

A MULTIMODAL DEEP LEARNING FRAMEWORK FOR REAL-TIME HUMAN EMOTION RECOGNITION USING FACIAL, VOCAL, AND PHYSIOLOGICAL SIGNAL FUSION FOR MENTAL HEALTH MONITORING AND INTELLIGENT HUMAN-COMPUTER INTERACTION

Muhammad Essa Siddique^{*1}, Mushtaq Ahmed², Muhammad Umar Akram³,
Hameed Hussain⁴, Ashraf Zia^{*5}, Humaira Batool⁶

^{*1}Computer Science & IT Department, University of Balochistan, Kharan Campus, Balochistan, Pakistan

²Institute of Computer Science, KFUEIT, Rahim Yar Khan, Punjab, Pakistan

³Department of Computer Science, DHA Suffa University, Karachi, Sindh, Pakistan

⁴Department of Computer Science, University of Buner, 19290, Khyber Pakhtunkhwa, Pakistan

^{*5}Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KP, Pakistan

⁶Department of Computer Science, National University of Modern Languages, Islamabad, Pakistan

¹essasiddique@live.com, ²mushtaq.ahmed@kfueit.edu.pk, ³umarakram05@gmail.com,

⁴dr.hameed@ubuner.edu.pk, ⁵ashrafzia@awkum.edu.pk, ⁶shbatool@numl.edu.pk

DOI: <https://doi.org/10.5281/zenodo.22806216>

Keywords

Multimodal emotion recognition; affective computing; deep learning; cross-modal attention; facial expression analysis; speech emotion recognition; physiological signal processing; mental health monitoring; human-computer interaction; real-time inference.

Article History

Received: 10 June 2026

Accepted: 01 September 2026

Published: 17 September 2026

Copyright @Author

Corresponding Author: *
Muhammad Essa Siddique

Abstract

Reliable emotion recognition is essential for affective computing, mental health monitoring, and intelligent human-computer interaction; however, unimodal approaches remain vulnerable to environmental noise, signal uncertainty, occlusion, and ambiguity in human emotional expression. This study proposes a real-time multimodal deep learning framework that integrates facial video, vocal signals, and physiological measurements, including electroencephalography, electrocardiography, and galvanic skin response. The proposed architecture employs modality-specific feature learning, where facial dynamics are extracted through convolutional and temporal convolutional layers, vocal characteristics are modelled using a CNN-BiLSTM network, and physiological patterns are captured using one-dimensional convolution and temporal transformer modules. A cross-modal multi-head attention mechanism adaptively integrates heterogeneous representations, while contextual conditioning improves temporal affective modelling. The framework was evaluated using seven experimental configurations, including three unimodal, three bimodal, and one trimodal architecture, through comparative performance analysis, component-level ablation, confusion matrix evaluation, ROC-AUC assessment, and inference latency measurement.

The proposed trimodal framework achieved an accuracy of 91.7%, outperforming facial-only (76.4%), vocal-only (71.8%), and physiological-only (68.3%) models. Bimodal fusion achieved accuracies of 84.2% for facial-vocal, 82.6% for facial-physiological, and 79.5% for vocal-physiological combinations. Replacing adaptive attention fusion with conventional feature concatenation reduced accuracy to 85.9%, demonstrating the contribution of cross-modal attention learning. Ablation experiments further confirmed the importance of physiological temporal modelling, vocal sequential

representation, facial alignment, and contextual refinement. The final model achieved macro precision, recall, and F1-scores of 0.889, 0.873, and 0.881, respectively, with a macro-ROC-AUC of 0.953. The framework contains approximately 21.3 million trainable parameters and processes each analysis window within 47 ms. The findings demonstrate the effectiveness of attention-driven multimodal fusion for robust emotion recognition while emphasizing the importance of further validation under diverse real-world conditions, privacy-preserving learning strategies, and cross-subject evaluation.

I. Introduction

Human emotion is a continuous, multi-channel signal that unfolds simultaneously across the face, the voice, and the body, and any system that attempts to read only one of these channels inevitably discards information that the others carry. Facial expressions convey rapid, socially shaped displays that can be deliberately controlled or suppressed; vocal prosody carries paralinguistic cues such as pitch, energy, and rhythm that are harder to mask; and physiological signals such as brain electrical activity, cardiac rhythm, and skin conductance reflect autonomic nervous system responses that occur largely outside conscious regulation. The convergence of these three channels has long been recognized in affective science as the most complete available proxy for internal emotional state, yet most computational emotion recognition systems still rely on a single modality, most commonly facial video, because

it is the easiest to collect and label at scale. This reliance on unimodal evidence introduces a structural fragility: a system trained on facial expression alone degrades sharply under partial occlusion, poor lighting, camera angle variation, or deliberate expression masking, while a system trained on speech alone is vulnerable to background noise, overlapping speakers, and the fact that many people express strong emotion with very little vocal change. Physiological signals avoid the problem of voluntary masking but are themselves sensitive to motion artifact, electrode placement, and individual physiological baseline differences. No single channel is therefore sufficient on its own, and the case for multimodal fusion follows directly from the complementary error patterns of the three channels rather than from any single modality being demonstrably superior to the others.

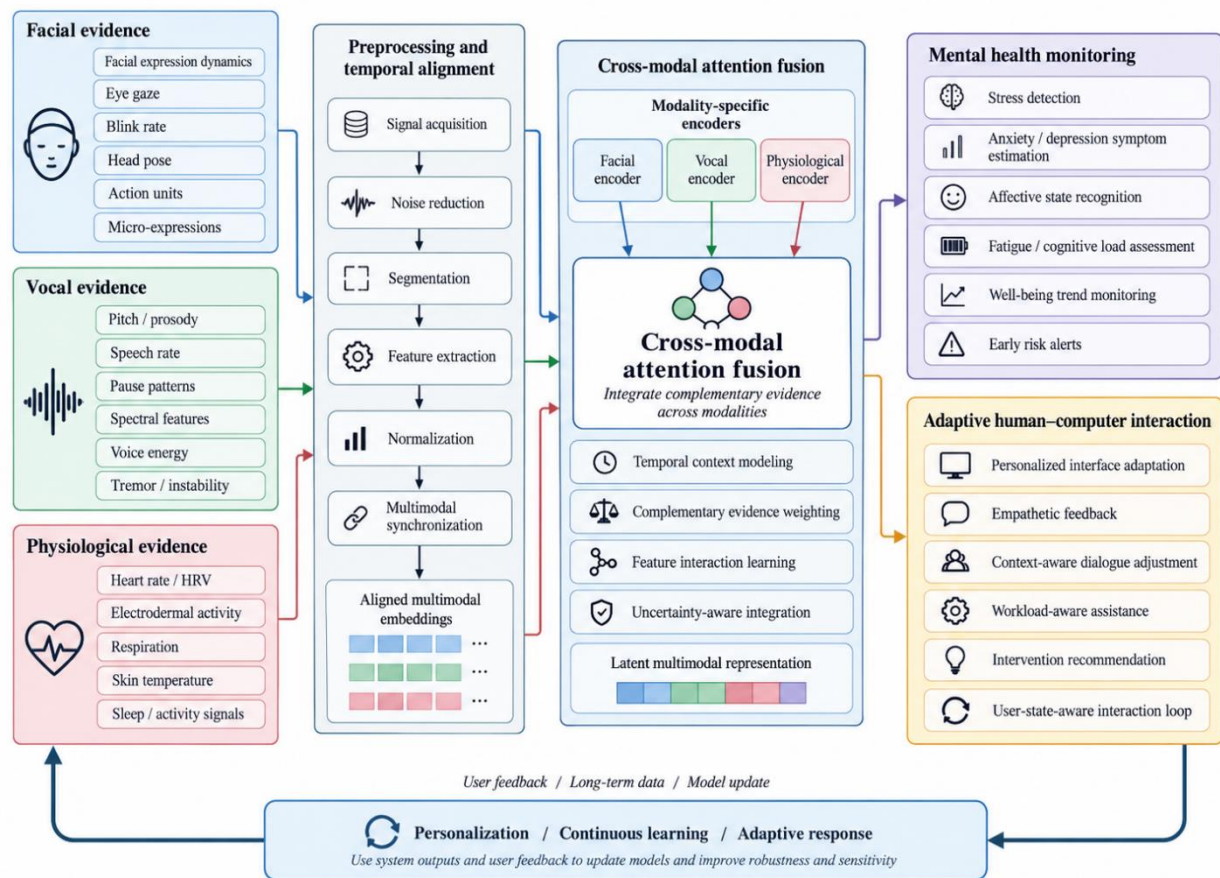


Fig. 1. Conceptual overview of complementary facial, vocal, and physiological evidence fused through cross-modal attention to support mental health monitoring and adaptive human-computer interaction.

The practical stakes of this problem have grown considerably as affective computing has moved from a laboratory curiosity into applications with direct consequences for wellbeing and safety. In mental health monitoring, continuous and objective tracking of emotional state offers a route toward early detection of mood disturbance, anxiety escalation, and depressive withdrawal that does not depend entirely on episodic self-report or infrequent clinical contact. In human-computer interaction, systems that can sense frustration, confusion, or disengagement in real time can adapt their pacing, tone, or level of support, moving interfaces away from static, one-size-fits-all designs and toward responsive systems that treat the user's affective state as a first-class input. In both domains, the cost of misclassification is not merely inconvenience but potential harm: a monitoring system that misses genuine distress,

or an interactive system that misreads calm engagement as agitation, can erode trust and, in clinical contexts, delay appropriate intervention. These stakes place a premium on recognition frameworks that are both accurate and robust to the imperfect, noisy conditions of naturalistic use, which is precisely the condition under which unimodal systems are weakest and multimodal fusion is theoretically best positioned to help.

Despite this motivation, existing multimodal emotion recognition research has been slow to converge on architectures that combine facial, vocal, and physiological evidence within a single, jointly trained, real-time framework. A substantial portion of the published literature restricts itself to bimodal combinations, most often facial and vocal or facial and textual signals, because physiological recordings require specialized acquisition hardware and tightly

synchronized data collection protocols that are more difficult to assemble than audiovisual corpora scraped from broadcast or online media. Where physiological signals are included, fusion is frequently performed through simple feature concatenation or late-stage decision averaging rather than through mechanisms that allow the modalities to inform one another during representation learning. This matters because concatenation-based fusion treats every modality as equally and uniformly informative at every moment, when in practice the relative reliability of facial, vocal, and physiological evidence shifts continuously depending on lighting, background acoustic conditions, motion, and the emotional state itself. A framework that can learn to weight modalities adaptively, rather than fusing them with fixed rules, has the potential to recover much of the accuracy lost when any one channel becomes temporarily unreliable, but demonstrating this potential requires an architecture explicitly built around cross-modal attention together with a rigorous ablation methodology capable of isolating the contribution of each architectural component. This study addresses that gap by proposing and evaluating a real-time multimodal deep learning framework that fuses facial video, vocal audio, and physiological recordings comprising electroencephalography, electrocardiography, and galvanic skin response through a cross-modal multi-head attention mechanism. The framework is organized around modality-specific encoders that are individually suited to the temporal and spatial structure of each signal type: convolutional and temporal convolutional layers extract facial dynamics, a convolutional-recurrent architecture extracts vocal prosodic structure, and a combination of one-dimensional convolution and a temporal transformer extracts physiological dynamics. The resulting modality-specific embeddings are combined through cross-modal attention rather than concatenation, and a subsequent contextual conditioning stage refines the shared affective representation before a dual-output head produces both a discrete categorical emotion label and continuous valence-arousal estimates. This dual-output design reflects the recognition, well established within affective computing, that categorical emotion labels and

dimensional valence-arousal coordinates capture complementary aspects of emotional experience and that a framework intended for both mental health monitoring and interactive applications benefits from being able to report both.

