Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MRF-CMFT: Mamba-Based Quality-Gated Multi-View Fusion Framework for EEG–ECG Affect Recognition on Low-Quality Data

Habiba Mostafa
Medical Informatics and Neural-Dynamics (MIND) Lab
Nedaa M. Elsherbini
Medical Informatics and Neural-Dynamics (MIND) Lab
Mervat Samy
Mansoura University
Ahmed M. Anter
ahmed_anter@fcis.bsu.edu.eg

Egypt-Japan University of Science and Technology

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11411925/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MRF-CMFT: Mamba-Based Quality-Gated Multi-View Fusion Framework for EEG–ECG Affect Recognition on Low-Quality Data

Habiba Mostafa[d], Nedaa M.Elsherbini[d], Mervat Samy[d,b], Ahmed M. Anter[a,c,d,*]

[a]*Egypt-Japan University of Science and Technology (E-JUST), 21934, New Borg El-Arab City, Alexandria, Egypt*
[b]*Department of Clinical Pharmacology, Faculty of Medicine, Mansoura University, 35516, Mansoura, Egypt*
[c]*Faculty of Computers and Artificial Intelligence, Beni-Suef University, 62511, Benisuef, Egypt*
[d]*Medical Informatics and Neural-Dynamics (MIND) Lab, Cairo, Egypt*

Abstract

Physiological affect recognition is inherently a multimodal fusion problem, yet its reliability is limited by noisy measurements, heterogeneous representations, and temporally varying signal quality. We propose Multi-view Representation Fusion with Quality-aware Cross-Modal Feature Transformation (MRF-CMFT), a quality-aware framework for EEG–ECG affect recognition. Each modality is represented using three complementary views: short-time Fourier transform power maps, Morlet wavelet scalograms, and compact handcrafted feature representations. These views are encoded by convolutional sequence encoders with a Mamba-based backbone and integrated through adaptive quality-aware view fusion. Cross-modal attention captures interactions between EEG and ECG, while a quality-aware modality gate dynamically regulates their contributions according to estimated signal reliability. A temporal Mamba aggregator then integrates consecutive windows to produce trial-level predictions. The MRF-CMFT was evaluated on the DREAMER and AMIGOS datasets using trial-grouped subject-dependent, subject-independent, and leave-one-subject-out protocols to prevent information leakage. On DREAMER, the framework achieves subject-independent

*Corresponding Author: Ahmed M. Anter (Ahmed_Anter@fcis.bsu.edu.eg)

AUC of 0.640, 0.534, and 0.581 for valence, arousal, and dominance, respectively; the corresponding AMIGOS AUC are 0.496, 0.561, and 0.623. The mean subject-dependent AUC is 0.614 on DREAMER and 0.616 on AMIGOS. Under the evaluated corruption conditions, the quality-aware fusion mechanism exhibits graceful degradation, with dominance AUC on AMIGOS remaining above 0.83 across all tested scenarios. These findings indicate that explicitly modelling representation reliability and cross-modal complementarity supports robust physiological affect recognition under subject variability and degraded sensing conditions.

Keywords: Multimodal information fusion, Physiological emotion recognition, EEG–ECG fusion, Quality-aware fusion, Mamba, Cross-modal learning

---

1. Introduction

Affective computing seeks to enable machines to perceive and interpret human emotional states [1, 2]. Physiological signals have emerged as promising sensing modalities for affect recognition, as they provide implicit measurements of neural and autonomic responses that are less susceptible to deliberate behavioral masking than facial expressions, speech, or language [3]. Among these signals, electroencephalography (EEG) captures neural activity with high temporal resolution, whereas electrocardiography (ECG) reflects cardiovascular dynamics associated with autonomic regulation. Their complementary physiological characteristics make joint EEG–ECG analysis particularly attractive for multimodal affect recognition. Public datasets such as DEAP [4], DREAMER [5], and AMIGOS [6] have further supported progress in this field by providing multimodal physiological recordings accompanied by self-reported affective labels.

Despite substantial progress, reliable physiological affect recognition remains challenging under realistic acquisition conditions [7, 8]. EEG and ECG exhibit substantial inter-subject variability, non-stationarity, and heterogeneous temporal and spectral characteristics. In addition, signal quality may vary throughout a recording because of electrode displacement, poor sensor contact, motion artifacts, physiological interference, missing channels, and unreliable measurements. These factors affect the modalities unevenly and may change across analysis windows. Consequently, effective multimodal recognition requires more than simply combining heterogeneous signals; it

requires mechanisms that can determine which observations and representations provide reliable information at a given point in time.

