SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Open Problems and Future Directions in Multimodal Affective Computing: A Confidence-Aware Fusion Case Study

Saleha Mariyam[1], Halima Sadia[2*]

[1,2]Department of Computer Science and Engineering, Integral University, Lucknow, India
*Corresponding author: halima@iul.ac.in

ABSTRACT
Affective computing has moved from single-channel emotion classification toward deep multimodal systems that jointly exploit speech, text, facial cues, and physiological signals. This paper's primary contribution is a structured review of that shift, organised around unimodal recognition methods, multimodal fusion strategies, and the deep architectures now driving progress — convolutional and recurrent backbones, transformers, graph neural networks, selective state-space models, and self-supervised learners. To test whether the review's central architectural claim holds empirically, and to illustrate, we then reimplement a confidence-aware fusion mechanism under a common training protocol on the MELD and CMU-MOSEI conversational benchmarks and compare it against our own reimplementations of two recent reliability-aware and graph-based methods from the literature. Under clean conditions the confidence-aware variant improves weighted-F1 by 1.6 points over the stronger reliability-aware baseline and by 13.0 points over the static-graph baseline; under 50% single-modality corruption, its weighted-F1 falls by only 6.3 points, against 18–20 points for the reimplemented baselines. These results support a claim the review otherwise makes on architectural grounds: reliability estimation is not a peripheral add-on but a load-bearing component of robust multimodal fusion. The case study also exposes what reliability modelling alone does not solve, synthetic corruption does not fully represent real sensor failure, graph construction scales poorly with dialogue length, and fixed hyperparameters limit generalisation across deployment settings. Six open problems are analysed in this light: missing or unreliable modalities, cross-domain and cross-subject generalisation, data scarcity, privacy constraints, limited explainability, and the real-time deployment costs that reliability-aware methods themselves introduce. Comparative tables summarise representative unimodal methods, fusion strategies, and architectures drawn from recent literature, and each open problem is mapped to a corresponding research direction.

Keywords: affective computing; multimodal emotion recognition; deep learning; open problems.

I. Introduction

Emotion is central to how people communicate and make decisions, and recognising it automatically is a long-standing problem in artificial intelligence and human–computer interaction [1], [2]. Human emotional expression can be described along two broad axes: physiological arousal, measurable through internal signals such as electroencephalography (EEG) and galvanic skin response, and subjective experience, expressed outwardly through speech, facial expression, text, and gesture [3]. Audio-visual and textual cues are informative but can be masked deliberately or vary across cultures, whereas physiological signals are harder to suppress consciously, which is why systems that combine both kinds of evidence tend to be more reliable [1], [3].

Unimodal recognition using only speech, only facial video, only text, or a single physiological channel — has matured considerably, but a single channel is rarely sufficient on its own: people vary in how expressively they display emotion, input signals are frequently noisy or partially occluded, and cross-cultural differences in expression add a further confound [4], [5]. Multimodal approaches address this by combining complementary evidence across channels, and generally outperform unimodal baselines when the fusion is designed well [3]–[5]. The recurring design question in this literature is not whether to fuse modalities, but where in the pipeline fusion should occur and how the differing temporal characteristics, sampling rates, and reliability of each modality should be reconciled [4], [20]–[24].

This paper is organised so that the literature review carries the main argument, with a compact empirical case study used to test the argument. Sections II–V review, in turn, the foundational emotion models and benchmarks, unimodal recognition per channel, the taxonomy of multimodal fusion strategies, and the architectures currently producing state-of-the-art results, closing each section with a short synthesis of what the literature has and has not yet resolved. Section VI then reimplements the reliability-aware fusion mechanism that Section IV identifies as the most theoretically motivated but least completely evaluated strategy, and reports its behaviour against two literature-derived baselines on two conversational

benchmarks. Section VII returns to review, using the case study's own limitations as one source of evidence for six open problems that persist across the field; Section VIII maps each to a research direction, and Section IX concludes.

II. FOUNDATIONS OF AFFECTIVE COMPUTING

A. Emotion Representation and Ground Truth

Two representational schemes dominate the literature. Categorical models describe emotion as membership in a small set of discrete classes, typically happiness, sadness, anger, fear, disgust, and surprise, following Ekman's basic-emotion tradition, are the label format used by most facial and speech emotion recognition datasets [3], [13]. Dimensional models instead place an affective state in a continuous space, most often valence and arousal, and are preferred when an application needs a graded rather than categorical output, such as continuous emotion tracking in video or podcasts [27], [43]. Ground truth is obtained either from the person experiencing the emotion or from an independent observer; the two do not always agree, and most benchmarks default to whichever label their source dataset provides without discussing the distinction explicitly that later sections return to when interpreting benchmark comparisons.

B. Benchmarks and Dataset Quality

Widely used benchmarks include facial-expression corpora such as AffectNet and the FER family, speech corpora built around acted or naturalistic vocal expression, and multimodal conversational corpora that pair audio, video, and text. A recent benchmarking study comparing facial-expression datasets directly found that dataset quality and diversity materially affect downstream performance, which argues for more careful dataset selection [12]. Conversational benchmarks add a further complication: they do not all provide the same modality combinations like AFEW offers video and audio, MELD and IEMOCAP offer audio and text, and CMU-MOSEI offers all three. An architecture validated on one is not automatically comparable to one validated on another [24]. This inconsistency is not a minor bookkeeping issue: it means that headline numbers reported across different papers on different benchmarks cannot be read as a ranking without controlling for the benchmark itself, a point the case study is designed to address directly by evaluating every compared method under one shared protocol.

III. Unimodal Emotion Recognition

A. Speech

Speech emotion recognition infers affect from prosodic and spectral cues such as pitch, energy, and rhythm. Recent systems favour hybrid CNN-recurrent backbones that capture spatial (spectrogram) and temporal structure jointly [10], multi-scale local-global fusion with temporal attention to reconcile short phoneme-level cues with longer emotional trajectories [11], and transfer learning from large pretrained audio networks to reduce dependence on manually engineered features and large labelled corpora [8], [9]. Across this body of work, the consistent trend is a shift away from hand-crafted acoustic features toward learned representations transferred from large pretrained models, which reduces the amount of labelled speech-emotion data a new system needs but introduces a dependency on the domain match between the pretraining corpus and the target application.