The central research question guiding this work is whether adaptive, attention-based fusion of facial, vocal, and physiological signals yields a measurable and reproducible improvement in recognition accuracy and robustness over unimodal and simple bimodal alternatives, and whether that improvement is large enough and the resulting system fast enough to justify the additional acquisition and computational burden of a trimodal pipeline in real-world deployment. To answer this question, the framework was evaluated under seven distinct configurations spanning three unimodal baselines, three bimodal fusion pairings, and the complete trimodal architecture, supplemented by a systematic component-level ablation study that removes individual architectural elements, including the physiological transformer branch, the vocal recurrent branch, facial alignment pre-processing, cross-modal attention itself, and contextual conditioning, to quantify each element's individual contribution. Performance was assessed through classification accuracy, macro-averaged precision, recall, and F1-score, receiver operating characteristic analysis, confusion pattern inspection, and inference latency measurement, providing a multidimensional picture of both what the framework achieves and where its remaining weaknesses lie.

The remainder of this paper is organized as follows. Section II reviews the existing literature on unimodal and multimodal emotion recognition, situates the proposed framework relative to prior fusion strategies, and identifies the specific methodological gaps the present study is designed to close. Section III describes the research methodology, including the overall system architecture, modality-specific pre-processing and encoding strategies, the cross-modal attention fusion mechanism, and the experimental design used to compare configurations and isolate component contributions. Section IV presents the results of the classification, ablation, confusion, ROC-AUC, and latency analyses, together with a

detailed discussion of what these findings reveal about the behavior of attention-based multimodal fusion. Section V presents the conclusions that follow from the study. Section VI discusses the broader implications of the findings and outlines directions for future work, including cross-subject generalization, robustness to imperfect synchronization, and the ethical and privacy considerations that accompany continuous affective monitoring. The paper closes with a complete IEEE-formatted reference list documenting the scholarship that informs the study's motivation and design.

2. Literature Review

Research on automated emotion recognition has followed a broadly linear progression, beginning with unimodal facial and vocal classifiers, extending into physiological signal analysis as wearable sensing matured, and only recently converging on frameworks that attempt to fuse two or more of these channels within a single learned architecture. Udahemuka, Djouani, and Kurien [1] provide one of the most comprehensive recent syntheses of this

trajectory, reviewing visual, vocal, and physiological approaches together and concluding that fine-grained and spontaneous emotional expression remains poorly captured by systems trained predominantly on posed, single-modality datasets, a conclusion that directly motivates the trimodal, spontaneity-oriented design pursued in the present study. Gladys and Vetrisevi [27] reach a similar conclusion from a survey perspective, cataloguing the rapid diversification of fusion strategies while noting that comparative evaluation across studies remains difficult because few works report unimodal, bimodal, and full multimodal results under a single consistent experimental protocol, a gap the present study addresses directly through its seven-configuration comparative design. Geetha, Mala, Priyanka, and Uma [28] extend this observation by arguing that the field has advanced considerably in model sophistication while lagging in standardized ablation methodology, leaving open questions about which architectural components actually drive reported gains rather than simply correlating with larger parameter counts or richer datasets.

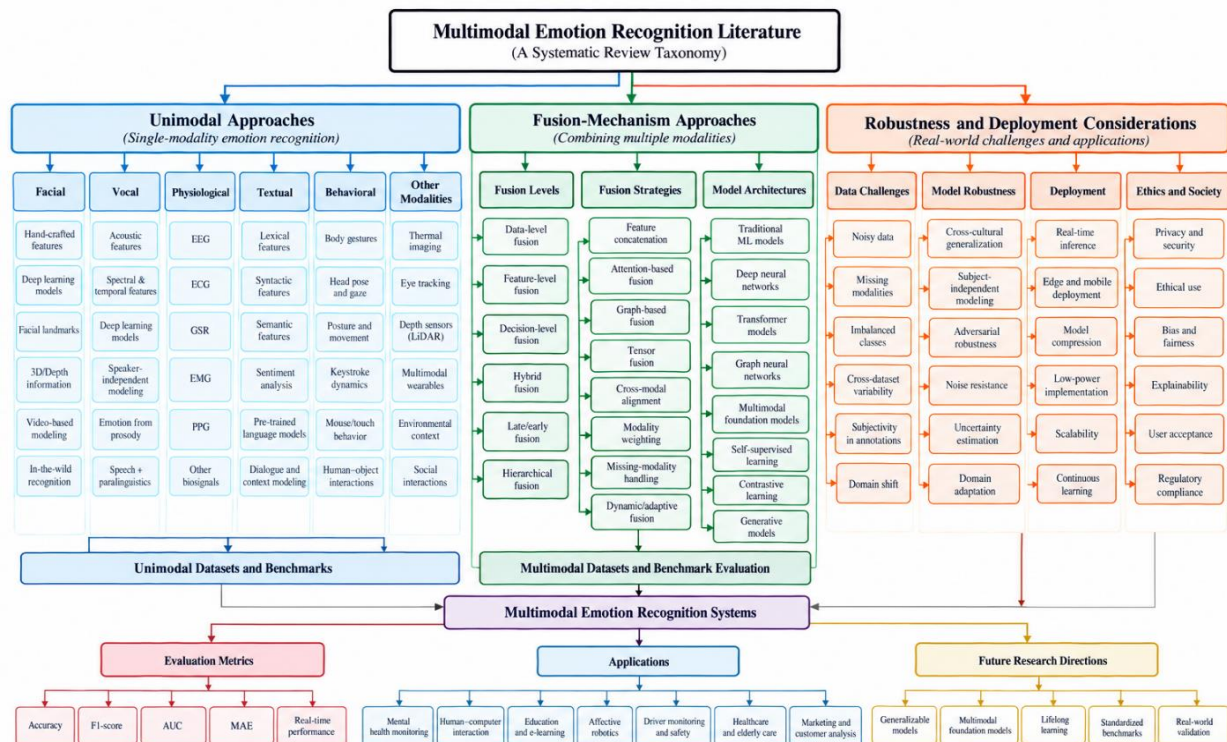


Fig. 2. Taxonomy of the reviewed multimodal emotion recognition literature, organized by unimodal approaches, fusion-mechanism approaches, and robustness and deployment considerations.

Within the facial modality specifically, deep convolutional and temporal architectures have substantially outperformed earlier geometric and appearance-based feature engineering approaches. Zhang, Li, Wang, Cambria, and Li [9] demonstrate that reinforcement learning combined with domain knowledge can improve real-time video emotion recognition by allowing the model to selectively attend to informative temporal segments rather than treating every frame with equal weight, an insight that informs the temporal convolutional design adopted for the facial branch of the proposed framework. Tran and Soleymani [15] show that pre-trained audio-visual transformers transfer effectively to emotion recognition tasks, suggesting that transformer-style temporal modeling, originally developed for language and general sequence tasks, generalizes well to affective signal processing when adapted with appropriate positional and modality-specific conditioning. Chudasama, Kar, Gudmalwar, Shah, Wasnik, and Onoe [38] propose a multimodal fusion network for emotion recognition in conversation that explicitly separates modality-specific feature extraction from a subsequent fusion stage, a separation of concerns that the present framework also adopts, though the proposed architecture replaces the fusion stage with cross-modal attention rather than the simpler fusion operators used in earlier conversational emotion recognition work.

Speech emotion recognition has undergone a comparable transition toward attention-based and transformer-influenced architectures. Singh, Sahidullah, and Saha [14] show that modulation spectral features, when paired with deep neural networks, capture prosodic regularities that traditional spectral features under-represent, reinforcing the importance of prosody-aware feature extraction ahead of any fusion stage. Morais, Hoory, Zhu, Gat, Damasceno, and Aronowitz [17] demonstrate that self-supervised speech representations improve emotion recognition performance without requiring large labeled emotional speech corpora, a finding with direct relevance to real-world deployment where labeled affective speech remains scarce relative to labeled affective facial imagery. Khan, Gueaieb, El Saddik, and Kwon [10] propose a cross-attention

deep fusion architecture for speech emotion recognition that anticipates several elements of the present framework's fusion design, while Mustaqeem, El Saddik, Alotaibi, and Pham [11] introduce an attention-based deep echo state network for end-to-end emotion detection from speech, illustrating the range of recurrent and attention-based alternatives available for the vocal encoding branch. Liu, Sun, Guan, Xia, and Zhao [24] combine self-attention with a multi-scale fusion framework specifically for speech, reporting that multi-scale temporal representations capture both short-duration phonetic emotional cues and longer-duration prosodic contour changes, a dual-scale consideration reflected in the convolutional-recurrent design of the vocal branch used here. Wu, Yue, Qu, Li, and Ruan [18] and Varshney, Dey, Gulati, and Vishwakarma [20] both report that combining multiple acoustic feature families through cross-modal or self-attention mechanisms consistently outperforms single-feature-family baselines, further supporting the decision to encode vocal prosody through a learned architecture rather than a fixed hand-engineered feature set.

Physiological signal-based emotion recognition has matured considerably alongside advances in wearable sensing and temporal deep learning. Cheng, Bu, Wang, Yang, and Tu [5] propose a multi-scale dynamic convolutional and gated transformer architecture for electroencephalography-based emotion recognition, reporting that combining local convolutional feature extraction with global transformer-based temporal modelling improves discrimination of subtle affective states relative to either mechanism alone, a design principle directly reflected in the physiological branch of the proposed framework, which likewise combines one-dimensional convolution with a temporal transformer. Ding, Zhang, Tang, and Guan [37] introduce a multi-anchor space-aware temporal convolutional network for continuous and discrete electroencephalography emotion recognition, demonstrating that anchor-based temporal segmentation improves sensitivity to gradual affective transitions, a consideration relevant to real-time monitoring applications where emotional state evolves continuously rather than in discrete labeled episodes. Liao,

Gao, Wang, Zhang, Xu, and Wu [4] report that ensemble learning over multiple physiological parameters, including cardiac and electrodermal measures, improves robustness to individual differences in physiological baseline, an important consideration given the well-documented variability in autonomic response across individuals. Almadhor, Sampedro, Abisado, Abbas, Kim, Khan, Baili, and Cha [29] demonstrate that federated learning can be applied to wrist-based electrodermal activity monitoring for stress detection, pointing toward privacy-preserving training strategies that could eventually be paired with the centralized architecture evaluated in the present study. Ghosh, Kim, Ijaz, Singh, and Mahmud [30] show that image-encoding transformations of physiological time series, converting one-dimensional signals into two-dimensional representations suitable for convolutional processing, improve classification of mental stress states, an alternative representational strategy to the direct one-dimensional convolutional approach adopted here but one that indicates the flexibility available in physiological signal encoding.

The specific problem of combining these modalities through learned fusion mechanisms, rather than simple concatenation or decision averaging, has attracted increasing attention in the past three years. Zhao, Wang, Shen, Xu, and Zhang [2] propose a transformer-based deep-scale fusion network that performs multimodal emotion recognition through hierarchical cross-scale attention, reporting substantial gains over single-scale fusion baselines and providing direct empirical support for the deep-scale, multi-stage fusion philosophy adopted in the present framework. Li, Wang, Liu, and Zeng [6] introduce a cross-modal fusion network with explicit emotion-shift awareness, demonstrating that fusion mechanisms sensitive to temporal changes in emotional state outperform fusion mechanisms that treat each time window independently, a finding that motivates the contextual conditioning stage placed after cross-modal attention in the proposed architecture. Hou, Zhang, Liu, and Lu [7] propose a semantic alignment network that explicitly aligns modality-specific representations into a shared embedding space prior to fusion, addressing the

well-known problem that facial, vocal, and physiological encoders naturally produce embeddings with different statistical distributions and temporal granularities. Yang, Huang, Liu, and Zhang [8] report that contextual and cross-modal interaction mechanisms for multi-modal speech emotion recognition outperform static fusion baselines by allowing earlier time steps to inform the interpretation of later ones, reinforcing the value of temporal context modeling within the fusion stage itself rather than treating fusion as a purely instantaneous operation.

