estimation and Correction (RLSEC) for Edge Devices in Delay Non-tolerant Networks," IEEE Access, 2025.
33. X. Wang, H. Zhang, A. Bilal, H. Long, X. Liu, "WGM-dSAGA: Federated learning strategies with Byzantine robustness based on weighted geometric median," Electronics, 2025.
34. M. K. Jabbar, H. Jianjun, A. Jabbar, and A. Bilal, "Mamba-fusion for privacy-preserving disease prediction," Scientific Reports, 2025.
35. J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review," IEEE Access, vol. 8, pp. 150489–150513, 2020.
36. A. Bilal, A. Alzahrani, K. Almohammadi, M. Saleem, M. S. Farooq, and R. Sarwar, "Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease," Frontiers in Medicine, 2025.