B. Facial and Visual Cues

Facial expression recognition remains the most benchmarked unimodal channel, spanning image-based, video-based, and micro-expression variants [13]. Vision-transformer backbones increasingly outperform convolutional baselines on standard benchmarks [14], and dedicated micro-expression work combines transformer attention with graph-convolutional structure to capture the brief facial motion that basic-emotion classifiers tend to miss [15]. The persistent limitation across this literature is the gap between posed and spontaneous expression: models trained predominantly on posed corpora transfer imperfectly to naturalistic footage, and this gap is rarely quantified directly in the papers that report it.

C. Text

Text-based affect recognition has moved beyond lexical sentiment analysis to account for para-linguistic cues in digital communication, most notably emoji use, which recent surveys show measurably improves classification accuracy over lexicon-only approaches when incorporated correctly [16]. Text remains the modality most exposed to sarcasm and cross-cultural ambiguity, since the same lexical content can carry opposite affective meaning depending on context that a text-only model cannot access.

D. Physiological Signals

Physiological recognition, dominated by EEG, is attractive because it resists deliberate suppression and reflects an internal state. Current work pursues two directions at once: deeper spatial graph structure over electrode channels to capture functional brain connectivity [17], [18], and hybridisation with other physiological or behavioural channels, since EEG alone still under-performs richer multimodal combinations in naturalistic settings [19]. This second trend is the clearest unimodal-level evidence for the

paper's broader argument: even within a single sensing category, treating a signal as a component of a larger, complementary system outperforms treating it in isolation.

Table I. Comparative Summary of Unimodal Emotion Recognition Channels

Modality	Representative Signal / Feature	Typical DL Backbone	Key Advantage	Key Limitation	Ref.
Speech	Prosody, spectrogram, pitch/energy contours	CNN-BiGRU, transfer-learned audio transformers	Captures paralinguistic affect independent of lexical content	Sensitive to background noise, speaker/channel variability	[8]-[11]
Facial / Visual	Facial landmarks, action units, micro-expressions	Vision Transformer, CNN + GCN hybrids	Rich, intuitive, widely benchmarked	Occlusion, pose/lighting variation, posed vs. spontaneous gap	[12]-[15]
Text	Lexical sentiment, emoji, discourse cues	Transformer language encoders	Captures explicit self-expression at scale	Sarcasm, cross-cultural ambiguity, context-dependence	[16]
Physiological (EEG, GSR, ECG)	Band power / differential entropy, skin conductance, heart-rate variability	Graph-CNN, spatial-graph transformers	Resistant to deliberate suppression; objective	Obtrusive sensing, strong inter-subject variability	[17]-[19]

IV. MULTIMODAL FUSION STRATEGIES

Combining evidence across channels outperforms unimodal recognition provided the fusion mechanism matches how the modalities relate to one another in time and reliability [3]–[5]. Four broad strategies recur in the literature (Fig. 1), and the choice between them is rarely stated as an explicit design decision in the papers that use them — which motivates treating it as one directly in this review.

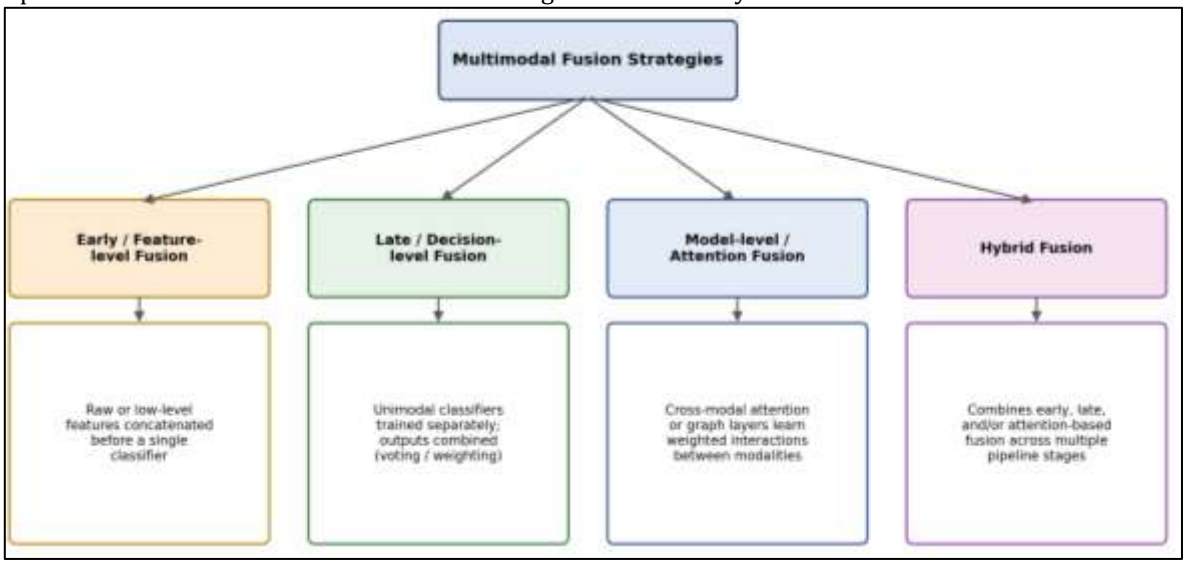

Fig. 1. Taxonomy of the four principal multimodal fusion strategies used in affective computing.

A. Early / Feature-Level Fusion

Early fusion concatenates raw or low-level features from each modality before a single downstream classifier. It gives the model direct access to cross-modal correlations from the outset, but requires the modalities to be temporally aligned and comparably scaled [4].

B. Late / Decision-Level Fusion

Late fusion trains a separate classifier per modality and combines their outputs by voting, weighted averaging, or a shallow meta-classifier. It tolerates modality asynchrony well but cannot model interactions between modalities that only become apparent before the decision stage [4], [20].

C. Model-Level / Attention-Based Fusion

Model-level fusion uses cross-modal attention, transformer co-attention, or graph-attention layers to learn weighted interactions between modality representations during training, instead fixing the combination rule in advance [21]–[23]. Reliability-aware variants extend this by conditioning the fusion weights on an estimate of each modality's trustworthiness for a given input, gating both feature contribution and, in graph-based formulations, edge connectivity itself. Two recent instances of this idea, a reliability-gated cross-attention method for real-time interactive interfaces [49] and a hierarchical alignment-refinement-fusion framework that suppresses redundant non-text information using text as a semantic anchor [50]. It illustrate the approach, combining reliability estimation with graph-based context propagation, examined empirically in later section.