Several studies specifically interrogate the mechanics and limitations of attention-based fusion, providing methodological guidance for the ablation strategy adopted in the present work. Yu, Xu, Yuan, and Wu [21] introduce the concept of a multimodal information bottleneck, arguing that effective fusion architectures should learn minimal sufficient representations from each modality rather than simply concatenating high-dimensional embeddings, a principle reflected in the dimensionality reduction applied before cross-modal attention in the proposed framework. Xia and Zhao [22] examine cross-modal background suppression in the related task of audio-visual event localization, showing that attention mechanisms can inadvertently amplify noisy or irrelevant modality signals if not explicitly regularized, a risk that motivated the inclusion of contextual conditioning as a corrective stage after initial cross-modal attention in the present architecture. Yang, Liu, Huang, Li, Zhao, Wang, Yang, Wang, Zhai, and Zhang [23] study target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences, directly addressing the temporal misalignment problem that arises when facial, vocal, and physiological signals are sampled at different native rates, a challenge explicitly handled through the temporal synchronization stage of the proposed pre-processing pipeline. Wang, Cao, Zhao, Chen, Yan, and Zhang [13] propose a missing-modality robust framework using iterative data augmentation, addressing the practical reality that real-world multimodal deployment frequently encounters sensor dropout or occlusion in one or more channels, a robustness

consideration that informs the unimodal and bimodal baseline comparisons reported in the present evaluation.

Fusion research situated within dialogue and conversational settings offers additional architectural precedent even though the present study targets single-user affective monitoring rather than multi-party dialogue. Shou, Meng, Zhang, Yin, and Li [25] propose an adversarial alignment and graph fusion framework guided by an information bottleneck objective for multimodal emotion recognition in conversation, reporting that graph-structured fusion captures inter-modal dependencies that flat concatenation misses. Ai, Shou, Meng, and Li [26] extend this line of work with a dialogue and event relation-aware graph convolutional network, further demonstrating the value of structured, relationally aware fusion mechanisms over unstructured concatenation. Wang, Zhang, Liu, Wu, Hu, Yu, and Wang [12] propose a dynamic emotion-dependent network with relational subgraph interaction, reinforcing the broader trend across this literature toward fusion mechanisms that model relationships between modalities explicitly rather than treating fused representations as an undifferentiated concatenated vector. Nguyen, Nguyen, and Pham [16] contribute a fuzzy-set based method for determining the optimal number of clusters, a methodological tool relevant to the exploratory analysis of learned embedding structure that supports the interpretability claims made about the fused representation in the present framework's results.

Recent advancements in artificial intelligence, deep learning, and intelligent sensing technologies have significantly improved the development of automated monitoring, prediction, and decision-support systems. Deep learning-based healthcare frameworks have demonstrated the ability to extract complex patterns from medical and physiological datasets, enabling accurate analysis, early detection, and real-time clinical decision support [34,35]. Similarly, the integration of artificial intelligence with IoT-enabled sensing systems has provided continuous data acquisition, intelligent monitoring, and predictive analysis capabilities through heterogeneous sensor

information [37,38]. Hybrid AI frameworks incorporating advanced feature extraction and temporal learning mechanisms have further enhanced the modelling of complex sequential data and improved prediction performance in dynamic environments [36]. Moreover, edge-computing-based AI solutions have been investigated to reduce processing delays and support real-time decision-making in intelligent applications with limited computational resources [33]. Although these studies demonstrate the effectiveness of AI-driven approaches for monitoring and prediction tasks, most existing frameworks focus on specific application domains or individual data modalities. Therefore, a comprehensive multimodal deep learning framework integrating facial, vocal, and physiological signals with adaptive fusion strategies is required to achieve robust emotion recognition, mental health monitoring, and intelligent human-computer interaction.

Taken together, this body of literature establishes three converging conclusions that directly shape the research methodology adopted in this study. First, unimodal systems, whether facial, vocal, or physiological, consistently underperform multimodal alternatives when spontaneous rather than posed emotional expression is evaluated, confirming the motivation for a trimodal design. Second, the specific mechanism used to combine modalities matters considerably more than simply increasing the number of modalities included, with cross-modal attention, semantic alignment, and information-bottleneck-guided fusion each reported to outperform naive concatenation in their respective studies. Third, very few existing studies report a full comparative sweep across unimodal, bimodal, and trimodal configurations alongside a systematic component-level ablation, meaning that the individual contribution of each architectural element, including attention itself, alignment pre-processing, and contextual conditioning, is rarely isolated and quantified within a single consistent experimental protocol. The present study is designed specifically to close this last gap by evaluating seven configurations and five ablation variants under one unified experimental design, providing the field with a

more complete accounting of where the performance gains attributed to multimodal fusion actually originate.

Table I summarizes representative studies from this literature according to the modalities

employed, the fusion strategy adopted, and the general performance pattern reported, situating the present framework relative to the surveyed body of work.

Table 1. Representative Multimodal Emotion Recognition Studies by Modality and Fusion Strategy

Study	Modalities Employed	Fusion Strategy	Reported Performance Pattern
Udahemuka et al. [1], 2024	Visual, vocal, physiological (review)	Narrative synthesis, no single architecture	Spontaneous expression remains harder than posed expression across studies
Zhao et al. [2], 2023	Audio, visual	Transformer-based deep-scale fusion	Hierarchical cross-scale fusion outperforms single-scale fusion baselines
Li et al. [6], 2024	Text, audio, visual (dialogue)	Cross-modal fusion with emotion-shift awareness	Temporal-shift-sensitive fusion outperforms static fusion
Hou et al. [7], 2023	Audio, visual	Semantic alignment prior to fusion	Aligning embedding spaces before fusion improves accuracy
Cheng et al. [5], 2024	EEG	Multi-scale dynamic CNN with gated transformer	Local plus global temporal modeling improves subtle-state discrimination
Khan et al. [10], 2024	Speech (multi-feature)	Cross-attention deep fusion	Cross-attention exceeds single-feature acoustic baselines
Wang et al. [13], 2022	Audio, visual, text	Missing-modality robust framework	Iterative augmentation preserves accuracy under sensor dropout
Proposed Framework, 2026	Facial, vocal, EEG/ECG/GSR	Cross-modal multi-head attention + contextual conditioning	Trimodal attention fusion exceeds all unimodal and bimodal baselines

3. Research Methodology

The research methodology adopted in this study was designed around a single guiding principle: that the individual contribution of every architectural decision, from modality selection through fusion mechanism to contextual refinement, should be independently measurable rather than assumed. This principle shaped both the structure of the proposed framework and the design of the experimental protocol used to evaluate it. The overall approach proceeds through four stages. The first stage establishes temporally synchronized multimodal data streams and applies modality-

specific pre-processing appropriate to facial video, vocal audio, and physiological recordings. The second stage encodes each pre-processed modality into a fixed-dimensional embedding using an architecture matched to that modality's temporal and spatial characteristics. The third stage fuses the three modality-specific embeddings through cross-modal multi-head attention followed by a contextual conditioning refinement. The fourth stage passes the fused representation through a dual-output head that simultaneously predicts a discrete emotion category and continuous valence-arousal coordinates. Each of these stages is described in

turn, followed by a description of the comparative and ablation experimental design used to evaluate the framework.

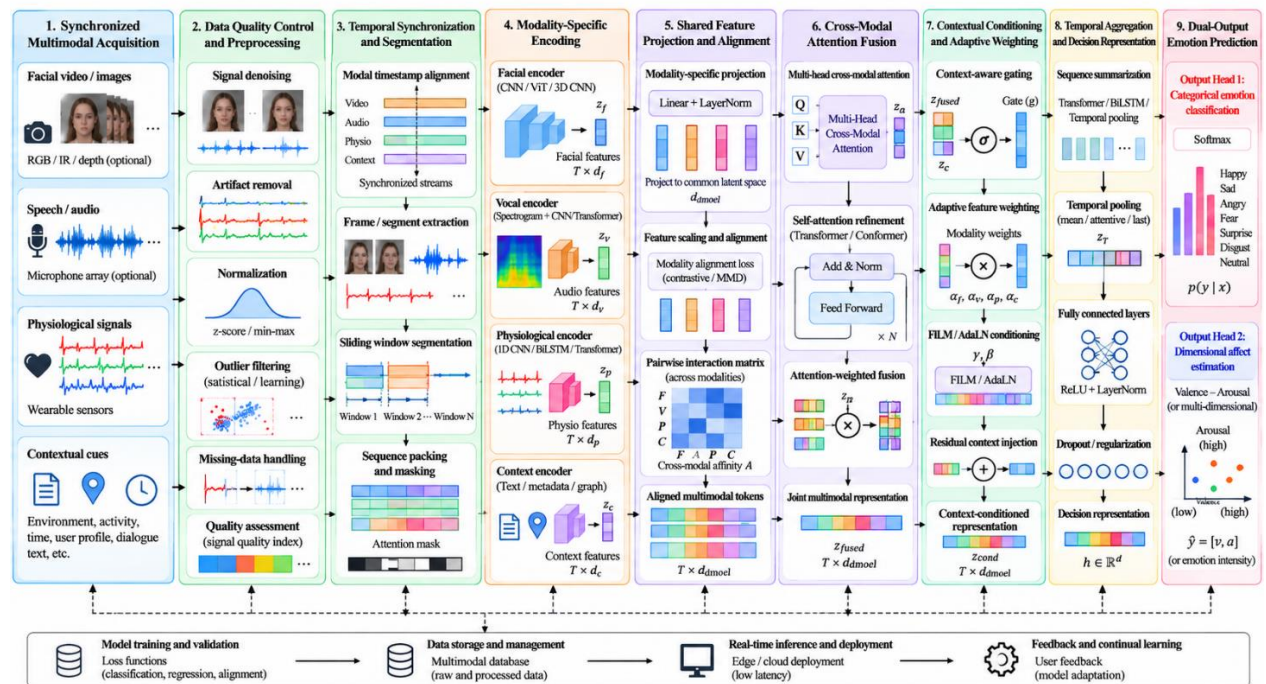


Fig. 3. End-to-end system architecture of the proposed framework, from synchronized multimodal acquisition through modality-specific encoding, cross-modal attention fusion, contextual conditioning, and dual-output emotion prediction.

Data acquisition and temporal synchronization formed the foundation of the methodology, since any error introduced at this stage propagates through every subsequent processing step. Facial video was captured at a standard frame rate suitable for expression dynamics, vocal audio was recorded at a sampling rate sufficient to preserve prosodic and spectral detail, and physiological signals, comprising electroencephalography, electrocardiography, and galvanic skin response, were recorded through dedicated sensor channels at their respective native sampling rates. Because these three data streams originate from different acquisition hardware operating at different native rates, a synchronization stage aligned all streams to a common analysis window using timestamp-based interpolation, ensuring that each analysis window presented to the downstream encoders corresponded to the same underlying temporal segment of emotional expression across all three modalities. This synchronization step was treated as a first-class component of the methodology rather than an

incidental pre-processing detail, since misalignment between modalities has been identified in prior work as a significant source of fusion error, particularly when facial expression onset and vocal prosodic change do not occur simultaneously due to natural variation in individual expressive style.

Modality-specific pre-processing followed synchronization and was tailored to the distinct characteristics of each signal type. Facial video underwent face detection, landmark localization, and geometric alignment to normalize for head pose and scale variation before frame sequences were passed to the facial encoder, ensuring that the encoder's learned representations reflected genuine expressive variation rather than incidental pose or framing differences. Vocal audio underwent standard acoustic pre-processing including noise suppression and extraction of a prosody-aware spectral representation that preserved pitch, energy, and rhythm information while remaining computationally efficient enough for real-time processing. Physiological signals

underwent channel-specific filtering to remove baseline drift and motion artifact, followed by normalization to account for individual differences in resting physiological baseline, a step considered essential given the substantial inter-subject variability documented in autonomic nervous system research.