Signal representation choice introduces a second information-fusion challenge. Handcrafted descriptors, including differential entropy, Hjorth parameters, spectral entropy, and relative band power, provide compact statistical and frequency-domain summaries of physiological activity. In contrast, time–frequency representations such as short-time Fourier transform (STFT) spectrograms and continuous wavelet transform (CWT) scalograms preserve localized temporal-frequency structures that may be lost in summary statistics. These representations therefore provide complementary views of the same underlying physiological process. However, many existing approaches rely on a limited set of representations or a single learned feature space, which may restrict the diversity of information available to subsequent fusion mechanisms.

Multimodal fusion strategies have consequently evolved from conventional early feature-level and late decision-level fusion toward attention-based and cross-modal architectures [9, 10]. Although these approaches can learn interactions between heterogeneous sources, many rely on fixed or globally learned modality contributions and have limited ability to respond to transient changes in signal reliability. In physiological sensing, this limitation may allow artifact-corrupted observations to influence the fused representation even when another modality provides cleaner information. Quality-aware fusion addresses this limitation by conditioning the integration process on estimated signal reliability, thereby attenuating unreliable observations and emphasizing more informative views and modalities.

A further challenge concerns the efficient modeling of long-range dependencies within physiological representations. Transformer architectures have demonstrated strong sequence-modeling capabilities through self-attention [11, 12]; however, their quadratic computational and memory complexity with respect to sequence length can become restrictive when processing long or high-resolution sequences. Structured state-space models provide an alternative sequence-modeling paradigm. In particular, Mamba introduces selective state-space dynamics that support efficient long-range sequence modeling with linear scaling in sequence length [13, 14]. These properties make Mamba a promising backbone for physiological representation sequences. Nevertheless, its integration with multi-representation and quality-aware EEG–ECG fusion, particularly under degraded input conditions, remains insufficiently investigated.

In addition to architectural considerations, the reliability of reported performance is an important methodological concern in physiological affect recognition. Physiological trials are commonly segmented into multiple analysis windows, and neighboring windows may be highly correlated because they originate from the same trial and share the same affective label. Random window-level partitioning can consequently place closely related windows from the same trial in both the training and testing sets. Under such a protocol, a model may exploit trial-specific characteristics rather than learn representations that generalize to unseen trials or subjects. This form of information leakage can substantially inflate performance and complicate comparisons across studies. Evaluation protocols that explicitly respect trial and subject boundaries are therefore essential for obtaining reliable estimates of generalization.

These issues are particularly important when assessing robustness to incomplete or degraded physiological observations. Practical acquisition may involve sensor noise, missing representations or modalities, temporal misalignment, and other unreliable inputs. Multi-view learning has investigated the fusion of noisy, incomplete, and heterogeneous views [15]; however, the combined problem of representation diversity, modality reliability, efficient temporal modeling, and leakage-controlled evaluation remains insufficiently addressed in physiological affect recognition.

To address these interconnected challenges, we propose the Multi Representation Frequency-aware Cross-Modal Fusion Transformer (MRF-CMFT), a quality-aware multimodal framework for EEG–ECG affect recognition under degraded sensing conditions. MRF-CMFT integrates three complementary representations for each physiological modality: handcrafted statistical and spectral descriptors, log-STFT power maps, and Morlet CWT scalograms. Each representation is processed by a convolutional feature extractor followed by a selective state-space sequence encoder based on Mamba. The resulting representation-specific features are integrated through quality-calibrated adaptive view fusion, while cross-modal attention captures interactions between EEG and ECG. A quality-aware modality gate further regulates the contribution of each modality according to its estimated signal quality. Finally, a temporal Mamba aggregator combines consecutive windows to generate trial-level predictions.

To provide a rigorous assessment of the proposed framework, we evaluate it on the DREAMER and AMIGOS datasets using leakage-controlled subject-independent (SI), trial-grouped subject-dependent (SD), and leave-

one-subject-out (LOSO) protocols. Additionally, an exploratory random window-level evaluation was performed to quantify the effect of trial-level leakage on reported performance. Robustness is assessed under controlled signal degradation, including noise, missing views, missing modalities, temporal misalignment, and incomplete observations. Component-level ablations are conducted to isolate the contributions of the proposed representations, quality-aware view fusion, cross-modal attention, and modality gating.