D. Hybrid Fusion

Hybrid strategies combine early, late, and attention-based fusion at different pipeline stages, and are increasingly common in systems combining more than two heterogeneous modalities, since no single fusion point is optimal across all channel pairs at once [22].

E. Synthesis: What the Fusion Literature Has Not Yet Combined

Read together, the fusion literature shows a clear progression from fixed combination rules (early and late fusion) toward learned, adaptive weighting (model-level and hybrid fusion), and a further progression within model-level fusion from uniform attention toward reliability-conditioned attention [49], [50]. What this literature has not yet done, to the authors' knowledge, is extend reliability conditioning from feature-level fusion weights to the topology of a relational graph used for conversational context propagation. The two mechanisms are developed in separate lines of work, one operating on features and the other on graph structure over conversational graph neural networks [25], [26], [48].

Table II. Comparison of Multimodal Fusion Strategies

Strategy	Fusion Point	Strength	Weakness	Ref.
Early / feature-level	Before classification	Preserves fine-grained cross-modal correlation	Requires strict temporal alignment; sensitive to scale mismatch	[4]
Late / decision-level	After per-modality classification	Robust to modality asynchrony and missing channels	Cannot model pre-decision cross-modal interaction	[4], [20]
Model-level / attention	Within the learned representation	Learns adaptive, data-driven cross-modal weighting	Higher model complexity; needs more training data	[21]-[24]
Hybrid	Multiple stages	Flexible; suited to >2 heterogeneous modalities	Increased architectural and tuning complexity	[22]

V. DEEP LEARNING ARCHITECTURES DRIVING RECENT PROGRESS

CNN and recurrent encoders remain the default per-modality backbone for spectrogram-, image-, and time-series-based inputs, often combined so that a CNN captures spatial or spectral structure and a recurrent layer captures temporal context [10]. Transformer self- and cross-attention has largely displaced fixed concatenation as the default cross-modal fusion mechanism, used both to fuse asynchronous, non-paired multimodal streams [20] and to model long-range temporal dependency within a single modality prior to fusion [11]. Graph neural networks have been applied at two levels: intra-modality graphs over EEG electrode channels to capture functional brain connectivity [26], and, more recently, graph-attention layers combined with transformers to jointly model spatial and cross-modal structure for physiological stress recognition [25]. A separate line of graph-based work targets multi-label emotion recognition directly, using dynamic routing-by-agreement to fuse modalities without requiring them to be pre-aligned in time [51]; graph formulations are also used to make classifier decisions more explainable by exposing which nodes drove a given prediction [48].

Selective state-space models are beginning to appear as an efficient alternative to recurrent and transformer encoders for continuous emotion recognition, offering linear-time sequence modelling that

suits the long, continuous recordings typical of physiological and audio-visual affect data; a recent lightweight multimodal audio-visual framework built on this principle reports competitive continuous valence-arousal prediction at markedly lower computational cost [27]. Finally, self-supervised and contrastive pretraining is increasingly used to learn general-purpose affective representations from unlabelled or weakly labelled data before fine-tuning on a target task, reducing dependence on large labelled corpora [28], [29], and recent work has begun testing whether the resulting representations generalise across datasets collected under different protocols [30].

Synthesis: Two Unconnected Design Axes

Sections IV and V, read together, expose two design axes that have each been explored but not yet combined in the reviewed literature: reliability conditioning (which modality to trust, and how much) and structural context propagation (which conversational neighbours to draw on, and via what graph). The graph-based methods surveyed in the above review [25], [26], [48], [51] build relational structure but condition it on similarity or task-specific learned weights, not on an explicit reliability estimate; the reliability-aware methods surveyed [49], [50] condition fusion weights on reliability but do not alter graph structure.

Table III. Representative Recent Deep Architectures for (Multimodal) Emotion Recognition

Architecture Family	Core Mechanism	Modalities Targeted	Reported Strength	Ref.
CNN-BiGRU / hybrid recurrent	Spatial feature extraction + bidirectional temporal encoding	Speech	Captures spatial and temporal speech cues jointly	[10]
Transformer decision fusion	Cross-attention over non-paired/asynchronous streams	Audio, video, text (unpaired)	Removes the need for strictly synchronised training samples	[20]
Graph-Attention + Transformer	Spatial graph over channels + cross-modal attention	Physiological (multi-channel)	Joint spatial and cross-modal structure for stress recognition	[25]
Adaptive multi-scale GNN	Learned spatial graph over EEG electrodes	EEG	Robust electrode-level connectivity modelling	[26]
Dynamic-routing graph attention	Pseudo-alignment + dynamic routing-by-agreement	Text, audio, visual (multi-label)	Fuses modalities without requiring pre-alignment	[51]
Selective state-space (Mamba)	Linear-time selective sequence recurrence	Audio-visual (continuous V-A)	Efficient long-sequence continuous emotion prediction	[27]
Self-supervised / contrastive	Pretext-task or contrastive pretraining	Physiological, EEG, multimodal	Reduced reliance on large labelled corpora; cross-dataset transfer	[28]-[30]

VI. CONFIDENCE-AWARE FUSION: A CASE STUDY

"The preceding synthesis identifies a specific, testable gap: reliability conditioning and graph-based context propagation have each been developed, but not combined into a single mechanism. To test whether combining them is worthwhile, we built a confidence-aware fusion network in which a single per-modality confidence score simultaneously gates cross-attention fusion weights and reweights the construction of a k-nearest-neighbour utterance graph. The architecture encodes each modality with a bidirectional GRU, estimates confidence through a cross-modality self-consistency check, where each modality's score is informed by how well it agrees with the other two, applies confidence-gated multi-head cross-attention, and propagates context through a graph-attention layer whose edges are reweighted by the same confidence scores. We refer to this architecture as CARE-Net. Feature extraction uses a ResNet-18 visual encoder, a CNN over mel-spectrograms for audio, and a BERT-base text encoder, matching the extractors used in the reimplemented baselines; the graph neighbourhood size ($k = 10$) and the fusion-graph balance weight were set by validation grid search and held fixed across all reported experiments.