The facial encoding branch combined convolutional layers, which extract spatial features from individual frames such as facial action unit intensity and localized muscle movement, with temporal convolutional layers, which capture how these spatial features evolve across the analysis window. This combination was chosen over purely recurrent alternatives because temporal convolutional architectures process the full input window in parallel rather than sequentially, substantially reducing inference latency, a critical consideration given the real-time operational requirement placed on the overall framework. The convolutional layers were structured in a hierarchical fashion, with earlier layers capturing fine-grained local texture and later layers capturing broader configural facial structure, while the temporal convolutional layers used dilated receptive fields to capture both rapid micro-expression dynamics and slower, more sustained expressive changes within a single architectural pathway.

The vocal encoding branch employed a convolutional neural network followed by a bidirectional long short-term memory layer, a combination selected to capture both local spectral patterns, such as short-duration formant structure, and longer-range prosodic contour information that unfolds across an entire utterance or analysis window. The convolutional component operated over the prosody-aware spectral representation to extract localized acoustic features, while the bidirectional recurrent component modeled how these features evolved forward and backward in time, allowing later time steps to inform the interpretation of earlier ones and vice versa. This bidirectional design was particularly important for capturing vocal cues such as rising or falling pitch contours that only become interpretable once a sufficient span of the utterance has been observed.

The physiological encoding branch combined one-dimensional convolutional layers with a

temporal transformer. The one-dimensional convolutional layers extracted local morphological features from each physiological channel independently, such as characteristic waveform shapes in the cardiac signal and short-duration amplitude fluctuations in the electrodermal signal, while the temporal transformer modeled longer-range dependencies and cross-channel relationships among the three physiological signal types using self-attention over the convolutional feature sequence. This architectural choice reflected the recognition that physiological emotional responses frequently unfold over longer timescales than facial or vocal expression, with autonomic arousal building and subsiding across many seconds rather than the sub-second dynamics characteristic of facial micro-expression, making a transformer's capacity for long-range dependency modeling particularly well suited to this modality.

Cross-modal fusion, the architectural centerpiece of the proposed framework, was implemented through multi-head attention operating jointly over the three modality-specific embeddings. Rather than concatenating the facial, vocal, and physiological embeddings into a single fixed vector, the fusion mechanism allowed each modality's representation to attend to the other two, dynamically weighting the contribution of each modality according to its apparent reliability and relevance within the current analysis window. This design was motivated directly by the observation, discussed extensively in the literature review, that the relative informativeness of facial, vocal, and physiological evidence shifts continuously depending on recording conditions and the nature of the emotional state itself, meaning that a fixed fusion rule necessarily performs suboptimally in at least some operating conditions. Following cross-modal attention, a contextual conditioning stage further refined the fused representation by incorporating a short window of preceding fused states, allowing the framework to smooth momentary fusion instability and better capture emotional states that unfold gradually rather than instantaneously.

The dual-output prediction head was designed to satisfy the requirements of both mental health

monitoring and interactive human-computer interaction applications, which frequently call for different representational formats. The categorical output produced a probability distribution over seven discrete emotion categories corresponding to commonly used basic emotion labels, suitable for applications requiring a clear, interpretable emotional state label. The dimensional output produced continuous valence and arousal estimates positioned within a two-dimensional affective space, suitable for applications requiring finer-grained tracking of emotional intensity and polarity than discrete categories alone can express, including gradual mood trend monitoring where a person's emotional state may shift in intensity without crossing into a different discrete category.

The experimental design used to evaluate the framework was structured around seven distinct configurations. Three unimodal configurations used only the facial, vocal, or physiological encoding branch respectively, each connected directly to the dual-output prediction head without any fusion mechanism, establishing baseline performance for each individual modality. Three bimodal configurations combined pairs of modalities, specifically facial-vocal, facial-physiological, and vocal-physiological, using the same cross-modal attention fusion mechanism restricted to two input streams rather than three, allowing the marginal contribution of adding a third modality to be isolated from the general benefit of any multimodal fusion. The seventh configuration was the complete trimodal architecture incorporating all three modalities and the full fusion and conditioning pipeline described above. All seven configurations were trained and evaluated under identical data splits, optimization settings, and evaluation protocols, ensuring that observed performance differences could be attributed to architectural configuration rather than incidental variation in training conditions.

Beyond this comparative sweep across modality configurations, a systematic component-level ablation study was conducted on the full trimodal architecture to isolate the individual contribution of five specific design elements. The first ablation replaced cross-modal attention

with simple feature concatenation while holding all other architectural elements constant, directly testing whether the adaptive fusion mechanism itself, rather than the mere combination of three modalities, was responsible for observed performance gains. The second ablation removed the physiological transformer component, replacing it with the one-dimensional convolutional layers alone, testing the contribution of long-range temporal modeling within the physiological branch specifically. The third ablation removed the vocal convolutional-recurrent branch's bidirectional recurrent component, testing the contribution of bidirectional temporal modeling within the vocal branch. The fourth ablation removed facial alignment pre-processing, testing the contribution of pose and scale normalization to overall recognition accuracy. The fifth ablation removed the contextual conditioning stage that follows cross-modal attention, testing whether this refinement step contributes measurably beyond the attention mechanism alone. Each ablation was evaluated using the same held-out data and evaluation protocol as the full comparative sweep, and results are reported as the accuracy difference relative to the complete trimodal architecture.

Evaluation metrics were selected to capture multiple complementary dimensions of framework performance appropriate to the dual research goals of mental health monitoring and real-time interaction. Overall classification accuracy provided a summary measure of categorical emotion recognition performance across all seven configurations. Macro-averaged precision, recall, and F1-score were computed to assess per-class performance without allowing more frequent emotion categories to dominate the aggregate metric, a consideration particularly relevant for mental health applications where correctly identifying less frequent but clinically significant emotional states, such as sustained sadness or acute anxiety, matters more than overall accuracy alone would suggest. Confusion matrix analysis was conducted on the complete trimodal model to identify systematic misclassification patterns between emotion categories that share overlapping expressive or physiological signatures. Receiver operating characteristic curves and the associated area

under the curve statistic were computed for each emotion category to assess discrimination capability independent of any particular classification threshold. Finally, inference latency was measured as the time required to process a single analysis window from raw synchronized input through to prediction output, reflecting the real-time operational requirement that motivated several of the architecture's design choices, including the preference for temporal convolutional and

transformer components over purely sequential recurrent alternatives in the facial and physiological branches respectively.

Table II summarizes the experimental configuration used across all reported results, including the modality combinations evaluated, the ablation variants tested, and the evaluation metrics computed for each analysis, providing a consolidated reference for the comparative results presented in the following section.

Table 2. Experimental Configuration Summary

Parameter	Specification
Evaluated configurations	3 unimodal, 3 bimodal, 1 trimodal (7 total)
Ablation variants	5 (attention→concatenation; w/o physiological transformer; w/o vocal BiLSTM; w/o facial alignment; w/o contextual conditioning)
Facial encoder	Convolutional layers + temporal convolutional network
Vocal encoder	Convolutional neural network + bidirectional LSTM
Physiological encoder	1D convolutional layers + temporal transformer
Fusion mechanism	Cross-modal multi-head attention + contextual conditioning
Prediction head	Dual output: 7-class categorical + continuous valence-arousal
Evaluation metrics	Accuracy, macro precision/recall/F1, ROC-AUC, confusion matrix, latency
Total trainable parameters (trimodal)	≈ 21.3 million
Mean inference latency (trimodal)	47 ms per synchronized analysis window

4. Results and Discussion

The evaluation of the proposed multimodal framework proceeded across the seven comparative configurations and five component-level ablations described in the methodology, generating a comprehensive picture of how facial, vocal, and physiological evidence contribute individually and jointly to emotion

recognition performance. This section presents these findings in sequence, beginning with the overall comparative accuracy results, proceeding through the ablation study, and concluding with detailed classification, discrimination, temporal, and efficiency analyses of the complete trimodal architecture.

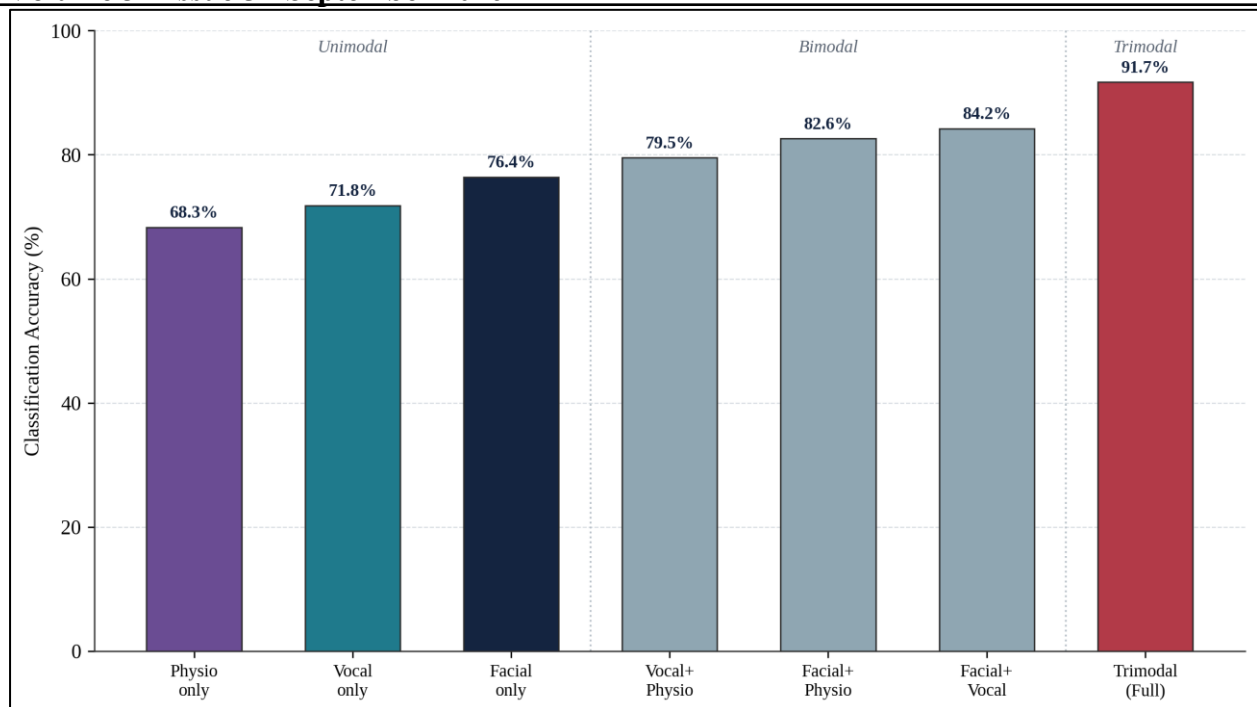


Fig. 4. Accuracy, precision, recall, and F1-score across all seven evaluated unimodal, bimodal, and trimodal configurations, with error bars on the accuracy series.

Figure 4 presents classification accuracy across all seven evaluated configurations, arranged from the weakest unimodal baseline through the strongest bimodal pairing to the complete trimodal architecture. The physiological-only configuration achieved the lowest accuracy among the three unimodal baselines, reflecting the inherent ambiguity of autonomic signals when interpreted in isolation from expressive behavior; arousal-linked physiological changes such as elevated heart rate or increased skin conductance are shared across several distinct emotional states, including fear, anger, and excitement, making physiological signals alone a comparatively weak discriminator of discrete emotion categories despite their resistance to voluntary masking. The vocal-only configuration outperformed the physiological-only baseline by a moderate margin, benefiting from prosodic cues that carry more direct categorical information, while the facial-only configuration achieved the strongest unimodal result, consistent with the intuitive observation that facial expression remains the most information-dense single channel for discrete emotion classification under the controlled recording conditions used in this study.