The main contributions of this work are as follows:

- **A multi-representation EEG–ECG fusion framework.** MRF-CMFT is proposed, which jointly exploits handcrafted, STFT, and Morlet CWT representations of EEG and ECG using convolutional encoders and a Mamba-based sequence-modeling backbone.
- **Quality-aware multi-view and cross-modal fusion.** Quality-calibrated adaptive view fusion and a quality-aware modality gate are introduced that dynamically regulate the contribution of representation views and physiological modalities according to their estimated reliability.
- **Robustness to degraded physiological inputs.** The framework is systematically evaluated under controlled corruption scenarios involving noisy, missing, and temporally misaligned information to determine whether quality-aware fusion promotes graceful degradation rather than catastrophic performance loss.
- **Leakage-controlled evaluation of physiological affect recognition.** A unified evaluation protocol is established based on trial-grouped subject-dependent, strict subject-independent, and LOSO evaluation, while separately quantifying the performance inflation associated with random window-level partitioning.

  The remainder of this paper is organized as follows. Section 2 reviews related research on EEG-based emotion recognition, multimodal physiological fusion, state-space sequence modelling, and evaluation protocols. Section 3 describes the proposed MRF-CMFT architecture, multi-view representation, quality-aware fusion mechanism, and training strategy. Section 4 presents the datasets, evaluation protocols, implementation details, baselines, and evaluation metrics. Section 5 reports the experimental results, including robustness, explainability,

cross-dataset, and state-of-the-art comparisons. Finally, Section 6 summarizes the findings, limitations, and directions for future work.

2. Related Work

2.1. EEG-Based Emotion Recognition

Electroencephalography (EEG) has been extensively investigated for emotion recognition because affective states are reflected in neural activity across multiple cortical regions and frequency bands. Early studies primarily relied on handcrafted representations, including differential entropy (DE), power spectral density (PSD), Hjorth parameters, and entropy-based descriptors. These features characterize complementary spectral, statistical, and signal-complexity properties and were commonly combined with conventional machine-learning classifiers, including support vector machines and ensemble methods, for affective-state discrimination [16, 17, 18, 3].

The emergence of deep learning shifted EEG emotion recognition toward automatic representation learning. Deep belief networks and stacked autoencoder architectures were initially used to learn hierarchical representations from EEG features, reducing reliance on manually designed classifiers. Convolutional neural networks (CNNs) subsequently demonstrated the ability to learn spatial, spectral, and local temporal patterns from raw EEG signals, frequency-domain features, and two-dimensional arrangements of channels and frequency bands [19, 17, 20]. CNN-based approaches have also been extended through randomized and ensemble designs to improve robustness and predictive performance [21]. In addition, multi-scale temporal convolutions and spatio-temporal feature-extraction strategies have been introduced to represent affective EEG activity at different temporal resolutions [22].

Recurrent neural networks (RNNs) provide a complementary approach by explicitly modeling temporal dependencies in EEG sequences. Long short-term memory (LSTM) networks are particularly suitable for this purpose because their recurrent state mechanism can preserve information over extended temporal intervals [23]. Hybrid CNN–LSTM architectures therefore combine local spatial–spectral feature extraction with subsequent modeling of temporal evolution. Such architectures have demonstrated the complementary roles of convolutional and recurrent processing in EEG emotion recognition [18, 24].

Graph neural networks offer a different perspective by representing EEG electrodes and their relationships explicitly. Dynamic graph convolutional

neural networks (DGCNNs) learn graph structures in which electrode channels are treated as nodes and inter-channel relationships are optimized jointly with the recognition task [25]. Subsequent studies have combined graph convolution with recurrent networks to capture both functional relationships among electrodes and temporal dependencies in EEG sequences. Dynamic graph construction and source-based graph representations have further been explored to obtain more flexible descriptions of functional relationships between brain regions [26].

Attention mechanisms have enabled models to assign different importance to EEG channels, frequency bands, or temporal segments. For example, attention has been integrated with graph convolution to jointly learn spatial electrode relationships and the relative relevance of channels and frequency components [27]. Transformer architectures extend this principle by enabling interactions between distant elements of a sequence through self-attention [11]. Recent EEG models have incorporated spatial–temporal attention, hierarchical brain-region representations, and graph-based token construction to capture electrode-level and temporal dependencies simultaneously [28, 29].