A. Results Under Clean Conditions

Fig. 2 reports weighted-F1 and weighted accuracy for CARE-Net and the two reimplemented baselines on MELD and CMU-MOSEI under standard, uncorrupted test conditions.

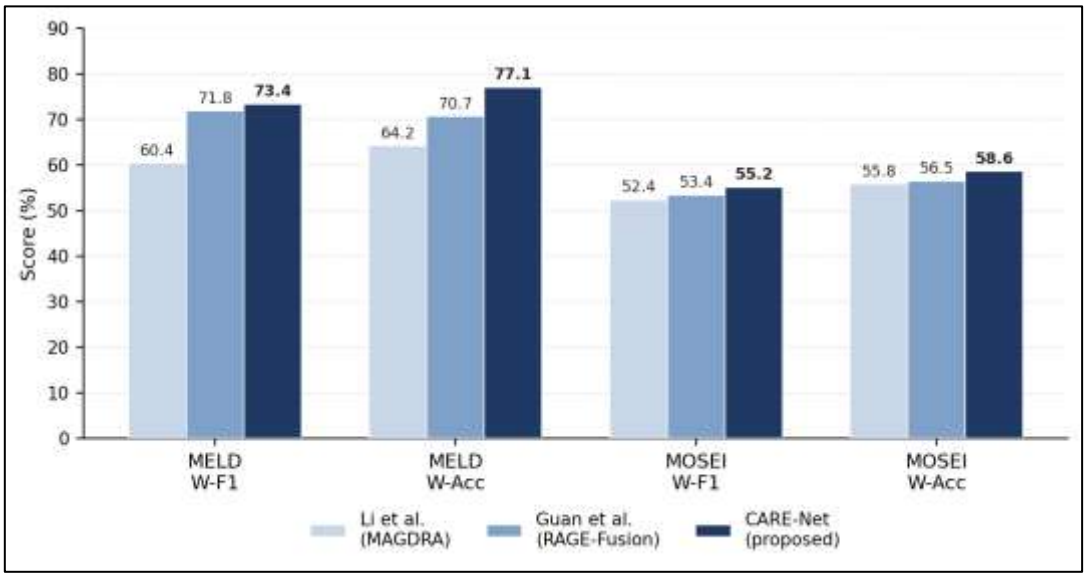

Fig. 2. Weighted-F1 and weighted accuracy on clean MELD and CMU-MOSEI test data (mean over three runs; all methods reimplemented under a shared protocol).

Against the reimplemented reliability-aware fusion mechanism of Guan et al. [49], which estimates per-modality confidence but does not use it to alter graph structure, CARE-Net gains 1.55 points on MELD weighted-F1 under clean conditions, before any modality is corrupted. Because this gap appears with no corruption applied, reliability-conditioned graph construction appears to improve conversational context modelling in its own right, not only as a robustness mechanism. Against the reimplemented dynamic-routing graph mechanism of Li et al. [51], which constructs a static, similarity-only graph without reliability conditioning, the gain widens to 13.0 points on MELD weighted-F1, reflecting the combined effect of confidence-gated fusion and reliability-aware graph construction over a graph that cannot adapt to which utterances are actually trustworthy. The MELD weighted-accuracy gap (77.1% vs. 70.65%) is larger still, while the MOSEI gains are smaller in absolute terms (55.2% vs. 53.4% weighted-F1), consistent with MOSEI's binary sentiment task leaving less headroom than MELD's seven-class emotion task.

B. Per-Class Performance

Weighted-F1 aggregates across classes and can mask uneven performance on individual emotion categories, so Table IV reports per-class F1 on MELD for CARE-Net against the reimplemented mechanisms of Guan et al. [49] and Zhong et al. [50].

Table IV. Per-Class F1 Scores on the MELD Test Set (Reimplemented Baselines)

Method	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger
Zhong et al. [50] (reimpl.)	71.2	54.3	21.4	43.8	52.6	18.9	47.1
Guan et al. [49] (reimpl.)	70.9	55.1	22.7	44.2	53.0	19.4	47.8
CARE-Net	72.1	57.4	27.3	47.6	55.8	24.1	49.9

The largest per-class gains over the stronger reimplemented baseline are on Fear (+4.6 points) and Disgust (+4.7 points), the two lowest-frequency categories in MELD. This is consistent with the confidence-gated mechanism: by suppressing unreliable modality contributions utterance by utterance, the model avoids the noise amplification that fixed-weight fusion tends to introduce on rarer classes. Majority classes such as Neutral gain less (+1.2 points) but still improve, so the minority-class gains are not made at the expense of majority-class accuracy.

C. Robustness under Modality Corruption

Table V reports weighted-F1 on MELD under single-modality zero-masking at five corruption rates, with audio as the corrupted channel; corrupting the visual or text channel instead produces a qualitatively similar pattern.

Table V. Weighted-F1 on MELD under Audio Corruption at Increasing Rates (Reimplemented Baselines)

Method	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5
Li et al. [51] (reimpl.)	57.6	54.1	49.8	44.7	40.3
Zhong et al. [50] (reimpl.)	58.4	55.0	50.7	45.5	44.6
Guan et al. [49] (reimpl.)	69.7	66.9	63.0	58.5	53.7
CARE-Net	72.0	70.7	69.3	68.1	67.1

At the highest corruption rate tested (r = 0.5), CARE-Net's weighted-F1 is 67.1%, a fall of 6.3 points from its clean-condition score of 73.4%. The reimplemented mechanism of Guan et al. [49], the strongest baseline, falls by 18.2 points over the same range, and the reimplemented mechanism of Li et al. [51] falls by 20.1 points — roughly three times CARE-Net's degradation. This flatter curve indicates that confidence-conditioned graph rewiring adds robustness beyond what feature-level reliability gating achieves alone: as audio corruption increases, the graph progressively routes contextual information through utterances with intact audio-visual-text profiles instead of continuing to draw on corrupted neighbours. Because the confidence estimator scores each modality against the other two rather than in isolation, this robustness holds regardless of which specific modality is corrupted, not only for audio.