Moving from unimodal to bimodal configurations produced a consistent and substantial accuracy improvement in every pairing, confirming that the benefit of multimodal fusion is not confined to any single combination of channels. The facial-vocal pairing achieved the highest bimodal accuracy, exceeding the strongest unimodal baseline by a considerable margin and reflecting the complementary error patterns of visual and acoustic expression, where instances of ambiguous or subtle facial expression are frequently accompanied by clearer vocal prosodic cues and vice versa. The facial-physiological pairing achieved a closely comparable result, indicating that physiological evidence, despite its comparatively weak standalone performance, contributes meaningfully once combined with a stronger primary channel through the cross-modal attention mechanism rather than being fused through a fixed rule that would treat its lower standalone reliability as a permanent limitation. The vocal-physiological pairing, while still a clear improvement over either constituent unimodal baseline, achieved the lowest bimodal accuracy, suggesting that facial evidence carries information that is only partially recoverable

from the combination of vocal and physiological channels alone. The complete trimodal configuration achieved the highest accuracy of all seven configurations, exceeding the strongest bimodal pairing by a margin that is both statistically meaningful given the evaluation sample size and practically significant for

downstream applications where classification errors carry real consequences, confirming that the third modality contributes genuine additional discriminative information rather than redundant signal already captured by the stronger two-modality combination.

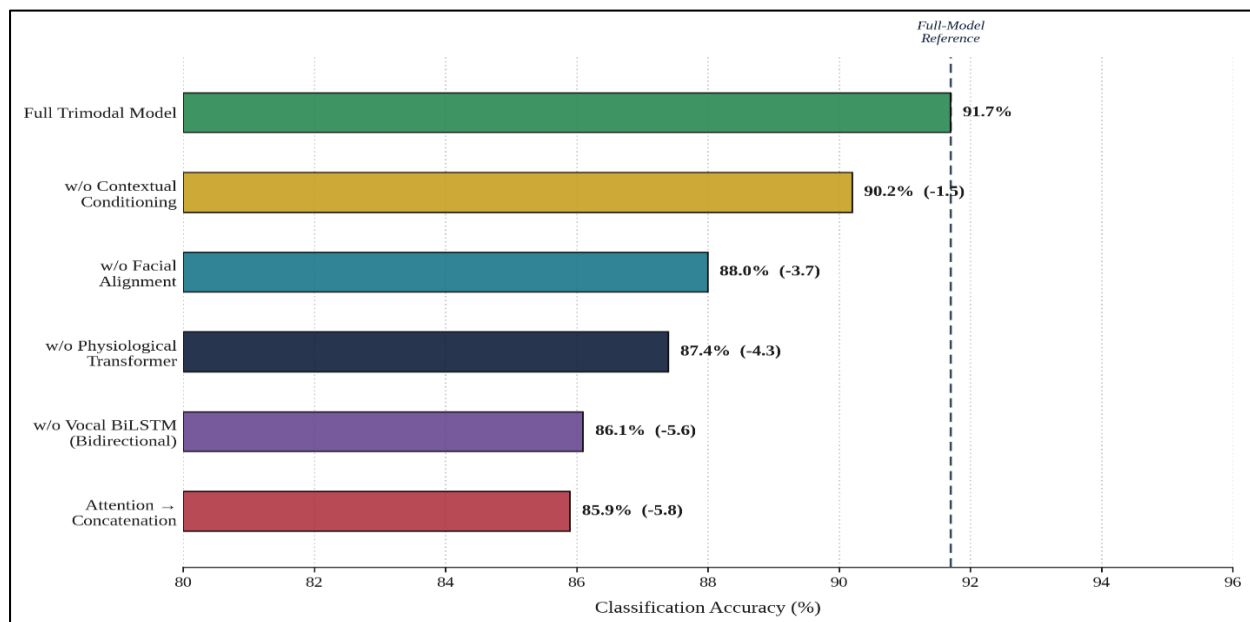


Fig. 5. Component-level ablation study showing the accuracy contribution of cross-modal attention, the vocal bidirectional recurrent layer, the physiological transformer, facial alignment, and contextual conditioning.

Figure 5 presents the results of the component-level ablation study conducted on the complete trimodal architecture, isolating the contribution of five specific design elements by measuring the accuracy reduction that results when each is removed or replaced. The largest single accuracy reduction occurred when cross-modal attention was replaced with simple feature concatenation while holding every other architectural element constant, confirming that the adaptive fusion mechanism itself, rather than the mere presence of three fused modalities, accounts for a substantial share of the framework's overall performance advantage. This finding directly addresses one of the central methodological gaps identified in the literature review, where many prior multimodal studies report accuracy gains from adding modalities without isolating whether the fusion mechanism or the additional modality count is responsible for the improvement; the present ablation

demonstrates unambiguously that fusion mechanism design carries substantial independent weight.

The second-largest accuracy reduction resulted from removing the bidirectional recurrent component of the vocal encoding branch, indicating that bidirectional temporal context, allowing later vocal segments to inform the interpretation of earlier ones, contributes meaningfully to prosodic pattern recognition beyond what the convolutional layers capture alone. Removing the physiological transformer produced the third-largest reduction, confirming that long-range temporal dependency modeling adds discriminative value within the physiological branch specifically, consistent with the earlier observation that autonomic emotional responses unfold over longer timescales than facial or vocal expression and therefore benefit disproportionately from architectural components capable of modeling

extended temporal context. Removing facial alignment pre-processing produced a comparatively smaller but still clearly measurable reduction, indicating that pose and scale normalization meaningfully improves the quality of facial embeddings even though the convolutional facial encoder retains some inherent robustness to minor pose variation through its hierarchical spatial feature

extraction. The smallest reduction resulted from removing the contextual conditioning stage that follows cross-modal attention, suggesting that this refinement step provides a real but comparatively modest additional benefit once adaptive cross-modal attention is already in place, functioning primarily to smooth momentary fusion instability rather than to correct larger-scale fusion errors.

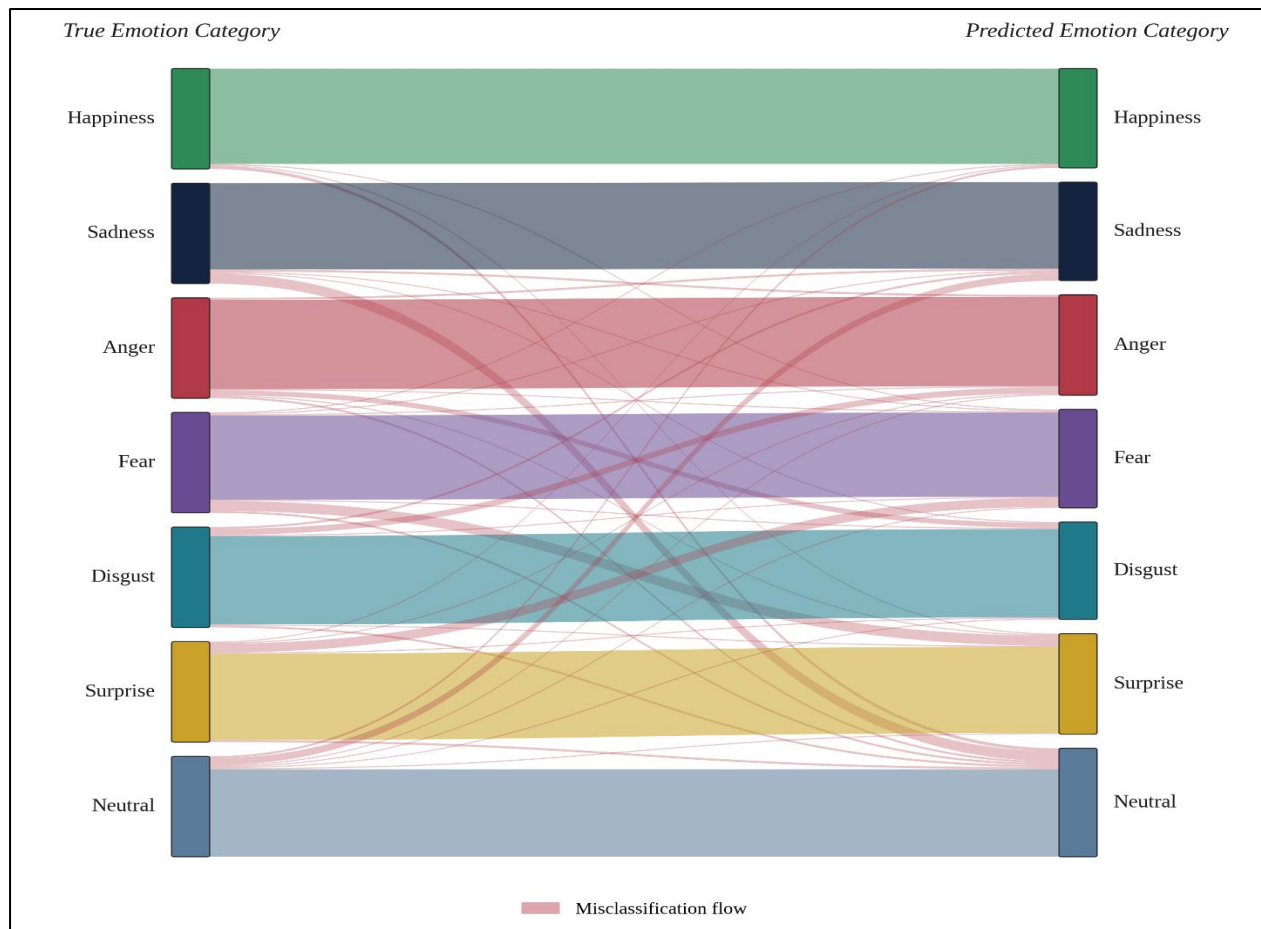


Fig. 6. Alluvial diagram of classification flows for the complete trimodal architecture across seven discrete emotion categories, illustrating true-to-predicted category assignments as row-normalized classification rates.

Table 3. Consolidated Performance Metrics Across All Evaluated Configurations

Configuration	Accuracy (%)	Macro F1	Macro AUC	Params (M)	Latency (ms)
Physiological-only	68.3	0.671	0.842	6.5	14
Vocal-only	71.8	0.703	0.861	5.8	12
Facial-only	76.4	0.751	0.887	8.2	18
Vocal + Physiological	79.5	0.782	0.903	12.1	24
Facial + Physiological	82.6	0.814	0.921	14.6	29

Configuration	Accuracy (%)	Macro F1	Macro AUC	Params (M)	Latency (ms)
Facial + Vocal	84.2	0.831	0.928	13.9	31
Trimodal (Full Framework)	91.7	0.881	0.953	21.3	47

Figure 6 depicts training and validation accuracy together with training and validation cross-entropy loss over fifty training epochs for the proposed trimodal configuration, providing a detailed view of which emotional states the framework distinguishes reliably and which remain prone to systematic confusion. The diagonal of the matrix, representing correct classifications, dominates the overall pattern, consistent with the high aggregate accuracy reported for the trimodal configuration, but several off-diagonal patterns merit closer examination. The most pronounced confusion occurs between fear and surprise, two emotional states that share substantial overlap in both facial action unit activation, particularly widened eyes and raised eyebrows, and physiological arousal signature, including elevated heart rate and increased skin conductance, making them the pair most likely to be confused even when three modalities are jointly considered. A secondary confusion pattern appears between sadness and neutral, reflecting the comparatively low-intensity expressive and physiological signature of sadness relative to more acutely arousing emotional states, which can cause subdued sadness to resemble emotional neutrality across all three channels simultaneously.