These developments have also encouraged hybrid architectures that combine graph convolution, recurrent processing, attention, and Transformer mechanisms. Graph convolutional and LSTM networks jointly model inter-channel relationships and temporal dynamics [30], while graph-attention and Transformer-based designs incorporate electrode connectivity, frequency-band importance, and temporal context within a common architecture [27]. Multi-scale spatio-temporal models provide an additional strategy for representing EEG activity at different temporal resolutions [31]. Collectively, these approaches show that EEG emotion recognition increasingly relies on complementary architectural components rather than a single feature-extraction mechanism.

Time–frequency analysis provides a further means of exposing discriminative EEG structure. Short-time Fourier transform (STFT) representations preserve localized spectral information over time, whereas wavelet-based representations provide multi-resolution analysis of transient and non-stationary oscillatory activity. Deep-learning models have increasingly used spectrograms and related time–frequency representations as two-dimensional inputs to CNNs and pretrained vision architectures, demonstrating the value of frequency-dependent information for EEG classification [19]. Nevertheless, much of the existing literature emphasizes one representation type or one dominant architectural paradigm. This leaves open the question of how

handcrafted descriptors, spectral features, and time–frequency views can be coordinated within a unified representation-learning and fusion framework.

2.2. Multimodal Physiological Emotion Recognition

Multimodal physiological emotion recognition extends EEG-based analysis by combining central neural measurements with peripheral indicators of autonomic activity. EEG reflects cortical processes, whereas electrocardiography (ECG) and heart-rate variability (HRV) descriptors provide information related to cardiovascular and autonomic regulation. Their joint use can therefore capture affective responses that are not fully represented by either modality alone. The broader motivation for combining neural and peripheral measurements is consistent with established work in affective computing and HRV analysis [1, 3, 32].

The DREAMER dataset, introduced by Katsigiannis and Ramzan [5], provides synchronized EEG and ECG recordings from 23 participants exposed to audiovisual emotional stimuli, together with self-assessed valence, arousal, and dominance ratings. AMIGOS provides a complementary benchmark containing multiple physiological modalities and affective responses collected under individual and group experimental conditions [6]. These datasets have subsequently supported the development and comparison of multimodal physiological affect-recognition systems under varied experimental settings.

Early multimodal methods commonly used feature-level or decision-level fusion, combining modality-specific features before classification or modality-specific predictions after separate processing. The distinction between early, intermediate, and late fusion has been formalized in the broader multimodal learning literature [9]. Although these strategies provide straightforward baselines, they may not adequately represent the nonlinear dependencies between heterogeneous physiological sources. Deep learning has therefore increasingly been used to learn modality-specific representations and fusion functions jointly.

Convolutional and recurrent architectures have been applied to learn complementary representations from multimodal physiological signals. One representative system combines a two-dimensional CNN for EEG with CNN- and LSTM-based processing for ECG and GSR, and evaluates multimodal fusion on DREAMER and AMIGOS. Its results demonstrate the potential benefit of combining heterogeneous physiological sources over using individual modalities alone [33]. End-to-end multimodal deep networks have likewise been

proposed to learn complementary representations from physiological and behavioral modalities without relying exclusively on handcrafted fusion rules [34].

Attention-based models provide a more flexible alternative by assigning data-dependent importance to modality-specific features, temporal segments, or cross-modal interactions. Recent multimodal physiological emotion-recognition models have introduced intra-modality and inter-modality attention to preserve modality-specific information while learning dependencies across sources. Evaluations on datasets such as DEAP and DREAMER demonstrate the usefulness of attention-based multimodal fusion [35].

Cross-modal attention provides an explicit mechanism through which the representation of one modality can condition the processing of another. Rather than simply concatenating EEG and peripheral features, these models learn intermediate interactions between heterogeneous physiological streams. For example, low-rank tensor fusion combined with cross-modal attention has been used to model interactions among EEG, EOG, EMG, and GSR representations for subject-independent emotion recognition [36]. Graph-based methods provide another complementary direction by modeling relationships within each physiological source before cross-modal integration. Recent EEG–ECG architectures have combined dynamic graph convolution, LSTM-based temporal processing, and self-attention to learn EEG functional connectivity and ECG features jointly [37].

More recent approaches have continued to combine CNNs, recurrent networks, and attention mechanisms for direct EEG–ECG analysis. For example, an attention-enhanced CNN–GRU architecture jointly processes EEG and ECG signals and learns time-domain, frequency-domain, and nonlinear representations for valence, arousal, and dominance recognition [38]. These studies demonstrate a progression from simple feature concatenation toward architectures that jointly learn modality-specific features, temporal dependencies, and cross-modal interactions.