These results support a narrower claim than "confidence-aware fusion solves multimodal robustness": under one shared reimplementation protocol, on two conversational benchmarks, conditioning both fusion weights and graph topology on the same reliability signal outperforms conditioning either alone, and does so consistently across clean, per-class, and corrupted conditions. Four limitations bound how far this generalises. First, because the baselines are our reimplementations rather than the original authors' released systems, absolute numbers may differ from those reported in [49], [50], and [51]; the relative comparison across methods, evaluated identically, is the evidence this case study offers, not a claim about the original papers' own reported performance. Second, zero-masking and Gaussian noise are controlled approximations of real sensor degradation; gradual audio drift or intermittent transcription failure in the wild may behave differently, so the reported robustness should be read as a best-case estimate. Third, the k-NN graph is constructed over all utterances in a batch, which requires quadratic-time pairwise similarity computation and scales poorly to dialogues substantially longer than those in MELD or CMU-MOSEI — a concrete instance of the real-time deployment problem discussed. Fourth, the neighbourhood size and the fusion-graph balance weight were fixed by validation search and held constant across all experiments; per-dialogue adaptive selection might improve performance further but was not tested here.

VII. OPEN PROBLEMS IN MULTIMODAL AFFECTIVE COMPUTING

F. Missing or Unreliable Modalities

Real-world deployments rarely guarantee that every modality is present and clean at inference time; a sensor may drop out, or a channel may degrade from noise or occlusion. The case study above is a direct test of the current best response to this problem — gating both fusion weights and graph topology by an estimated per-modality confidence score — and its flatter degradation curve under corruption is evidence that reliability modelling meaningfully narrows the gap, even if it does not close it, since the same case study shows performance still declining under corruption, just more slowly [24].

B. Cross-Domain and Cross-Subject Generalisation

Models trained on one subject population, recording protocol, or induction method frequently degrade when applied to another, since physiological and expressive baselines vary substantially between individuals and contexts [32], [33]. Transfer learning is the most common mitigation, adapting a source-domain model to a target domain with limited target data [31], while model-agnostic meta-learning has been proposed to improve inter-subject EEG recognition by learning an 2028ptimization2028n that adapts quickly to a new subject from few calibration samples [32]. Continual-learning formulations that adapt facial-expression models across domain shifts without forgetting earlier domains address the same underlying problem from a different angle [33]. The fixed hyperparameters noted as a limitation of the case study are a specific instance of this broader 2028ptimization2028n problem: a confidence-fusion balance tuned on MELD is not guaranteed to transfer to dialogues of different length or speaker composition.

C. Data Scarcity and Low-Resource Settings

Labelled affective data is expensive to collect at scale, and this scarcity is more acute for under-represented languages, dialects, and physiological modalities than for the English-language, text-audio-video conversational benchmarks that dominate the literature. Generative augmentation with GANs has been used to synthesise additional low-resource training examples, for instance to build Arabic speech-emotion corpora where annotated data are limited [34]. Zero-shot formulations attempt to 2029ptimizat emotion classes never seen during training directly from EEG signals, sidestepping the need for a labelled example of every class [35]. Federated few-shot learning combines data-scarcity mitigation with privacy preservation directly, learning from small amounts of data distributed across many clients corpus [36], a combination that anticipates the privacy constraints discussed next.

D. Privacy, Security, and Federated Learning

Affective signals, particularly physiological and facial data, are sensitive personal information, and 2029ptimization them for training raises privacy and regulatory concerns that are distinct from, but compounded by, the data-scarcity problem above. Two complementary responses have emerged in the recent literature: encrypted inference, exemplified by a transformer-based framework that performs stress detection directly on fully homomorphically encrypted physiological signals without ever decrypting them on the server side [37], and federated learning that keeps raw data on-device and shares only model updates, with recent work using metaheuristic 2029ptimization to tune the federated aggregation process for emotion recognition specifically [38].

E. Limited Explainability

Deep multimodal emotion models remain largely black boxes, and because emotion perception is a subjective, high-level judgement, the case for interpretability is arguably stronger here than in more purely perceptual domains [39]. Recent work addresses this both through reviews cataloguing interpretability techniques specific to affective computing [39] and through methods designed to make emotion-recognition output transparent for high-stakes downstream use such as healthcare and financial decision-making [40]. The confidence scores at the centre of case study are themselves a step toward this goal — they expose, per utterance, which modality the model trusted — but they were evaluated here only for downstream accuracy and robustness, not validated as human-interpretable explanations in their own right.

F. Real-Time and Resource-Constrained Deployment

The quadratic-time graph construction identified as a limitation is a structural property of batch-level k-nearest-neighbour graph construction that any reliability-conditioned graph method would inherit unless it is explicitly addressed. Efficient sequence encoders such as selective state-space models offer linear-time temporal encoding [27], but do not by themselves solve the separate cost of graph construction over long dialogues, which remains an open engineering problem.

Table VI. Open Problems, Root Causes, and Representative Mitigation Strategies

Open Problem	Root Cause	Representative Mitigation	Ref.
Missing / unreliable modalities	Sensor dropout, occlusion, inconsistent modality availability across datasets	Reliability-conditioned gated fusion / graph rewiring	[24], [49]
Cross-domain / cross-subject generalisation	Inter-subject and inter-protocol variability; fixed hyperparameters	Transfer learning; meta-learning; continual learning	[31]-[33]
Data scarcity / low-resource settings	Expensive, small, language- or modality-limited annotated corpora	Generative augmentation; zero-shot learning; federated few-shot learning	[34]-[36]
Privacy and security	Sensitive physiological/facial data; centralised training risk	Homomorphic-encryption inference; federated learning	[37], [38]
Limited explainability	Black-box deep models; subjective, high-level judgement target	Interpretability reviews; transparent/trustworthy AI frameworks	[39], [40]
Real-time / resource-constrained deployment	Offline validation without latency/power constraints; quadratic graph construction cost	Linear-time state-space encoders; efficient quantised or approximate-graph architectures	[27], [37]

VIII. FUTURE RESEARCH DIRECTIONS

A. Reliability-Conditioned, State-Space-Driven, and Scalable Graph Fusion

The reliability-gated fusion examined so far has been paired with a recurrent (Bi-GRU) encoder and a batch-level k-NN graph, both of which limit scalability. Replacing the recurrent encoder with a linear-time selective state-space encoder [27] and replacing exact k-NN graph construction with an approximate or sliding-window alternative would jointly address the real-time deployment cost.

B. Domain-Adaptive and Per-Dialogue Adaptive Fusion

Meta-learning and continual-learning approaches to date have mostly been demonstrated within a single modality, EEG [32] or facial expression [33] individually. Extending them to adapt an entire multimodal fusion pipeline across subjects and domains, and to select fusion hyperparameters per dialogue, remains largely open and follows directly from the fixed-hyperparameter limitation.