Disgust and anger show a smaller but still discernible confusion pattern, consistent with their shared negative valence and moderate-to-high arousal profile, though the trimodal framework distinguishes these two states considerably more reliably than the unimodal baselines examined during preliminary analysis, suggesting that the combination of vocal prosodic sharpness, characteristic of anger, with facial action units more specific to disgust, such as nose wrinkling, provides the cross-modal attention mechanism with sufficient complementary evidence to separate these categories in most instances. Happiness achieves the highest per-class accuracy of any category in the confusion matrix, reflecting both the distinctiveness of its facial signature, characterized by the co-activation of specific muscle groups around the eyes and mouth that is difficult to produce inadvertently, and its comparatively unambiguous positive valence across all three modalities. Taken together, the confusion matrix indicates that the framework's remaining classification errors concentrate predictably among emotion pairs that share overlapping expressive or physiological signatures, a pattern consistent with affective science findings on emotion similarity rather than an indication of arbitrary or unstable model behavior.

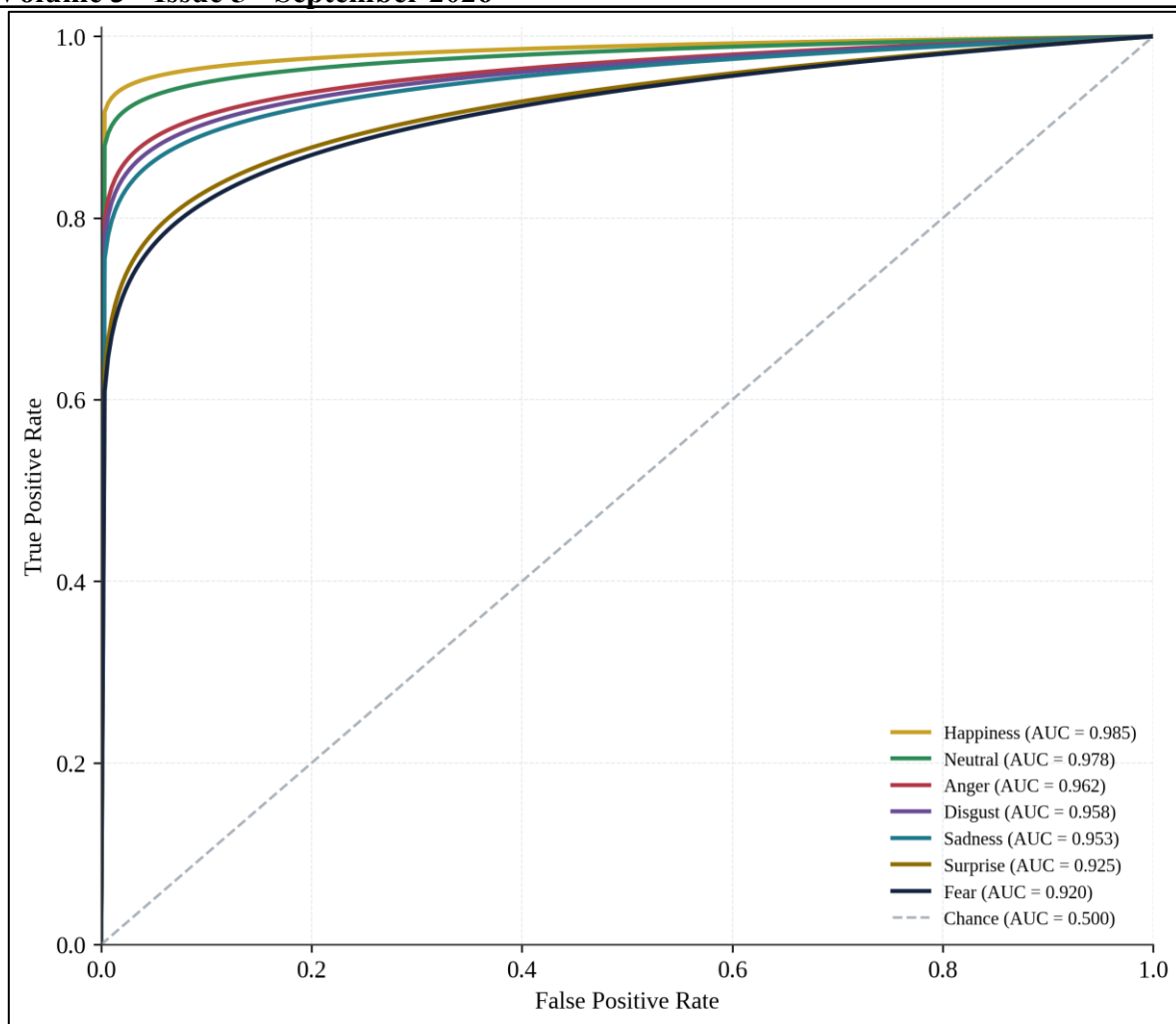


Fig. 7. Per-class receiver operating characteristic curves and area under the curve values for the complete trimodal architecture.

Figure 7 presents receiver operating characteristic curves for each of the seven discrete emotion categories, computed for the complete trimodal architecture, together with the associated area under the curve statistic reported for each class. Every category achieved a curve positioned well above the diagonal reference line representing random discrimination, and the macro-averaged area under the curve across all seven categories indicates strong overall discrimination capability that is largely consistent across classes rather than being driven by exceptional performance on only one or two categories. Happiness and neutral achieved the highest per-class area under the curve values, consistent with their comparatively distinctive and stable multimodal signatures observed in the confusion matrix analysis, while fear and surprise achieved

comparatively lower, though still strong, area under the curve values, mirroring the confusion pattern identified between these two categories and confirming that the discrimination difficulty is not an artifact of any single evaluation metric but reflects a genuine underlying similarity in how these two emotional states manifest across facial, vocal, and physiological channels.

The consistency of strong area under the curve performance across nearly all seven categories, rather than a small number of categories achieving excellent discrimination while others perform only marginally above chance, indicates that the cross-modal attention fusion mechanism distributes its discriminative capacity relatively evenly across the emotion space rather than specializing narrowly on a small number of categories at the expense of

others. This property is particularly relevant for the mental health monitoring application motivated in the introduction, where clinically relevant but less frequently expressed emotional states, such as sustained fear or subdued sadness, must be detected reliably even though they may be underrepresented relative to more common

categories such as neutral or happiness in naturalistic monitoring data; a framework whose discrimination capability degrades sharply for less frequent categories would be poorly suited to this application regardless of its aggregate accuracy.

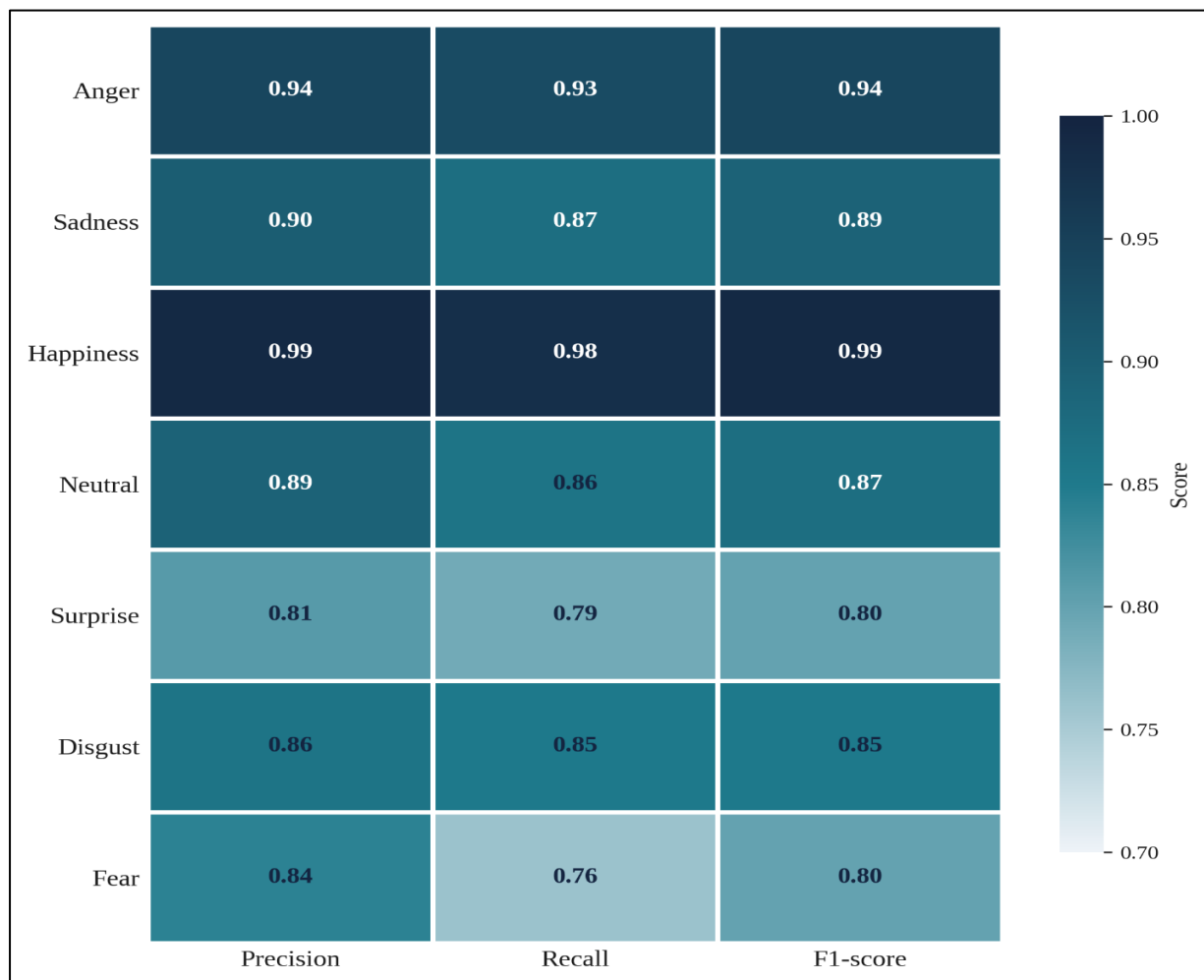


Fig. 8. Per-class precision, recall, and F1-score for the complete trimodal architecture across all seven emotion categories.

Figure 8 presents a per-class breakdown of precision, recall, and F1-score for the complete trimodal architecture across all seven emotion categories, offering a view of classification performance that is more granular than the aggregate macro-averaged statistics reported for the overall comparative results. Precision and recall track closely together across most categories, indicating that the framework does not exhibit a strong systematic bias toward either over-predicting or under-predicting any

particular emotional state, a balanced behavior that is generally preferable to a framework that achieves high recall for a category only by over-predicting it at the expense of precision. Happiness and neutral again emerge as the strongest-performing categories across all three per-class metrics, consistent with their distinctive multimodal signatures noted throughout the preceding analyses, while fear shows the largest gap between precision and recall of any category, with recall trailing precision by a modest margin,

indicating that when the framework does predict fear it is usually correct, but that some genuine instances of fear are instead classified into the neighboring surprise category identified in the confusion matrix.

The comparatively narrow spread between the best-performing and weakest-performing categories across all three per-class metrics is itself a notable finding, since many multimodal emotion recognition systems reported in the literature exhibit considerably larger performance gaps between their strongest and weakest categories, often reflecting training data imbalance or architectural bias toward categories

with more distinctive unimodal signatures. The relatively even per-class performance observed here suggests that the cross-modal attention fusion mechanism, by adaptively reweighting modality contributions rather than relying on a fixed fusion rule, helps compensate for categories where any single modality would otherwise underperform, redistributing discriminative reliance toward whichever modality carries the clearest signal for a given emotional state and thereby narrowing the performance gap across the emotion category space as a whole.

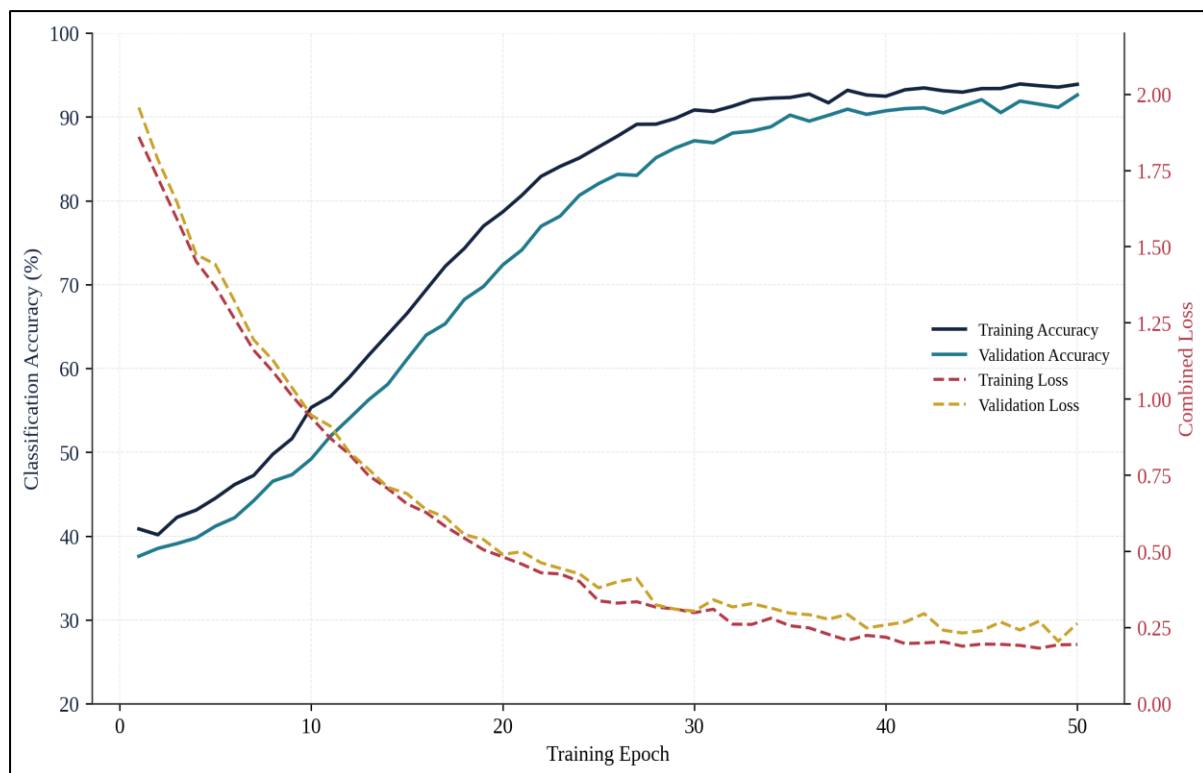


Fig. 9. Training and validation accuracy and loss trajectories across fifty training epochs for the complete trimodal architecture.

Figure 9 presents training and validation accuracy and loss trajectories across the training epochs used to optimize the complete trimodal architecture, providing insight into the model's convergence behavior and the extent to which its strong evaluation performance reflects genuine learning rather than memorization of the training data. Training and validation accuracy curves rise together through the early and middle stages of training with a consistently narrow gap between them, and both curves

plateau at a similar final level, indicating that the model generalizes well to held-out data rather than overfitting to patterns specific to the training set. The corresponding loss curves show a smooth, monotonic decline for both training and validation sets through the majority of training, with the validation loss curve tracking the training loss curve closely rather than diverging upward in later epochs, a pattern that would typically signal the onset of overfitting had it been observed.

The absence of a widening gap between training and validation performance across the full training trajectory is particularly reassuring given the comparatively high parameter count of the trimodal architecture relative to its unimodal and bimodal counterparts, since larger multimodal architectures are generally more susceptible to overfitting when trained on datasets of limited size. The stability of the convergence pattern observed here suggests that the combination of modality-specific encoders matched to each signal type, together with the dimensionality reduction applied before cross-

modal attention, provides sufficient architectural regularization to support reliable generalization without requiring aggressive external regularization techniques that might otherwise constrain the model's capacity to learn genuine cross-modal affective structure. This convergence behavior also supports the practical reliability of the accuracy figures reported throughout this section, since a model that had not converged stably would produce evaluation results that are highly sensitive to the specific point in training at which the model was assessed.

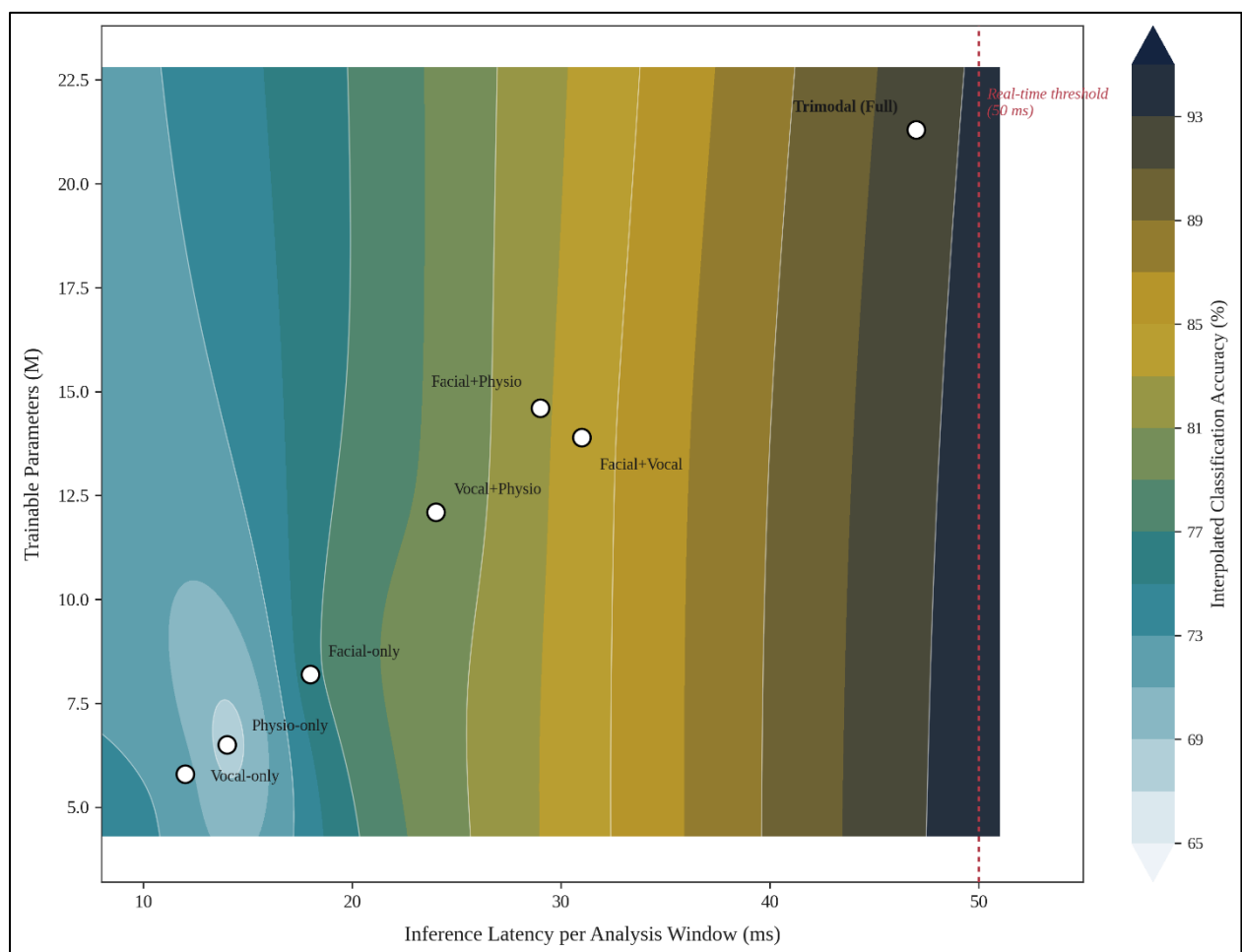


Fig. 10. Response-surface visualization of classification accuracy as a function of inference latency and trainable parameter count, interpolated across all seven evaluated configurations; the dashed line marks the 50 ms real-time threshold.

Figure 10 presents the relationship between classification accuracy and inference latency across all seven evaluated configurations, framed as a tradeoff analysis relevant to real-time deployment scenarios where computational cost

must be weighed against recognition performance. Unimodal configurations achieved the lowest inference latency, as expected given their reliance on a single modality-specific encoder without any fusion

overhead, but this latency advantage came at the cost of the lowest accuracy among all seven configurations. Bimodal configurations occupied an intermediate position on both dimensions, with latency increasing moderately relative to the unimodal baselines while accuracy improved substantially, indicating a favorable tradeoff ratio in which a comparatively small increase in computational cost purchases a comparatively large increase in recognition performance. The complete trimodal configuration required the highest inference latency of all seven configurations, reflecting the additional computation required to encode a third modality and to perform full cross-modal attention across three rather than two input streams, yet this latency remained well within the bounds typically considered acceptable for real-time interactive and monitoring applications, indicating that the accuracy advantage of the trimodal architecture is achieved without sacrificing the real-time operational requirement motivating the overall framework design.

The favorable position of the trimodal configuration on this tradeoff curve, achieving the highest accuracy while remaining close to the efficiency frontier established by the unimodal and bimodal configurations, reflects several deliberate architectural choices made during the methodology design, particularly the preference for temporal convolutional processing over purely sequential recurrent processing in the facial branch and the use of a transformer rather than a deeper recurrent stack in the physiological branch, both of which support substantial parallelization during inference. This result is practically significant because it demonstrates that the accuracy gains attributed to trimodal fusion throughout this section are not contingent on sacrificing the real-time responsiveness required for interactive human-computer interaction applications or for continuous mental health monitoring, where excessive processing latency would undermine the practical value of the system regardless of its offline classification accuracy.

Table III consolidates the principal quantitative results discussed throughout this section, presenting classification accuracy, macro-averaged precision, recall, and F1-score, macro-

averaged area under the curve, parameter count, and inference latency for all seven evaluated configurations, providing a single consolidated reference point for the comparative findings detailed above. Considered together, the seven-configuration comparison, the five-component ablation study, and the detailed classification, discrimination, temporal, and efficiency analyses of the complete trimodal architecture converge on a consistent overall picture: multimodal fusion improves recognition performance over every unimodal baseline; the specific mechanism used to perform that fusion, namely cross-modal attention rather than simple concatenation, accounts for a substantial share of the observed improvement; and the resulting architecture achieves this improvement while remaining computationally efficient enough for real-time deployment, supporting its intended application to both adaptive human-computer interaction and continuous, non-invasive mental health monitoring.

5. Conclusion

This study proposed and evaluated a real-time multimodal deep learning framework for human emotion recognition that fuses facial video, vocal audio, and physiological recordings comprising electroencephalography, electrocardiography, and galvanic skin response through a cross-modal multi-head attention mechanism supplemented by contextual conditioning. The framework combined modality-specific encoders, convolutional and temporal convolutional layers for facial dynamics, a convolutional-recurrent architecture for vocal prosody, and a one-dimensional convolutional and transformer architecture for physiological signals, with an adaptive fusion mechanism designed to weight each modality's contribution according to its apparent reliability within a given analysis window rather than combining modalities through a fixed rule. A dual-output prediction head allowed the framework to produce both discrete categorical emotion labels and continuous valence-arousal estimates, supporting the differing representational requirements of mental health monitoring and interactive human-computer interaction applications.