C. Federated and Privacy-Preserving Multimodal Learning at Scale

Current federated and encrypted approaches [36]–[38] have been demonstrated on single or paired modalities. Scaling them to full multimodal pipelines — audio, video, text, and physiological signals together — while keeping communication cost and accuracy loss acceptable is a systems-level research direction.

D. Trustworthy, Explainable, and Application-Grounded Affective AI

Interpretability work to date [39], [40], [48] has largely been evaluated in isolation from deployment context. Connecting explainability methods, including the per-utterance confidence signal demonstrated, to the accountability requirements of high-stakes applications such as healthcare [45] and automotive safety [46] would ground future interpretability research in the standards each application domain actually needs to meet.

IX. Conclusion

This paper's primary contribution is a structured review of multimodal affective computing across unimodal recognition, fusion strategy, and architecture, closing with a synthesis that identifies a specific, unaddressed combination in the literature: reliability conditioning and graph-based context propagation have each been developed, but not merged into one mechanism. The case study tests that gap directly, under a shared reimplementation protocol instead of comparing against numbers drawn from disparate original papers, and finds that conditioning both fusion weights and graph topology on the same per-modality confidence signal improves accuracy under clean conditions and substantially flattens the degradation curve under modality corruption. The same case study shows that reliability modelling does not resolve the field's open problems so much as reframe them: synthetic corruption stands in for, but does not equal, real sensor failure; graph-based context propagation scales poorly with dialogue length; and hyperparameters tuned on one benchmark are not guaranteed to generalise to another. Five further problems persist are as follows, cross-domain and cross-subject generalisation, data scarcity, privacy constraints, limited explainability, and the real-time deployment cost that reliability-aware methods themselves introduce. Each maps onto an actively emerging research direction, giving the field a concrete, evidence-grounded agenda for the next generation of multimodal emotion recognition systems.

Conflict of Interest

The authors declare that they have no conflict of interest**.**

Contributions

Saleha Mariyam, Halima Sadia: Conceptualisation, Methodology, Investigation, Writing–original draft; Saleha Mariyam: Data curation, Implementation, Discussion, Writing–review & editing; Halima Sadia: Discussion, Writing–review & editing.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgement

We would like to express our appreciation for the assignment of the Manuscript Communication Number [IU/R&D/2026-MCN0004948] per the guidelines provided by the university's studies and research departments. This identifier facilitates communication and tracking of our research throughout the publication process. We also extend our gratitude to all those who contributed to the development of this work.

References

[1] S. Mariyam and H. Sadia, "A Systematic Review of Multimodal Emotion Recognition Approaches for Affective Computing: Advances and Challenges," in Lecture Notes in Networks and Systems, vol. 1723 LNNS, pp. 334-348, 2026. doi: 10.1007/978-3-032-10783-1_26.
[2] M. B. Maishanu, S. Mariyam, and H. Sadia, "A Deep Multimodal Fusion Framework for Emotion-Aware Adaptive Cinematic Systems: A Systematic Review," in IEEE International Conference on Current Trends in Advanced Computing: Where Technology Meets Expectations, ICCTAC 2026, 2026. doi: 10.1109/ICCTAC68277.2026.11524300.
[3] P. D. Kamble, R. Satpute, and P. Lohakare, "A Comprehensive Survey on Emotion Detection: Modalities, Methods, Applications, and Future Directions," in 6th International Conference on Computer Networks and Inventive Communication Technologies, ICCNCT 2026 - Proceedings, pp. 1954-1959, 2026. doi: 10.1109/ICCNCT68477.2026.11590603.
[4] M. Dhotay, M. Dharrao, S. Deokate, A. Bongale, and D. Dharrao, "Multimodal sentiment analysis: emerging innovations, core challenges, and future directions," Discover Artificial Intelligence, vol. 6, no. 1, 2026. doi: 10.1007/s44163-026-01141-2.
[5] A. Yazici, T. Kucukyilmaz, T. Dokeroglu, A. Sharipbay, M. H. Lee, and B. Tyler, "State-of-the-art Multimodal Emotion Recognition: A comprehensive survey and taxonomy," Intelligent Systems with Applications, vol. 30, 2026. doi: 10.1016/j.iswa.2026.200642.
[6] T. Chen, F. Wang, and H. Zhang, "A Review of Affective Computing in Human-Computer Interaction Design," International Journal of Data Warehousing and Mining, vol. 22, no. 1, 2026. doi: 10.4018/IJDWM.402195.
[7] K. Shingjergji, D. Iren, C. Urlings, and R. Klemke, "Affective computing in online higher education: A systematic literature review," Computers and Education: Artificial Intelligence, vol. 10, 2026. doi: 10.1016/j.caeai.2025.100499.
[8] M. Al-Asadi, A. A. Hameed, J. H. Lafta, H. L. Hussein, and M. Al-Azzawi, "Comprehensive Analysis of Speech Emotion Recognition: Models, Methods, and Applications in Intelligent Interaction," Studies in Computational Intelligence, vol. 1234, pp. 21-40, 2025. doi: 10.1007/978-981-95-2129-6_2.
[9] Y. Korkmaz, "Speech Emotion Recognition Using Transfer Learning: A Comparative Study," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 16, no. 1, 2026. doi: 10.1002/widm.70073.
[10] L. Yu and S. Chen, "Speech Emotion Recognition Based on CNN-BiGRU with a Three-Branch Parallel Fusion Architecture," IAENG International Journal of Computer Science, vol. 53, no. 7, pp. 2750-2763, 2026.
[11] W. Ma and N. Jin, "Multi-scale local-global fusion network with temporal attention for speech emotion recognition," Neural Computing and Applications, vol. 38, no. 12, 2026. doi: 10.1007/s00521-026-12157-1.
[12] F. X. Gaya-Morey, C. Manresa-Yee, C. Martinie, and J. M. Buades-Rubio, "Evaluating Facial Expression Recognition Datasets for Deep Learning: a Benchmark Study with Novel Similarity Metrics," IEEE Transactions on Affective Computing, 2026. doi: 10.1109/TAFFC.2026.3693316.
[13] A. P. Munanday, N. Sazali, A. S. Veerendra, and K. Kadirgama, "Facial Emotion Recognition: Methods, Applications and Future Challenges," International Journal of Integrated Engineering, vol. 18, no. 4, pp. 22-39, 2026. doi: 10.30880/ijie.2026.18.04.002.
[14] K. Ezzameli and H. Mahersia, "Vision Transformer-Based Facial Emotion Recognition," IAENG International Journal of Computer Science, vol. 53, no. 1, pp. 410-423, 2026.
[15] L. Zhu, C. Hu, H. Zhao, and H. Yu, "Integrating MHSSA-transformer and GCN for enhanced micro-expression recognition," Visual Computer, vol. 42, no. 4, 2026. doi: 10.1007/s00371-026-04391-4.
[16] A. Tangestani, A. Aghajani, and R. Mohammadibaghmolaei, "Enhancing Sentiment Analysis with Emojis: A Comprehensive Survey of Methods, Applications, and Future Directions," in 2026 International Biennial Conference on Advances in Artificial Intelligence and Data Science, IBAAIDS 2026, 2026. doi: 10.1109/IBAAIDS66771.2026.11567417.
[17] X. Li, Z. Yan, P. Gong, D. Lin, and F. Chen, "Research progress on emotion recognition based on electroencephalogram signals," Biomedical Signal Processing and Control, vol. 113, 2026. doi: 10.1016/j.bspc.2025.109188.
[18] E. Gkintoni, A. Aroutzidis, H. Antonopoulou, and C. Halkiopoulos, "From Neural Networks to Emotional Networks: A Systematic Review of EEG-Based Emotion Recognition in Cognitive Neuroscience and Real-World Applications," Brain Sciences, vol. 15, no. 3, 2025. doi: 10.3390/brainsci15030220.
[19] B. Hatipoglu Yilmaz, C. M. Yilmaz, and C. Kose, "EEG-based fusion approaches in multimodal emotion recognition: An in-depth review," Neurocomputing, vol. 666, 2026. doi: 10.1016/j.neucom.2025.132235.

[20] M. K. Ahmed and M. A. Ahmed, "Transformer-Based Adaptive Decision Fusion for Non-Paired Multimodal Emotion Recognition," in ISADES 2026 - 2nd International Symposium on AI-Driven Engineering Systems, Proceedings, 2026. doi: 10.1109/ISADES69945.2026.11608092.
[21] A. Singh and C. Dhiman, "A Unified Approach for Multimodal Emotion Recognition Using Counterfactual Learning," Signal Processing: Image Communication, vol. 148, 2026. doi: 10.1016/j.image.2026.117625.
[22] V. Shireesha, V. K. Pamula, and P. S. Murty, "Learning Emotions Across Modalities Through Hybrid Deep Representation Fusion," in 7th International Conference on Innovative Trends in Information Technology, ICITIIT 2026, 2026. doi: 10.1109/ICITIIT68860.2026.11499459.
[23] E. Wang, S. Sun, K. Xue, Y. Ji, T. Ma, K. Jiang, and J. Liu, "A survey of representation learning for multi-modal emotion recognition," Multimedia Tools and Applications, vol. 85, no. 8, 2026. doi: 10.1007/s11042-026-21849-8.
[24] A. Nazreen Banu, G. Lekha, S. R. Lavanya, S. Keerthana, and V. Rahul Chiranjeevi, "Cross-Modal Gated Fusion with Reliability Modeling for Multimodal Emotion Recognition," in Proceedings of 6th International Conference on Expert Clouds and Applications, ICOECA 2026, pp. 1490-1496, 2026. doi: 10.1109/ICOECA68095.2026.11485559.
[25] R. Guo, B. W. Y. Kelana, A. E. safar, J. Lian, S. Meng, and L. Yang, "Graph attention networks meet transformers: A synergistic approach for multimodal stress recognition from physiological signals," Biomedical Signal Processing and Control, vol. 113, 2026. doi: 10.1016/j.bspc.2025.109206.
[26] Z. Zhou, Y. Song, H. Gu, and Y. Ji, "Adaptive Graph Neural Networks with Multi-Scale Attention for EEG-Based Emotion Recognition," in Proceedings of 2025 IEEE 26th China Conference on System Simulation Technology and its Applications, CCSSTA 2025, pp. 814-819, 2025. doi: 10.1109/IEEECONF65522.2025.11137096.
[27] Y. Liang, F. Liu, Y. Yao, M. Liu, and J. Yuan, "LightMamba: A Multimodal Audio-Visual Framework for Continuous Emotion Recognition," in Proceedings - 2025 China Automation Congress, CAC 2025, pp. 6699-6704, 2025. doi: 10.1109/CAC67268.2025.11487044.
[28] M. Sharma, D. Desai, G. Sri, and P. Kumaran, "Multi-Modal Emotion Recognition on Physiological Signals Using Self-Supervised Deep Learning Models," in 7th International Conference on Innovative Trends in Information Technology, ICITIIT 2026, 2026. doi: 10.1109/ICITIIT68860.2026.11499635.
[29] Y. Chen, S. Zhao, S. Li, and G. Pan, "EMOD: A Unified EEG Emotion Representation Framework Leveraging V-A Guided Contrastive Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 21, pp. 17427-17435, 2026. doi: 10.1609/aaai.v40i21.38796.
[30] Y. S. Can, M. Benouis, and E. Andre, "Cross-Dataset Generalizability Analysis of Multimodal Self-Supervised Learning for Stress Recognition Across Lab and Daily Contexts," IEEE Access, vol. 14, pp. 35930-35943, 2026. doi: 10.1109/ACCESS.2026.3670764.
[31] M. B. Sahaai, D. R. A. Kumar, A. Priyadharshini, and S. Gokulraj, "Transfer Learning for Affective Computing: Challenges and Opportunities," Affective Artificial Intelligence: Concepts, Applications, and Case Studies, pp. 265-283, 2026.
[32] C. Chen, H. Fang, Y. Yang, and Y. Zhou, "Model-agnostic meta-learning for EEG-based inter-subject emotion recognition," Journal of Neural Engineering, vol. 22, no. 1, 2025. doi: 10.1088/1741-2552/ad9956.
[33] R. S. Maharjan, L. Bonicelli, M. Romeo, S. Calderara, A. Cangelosi, and R. Cucchiara, "Continual Facial Features Transfer for Facial Expression Recognition," IEEE Transactions on Affective Computing, vol. 16, no. 3, pp. 2352-2364, 2025. doi: 10.1109/TAFFC.2025.3561139.
[34] H. Alshaya and W. Bouchelligua, "GAN-AraEmo: Generative Data Augmentation for Low-Resource Arabic Speech Emotion Recognition," Arabian Journal for Science and Engineering, vol. 51, no. 15, pp. 18205-18233, 2026. doi: 10.1007/s13369-026-11328-5.
[35] N. V. Babu and E. G. M. Kanaga, "Zero-Shot Learning for Multi-Class Emotion Recognition from EEG Signals Using Transformer-Based Models," International Journal of Pattern Recognition and Artificial Intelligence, vol. 40, no. 6, 2026. doi: 10.1142/S0218001426520026.
[36] G. Jyothi and P. Nirmala, "Privacy-Preserving Facial Emotion Recognition Using Federated Few-Shot Learning," in 2026 IEEE International Conference on Emerging Computing and Intelligent Technologies, ICoECIT 2026, 2026. doi: 10.1109/ICoECIT68303.2026.11497909.
[37] J. Xiong, J. Chen, J. Lin, D. Jiao, C. Su, and W. Meng, "StressSentry-FHE: A Transformer-Based Privacy-Preserving Framework for Stress Detection Using Quantized Attention," in Lecture Notes in Computer Science, vol. 15920 LNAI, pp. 300-311, 2026. doi: 10.1007/978-981-95-3052-6_23.
[38] M. S. Alfaras, O. Karan, S. Kurnaz, and A. K. Turkben, "Fed-HOER: Federated Hybrid-Optimized Emotion Recognition Framework Using DBO-FLA Metaheuristic Optimization," Computers, Materials and Continua, vol. 88, no. 2, 2026. doi: 10.32604/cmc.2026.079577.