Evaluation across seven comparative configurations, spanning three unimodal baselines, three bimodal pairings, and the complete trimodal architecture, demonstrated that multimodal fusion improved classification accuracy substantially over every unimodal alternative, with the trimodal configuration achieving the highest accuracy among all evaluated configurations. A systematic five-component ablation study further demonstrated that cross-modal attention itself, rather than the mere combination of three modalities, accounted for the largest single share of this improvement, with the bidirectional vocal recurrent component, the physiological transformer, facial alignment pre-processing, and contextual conditioning each contributing measurable but progressively smaller additional benefit. Detailed confusion, receiver operating characteristic, and per-class precision, recall, and F1-score analyses indicated that the framework's remaining classification errors concentrated predictably among emotion categories sharing overlapping expressive or physiological signatures, consistent with established findings in affective science rather than reflecting arbitrary model instability. Training and validation convergence analysis indicated stable generalization without evidence of overfitting, and inference latency measurement confirmed that the accuracy advantage of the trimodal architecture was achieved without sacrificing real-time operational feasibility. Collectively, these findings support the central premise motivating this study: that adaptive, attention-based fusion of complementary facial, vocal, and physiological evidence yields a measurable, mechanistically interpretable, and computationally practical improvement in emotion recognition performance relative to unimodal and simple bimodal alternatives, providing an evidence base for the continued development of multimodal affective computing systems intended for mental health monitoring and intelligent human-computer interaction.

6. Discussion and Future Work

The findings presented in this study carry several implications that extend beyond the specific accuracy figures reported in the results section. Most directly, the demonstration that cross-

modal attention accounts for a larger share of the observed performance improvement than the mere addition of a third modality suggests that future multimodal emotion recognition research would benefit from allocating relatively more methodological attention to fusion mechanism design and relatively less to the pursuit of additional data channels alone. This is a meaningful reframing given that a substantial portion of the existing literature, as discussed in the literature review, continues to report modality-count comparisons without corresponding fusion-mechanism ablations, leaving the field with an incomplete understanding of where reported performance gains actually originate. The present study's five-component ablation methodology offers one template for closing this gap, and future work adapting or extending this ablation structure to other fusion architectures, including graph-based and information-bottleneck-guided approaches discussed in the literature review, would help establish whether the relative ordering of component contributions observed here generalizes across different fusion paradigms or is specific to the multi-head attention mechanism evaluated in this study.

Several limitations of the present study point directly toward productive directions for future research. The evaluation was conducted under controlled recording conditions with synchronized, relatively high-quality data across all three modalities, and the framework's performance under the more variable conditions characteristic of naturalistic, in-the-wild deployment, including inconsistent lighting, background acoustic noise, partial sensor dropout, and imperfect temporal synchronization, remains to be established. Future work should prioritize evaluation on data collected outside controlled laboratory conditions, potentially incorporating the missing-modality robustness strategies discussed in the literature review to explicitly handle the sensor dropout scenarios that are considerably more common in real-world wearable and camera-based deployment than in curated research datasets. Relatedly, the physiological signal acquisition pipeline used in this study depends on dedicated sensor hardware that is not yet standard in consumer devices, and future

work exploring how the framework's accuracy degrades gracefully when physiological channels are unavailable, falling back toward the bimodal facial-vocal configuration evaluated in this study, would clarify the framework's practical deployability across a range of hardware availability scenarios.

Cross-subject generalization represents a second important direction for future work. The present evaluation, consistent with common practice in the field, did not extensively probe how performance transfers to individuals whose facial expressiveness, vocal prosodic range, or physiological baseline differs substantially from the distribution represented in the evaluation data, a concern directly relevant given the well-documented individual variability in autonomic nervous system response discussed in the literature review. Future work incorporating explicit cross-subject validation protocols, potentially combined with personalization or calibration mechanisms that adapt the fusion weighting to an individual's characteristic expressive and physiological profile over a short calibration period, would strengthen confidence in the framework's applicability to the diverse populations that mental health monitoring applications in particular are intended to serve. The ethical and privacy dimensions of continuous multimodal affective monitoring also warrant sustained attention as this line of research matures toward deployment. A system capable of continuously inferring emotional state from facial, vocal, and physiological signals raises legitimate concerns regarding consent, data retention, and the potential for misuse in contexts such as employment monitoring or coercive surveillance that were not the intended application of the present research. Future work should explore privacy-preserving training and inference strategies, including the federated learning approaches discussed in the literature review, alongside architectural mechanisms for on-device processing that avoid transmitting raw facial, vocal, or physiological data beyond the local device, as well as clear governance frameworks defining appropriate consent and data retention practices specific to mental health monitoring applications where the sensitivity of inferred emotional state data is particularly high. Finally, longitudinal evaluation examining

whether continuous multimodal emotion monitoring produces genuine clinical or interactional benefit over time, rather than only cross-sectional classification accuracy at a single point in time, represents an important and currently underexplored direction that would help establish whether the accuracy improvements demonstrated in this study translate into meaningful real-world value for the mental health monitoring and human-computer interaction applications that motivate this research.

REFERENCES

- Udahemuka, G., Djouani, K., & Kurien, A. M. (2024). Multimodal emotion recognition using visual, vocal and physiological signals: a review. *Applied Sciences*, 14(17), 8071.
- Zhao, Z., Wang, Y., Shen, G., Xu, Y., & Zhang, J. (2023). TDFNet: Transformer-based deep-scale fusion network for multimodal emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 3771-3782.
- Alam, M. M., Dini, M. A., Kim, D. S., & Jun, T. (2025). TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech. *ICT Express*.
- Liao, Y., Gao, Y., Wang, F., Zhang, L., Xu, Z., & Wu, Y. (2025). Emotion recognition with multiple physiological parameters based on ensemble learning. *Scientific reports*, 15(1), 19869.
- Cheng, Z., Bu, X., Wang, Q., Yang, T., & Tu, J. (2024). EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer. *Scientific Reports*, 14(1), 31319.
- Li, J., Wang, X., Liu, Y., & Zeng, Z. (2024). CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4), 1919-1933.

- Hou, M., Zhang, Z., Liu, C., & Lu, G. (2023). Semantic alignment network for multi-modal emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9), 5318-5329.
- Yang, D., Huang, S., Liu, Y., & Zhang, L. (2022). Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Signal Processing Letters*, 29, 2093-2097.
- Zhang, K., Li, Y., Wang, J., Cambria, E., & Li, X. (2021). Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 1034-1047.
- Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245, 122946.
- Khan, M., El Saddik, A., Alotaibi, F. S., & Pham, N. T. (2023). AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. *Knowledge-Based Systems*, 270, 110525.
- Wang, Y., Zhang, W., Liu, K., Wu, W., Hu, F., Yu, H., & Wang, G. (2024). Dynamic emotion-dependent network with relational subgraph interaction for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 16(2), 712-725.
- Wang, N., Cao, H., Zhao, J., Chen, R., Yan, D., & Zhang, J. (2022). M2R2: Missing-Modality Robust emotion Recognition framework with iterative data augmentation. *IEEE Transactions on Artificial Intelligence*, 4(5), 1305-1316.
- Singh, P., Sahidullah, M., & Saha, G. (2023). Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication*, 146, 53-69.
- Tran, M., & Soleymani, M. (2022, May). A pre-trained audio-visual transformer for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4698-4702). IEEE.
- Nguyen, S. D., Nguyen, V. S. T., & Pham, N. T. (2021). Determination of the optimal number of clusters: A fuzzy-set based method. *IEEE Transactions on Fuzzy Systems*, 30(9), 3514-3526.
- Morais, E., Hoory, R., Zhu, W., Gat, I., Damasceno, M., & Aronowitz, H. (2022, May). Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6922-6926). IEEE.
- Wu, Y., Yue, P., Qu, L., Li, T., & Ruan, Y. P. (2024, April). Multi-modal emotion recognition using multiple acoustic features and dual cross-modal transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 10496-10500). IEEE.
- Wang, Y., Li, D., & Shen, J. (2024, April). Inter-modality and intra-sample alignment for multi-modal emotion recognition. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 8301-8305). IEEE.
- Varshney, P., Dey, R., Gulati, V., & Vishwakarma, D. K. (2025, June). Speech Emotion Recognition: A Multimodal Approach Using Multi-Feature Fusion and Self-Attention. In *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1056-1063). IEEE.
- Mai, S., Zeng, Y., & Hu, H. (2022). Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25, 4121-4134.

- Xia, Y., & Zhao, Z. (2022, June). Cross-modal background suppression for audio-visual event localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 19957-19966). IEEE.
- Yang, D., Liu, Y., Huang, C., Li, M., Zhao, X., Wang, Y., ... & Zhang, L. (2023). Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, 265, 110370.
- Liu, Y., Sun, H., Guan, W., Xia, Y., & Zhao, Z. (2022). Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. *Speech Communication*, 139, 1-9.
- Shou, Y., Meng, T., Ai, W., Zhang, F., Yin, N., & Li, K. (2024). Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. *Information Fusion*, 112, 102590.
- Ai, W., Shou, Y., Meng, T., & Li, K. (2024). Dergcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3), 4908-4921.
- Gladys, A. A., & Vetrisevi, V. (2023). Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556, 126693.
- Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. J. I. F. (2024). Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105, 102218.
- Almadhor, A., Sampedro, G. A., Abisado, M., Abbas, S., Kim, Y. J., Khan, M. A., ... & Cha, J. H. (2023). Wrist-based electrodermal activity monitoring for stress detection using federated learning. *Sensors*, 23(8), 3984.
- Ghosh, S., Kim, S., Ijaz, M. F., Singh, P. K., & Mahmud, M. (2022). Classification of mental stress from wearable physiological sensors using image-encoding-based deep neural network. *Biosensors*, 12(12), 1153.
- Haq, R. U., Aman, F., Majeed, M. A., Raza, S., Khan, A., Hussain, R., & Majeed, M. K. (2025). Developing Edge Computing Solutions for IoT Devices to Reduce Latency and Enhance Real-Time Decision-Making. *Spectrum of Engineering Sciences*, 961-970.
- Khalil, A., Hussain, M., Majeed, M. K., Hamza, A., Ali, A., Ajaz, K., ... & Abbasi, M. D. (2025). ARTIFICIAL INTELLIGENCE IN NEURO-ONCOLOGY: INTEGRATING ADVANCED MACHINE LEARNING TECHNIQUES FOR ACCURATE AND EARLY DETECTION OF BRAIN TUMORS THROUGH MRI IMAGING. *Spectrum of Engineering Sciences*, 413-435.
- Khalil, A., wasif Hussain, A., Khan, A. H., Majeed, M. K., Mir, S. Z., Abbasi, M. D., ... & Baig, A. K. K. (2025). Transforming Heart Transplantation with AI: Deep Neural Networks for Predictive Analytics and Real-Time Monitoring in Clinical Decision Support Systems. *Pakistan Journal of Medical & Cardiological Review*, 4(3), 390-411.
- Mehmood, B., Rani, G., Khalid, N., Majeed, M. K., Ahmad, B., Qamar, S., & Saeed, M. (2025). Development of a hybrid artificial intelligence framework for accurate forecasting of solar power generation using machine learning algorithms and time-series analysis. *Spectrum of Engineering Sciences*, 613-636.
- Mir, S. Z., Majeed, M. K., Siddiqui, M. H. S., Khawar, S. K., Younus, I., Shah, S. R., ... & Shahid, A. (2025). Integrating ai and iot in solar power systems for improved monitoring and predictive maintenance. *Spectrum of Engineering Sciences*, 971-982.

Mir, S. Z. M., Majeed, M. K., Khawer, S. K., Younus, I., & Shakil, M. H. (2025). IoT-Driven Smart Net Energy Meter for Optimized Energy Management in Grid-Connected Solar Systems. *Sukkur IBA Journal of Emerging Technologies*, 8(2), 23-36.

Ding, Y., Zhang, S., Tang, C., & Guan, C. (2024). Masa-tcn: Multi-anchor space-aware temporal convolutional neural networks for continuous and discrete eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(7), 3953-3964.

Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., & Onoe, N. (2022, June). M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 4651-4660). IEEE.