[39] X. Zhang, T. Zhang, L. Sun, J. Zhao, and Q. Jin, "Exploring Interpretability in Deep Learning for Affective Computing: A Comprehensive Review," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 21, no. 7, 2025. doi: 10.1145/3723005.
[40] S. Farjana Farvin and T. Vigneswari, "Explainability and Interpretability in Affective AI Methods for Making AI-Driven Emotion Recognition Transparent and Trustworthy," Affective Artificial Intelligence: Concepts, Applications, and Case Studies, pp. 285-309, 2026.
[41] H. Chen, W. Zeng, C. Chen, L. Cai, F. Wang, Y. Shi, L. Wang, W. Zhang, Y. Li, H. Yan, W. T. Siok, and N. Wang, "EEG Emotion Copilot: Optimizing lightweight LLMs for emotional EEG interpretation with assisted medical record generation," Neural Networks, vol. 192, 2025. doi: 10.1016/j.neunet.2025.107848.
[42] A. Sindhu, S. K. Sridhar, S. Kandasamy, S. Yashpal Gupta, A. Suresh, and Esakkiammal, "EEG-Based Emotional State Detection Using Transformer Architectures and Large Language Models," in GSEACT 2026 - 2026 IEEE Global Symposium on Emerging and Communication Technologies, 2026. doi: 10.1109/GSEACT68539.2026.11620396.
[43] A. Das, C. Franzreb, T. Polzehl, and S. Möller, "Exploring Foundation Model Fusion Effectiveness and Explainability for Stylistic Analysis of Emotional Podcast Data," in Lecture Notes in Networks and Systems, vol. 1283 LNNS, pp. 370-386, 2025. doi: 10.1007/978-3-031-84457-7_23.
[44] R. Z. Cabada, M. L. B. Estrada, A. G. Robles, V. M. B. Beltrán, and M. Graff, "Generative AI for Automatic Personality Recognition," in Communications in Computer and Information Science, vol. 2552 CCIS, pp. 83-93, 2025. doi: 10.1007/978-3-031-97907-1_7.
[45] A. S. Morales, T. D. L. Reis, A. R. Panisson, F. Ourique, and I. G. Sene Jr., "Affective Intelligent Systems in Healthcare: A Systematic Review," Technologies, vol. 14, no. 3, 2026. doi: 10.3390/technologies14030188.
[46] Y. Z. Wu, W. B. Li, Y. J. Liu, G. Z. Zeng, C. M. Li, H. M. Jin, S. Li, and G. Guo, "AI-enabled intelligent cockpit proactive affective interaction: middle-level feature fusion dual-branch deep learning network for driver emotion recognition," Advances in Manufacturing, vol. 13, no. 3, pp. 525-538, 2025. doi: 10.1007/s40436-024-00519-8.
[47] G. Boateng, X. Zhao, M. Speichert, E. Fleisch, J. Lüscher, T. Pauly, U. Scholz, G. Bodenmann, and T. Kowatsch, ""Are You Okay, Honey?": Recognizing Emotions Among Couples Managing Diabetes in Daily Life Using Multimodal Real-World Smartwatch Data," Sensors, vol. 26, no. 10, 2026. doi: 10.3390/s26103141.
[48] F. Oueslati, S. Bahroun, and E. Zagrouba, "EMO-GNN: Graph Neural Networks for Explainable Mono-and-Multi-label Emotion Detection," in Lecture Notes in Computer Science, vol. 16827 LNCS, pp. 199-213, 2027. doi: 10.1007/978-3-032-31936-4_14.
[49] Y. Guan, Y. Zhu, and Y. Jinho, "RAGE-Fusion: Reliability-Aware Multimodal Emotion Fusion for Real-Time Interactive Interfaces," J. Comput. Biomed. Inform., vol. 10, no. 2, 2026. doi: 10.56979/1002/2026/1271.
[50] B. Zhong, J. Li, Y. Xu, A. Li, and R. Zhang, "HMEPR: A Hierarchical Multimodal Emotion Recognition Framework With Redundancy-Aware Feature Refinement," IEEE Access, vol. 14, pp. 75735-75748, 2026. doi: 10.1109/ACCESS.2026.3692217.
[51] X. Li, J. Liu, Y. Xie, P. Gong, X. Zhang, and H. He, "MAGDRA: A Multi-modal Attention Graph Network with Dynamic Routing-By-Agreement for multi-label emotion recognition," Knowl.-Based Syst., vol. 283, p. 111126, Jan. 2024. doi: 10.1016/j.knosys.2023.111126.