A persistent limitation is that physiological modalities may differ in reliability across windows and recording conditions. Motion artifacts, electrode displacement, poor sensor contact, physiological noise, and missing measurements can selectively degrade one modality while leaving another comparatively informative. Consequently, fusion functions learned without explicit reliability information may assign inappropriate importance to corrupted observations. Recent models have addressed this issue indirectly through attention, cross-modal weighting, or experiments involving modality-specific

noise and missing information [35, 36]. However, explicit quality-aware fusion of multiple representations within EEG–ECG affect recognition, particularly under controlled missing, noisy, and degraded-input conditions, remains comparatively underexplored.

2.3. Multi-View and Quality-Aware Fusion

Multi-view learning considers multiple representations of a shared underlying target and seeks to exploit their complementary information [39, 40]. In practical settings, however, individual views may be noisy, incomplete, missing, or temporally misaligned. This distinction is important for physiological sensing because different representations of the same signal may have different reliability profiles under the same acquisition conditions.

Existing multi-view and multimodal methods have investigated view-specific encoders, attention-based weighting, gating networks, modality dropout, and incomplete-view learning to improve robustness to unreliable inputs [39, 40, 41, 42]. Related problems arise in multimodal medical and physiological data fusion, where heterogeneous measurements may contain missing observations, modality-specific noise, or varying signal quality [41]. Recent surveys further distinguish inter-modality missingness, in which an entire modality is absent, from intra-modality degradation, in which a nominally available view is present but corrupted [41, 43].

For physiological affect recognition, the views may include handcrafted descriptors, STFT spectrograms, and continuous wavelet transform (CWT) scalograms. Handcrafted descriptors provide compact statistical and spectral summaries, whereas STFT and CWT representations preserve localized temporal-frequency structure. Their combination can expose information that is not available from an individual view. However, equal weighting is not necessarily appropriate when one representation is strongly affected by noise or acquisition artifacts.

Quality-aware, uncertainty-aware, and reliability-aware fusion methods address this issue by incorporating information about input trustworthiness into the integration process. Instead of assigning fixed importance to each view or modality, these methods estimate quality or uncertainty and use it to modulate fusion weights. Such strategies have recently been investigated in multimodal medical imaging, clinical prediction, and outdoor health monitoring, where quality scores or uncertainty estimates guide the combination of heterogeneous measurements [44, 45, 43].

Despite this progress, the explicit integration of quality-aware multi-view fusion with EEG–ECG affect recognition under controlled low-quality conditions remains comparatively limited. In particular, fewer studies jointly consider representation diversity, modality complementarity, view-level reliability, and temporal variation in signal quality. This unresolved combination provides the specific information-fusion context for the framework developed in this work.

2.4. State-Space Models for Physiological Sequence Modeling

Transformer architectures have become a dominant paradigm for sequence modeling because self-attention enables direct interactions between distant elements of a sequence [11]. However, standard self-attention has quadratic computational and memory complexity with respect to sequence length. This limitation is relevant when physiological signals are transformed into high-resolution temporal or time–frequency representations and subsequently organized as long token sequences.

State-space models (SSMs) provide an alternative framework by representing temporal dependencies through latent state transitions. Early deep SSM research investigated continuous-time state-space formulations in recurrent and convolutional architectures, establishing a foundation for efficient long-range sequence modeling [46, 47]. The HiPPO framework introduced a principled mechanism for retaining long-term information through structured polynomial projections, while subsequent linear state-space layers demonstrated how continuous-time dynamics could be incorporated into trainable sequence models.

The Structured State Space sequence model (S4) advanced this direction through a structured parameterization that supports efficient computation while retaining long-range modeling capability [48]. S4 achieved strong performance on long-range sequence benchmarks and showed that SSMs could serve as competitive alternatives to recurrent and attention-based models for sequences containing thousands of time steps. Subsequent work further simplified and improved SSM architectures. In particular, S5 uses a single multi-input, multi-output state-space model rather than multiple independent SSMs while retaining efficient parallel sequence processing and long-range modeling capability [49].

Mamba introduced selective state-space modeling, in which the parameters governing state propagation depend on the input [13]. This mechanism

enables input-dependent control over which information is propagated or retained, offering an adaptive alternative to the fixed dynamics of earlier SSMs. Mamba also incorporates hardware-aware algorithms for efficient sequence processing, making it attractive for long-context applications.

The applicability of Mamba has since expanded beyond language and general sequence modeling. Vision Mamba, for example, investigated bidirectional state-space blocks as an alternative to self-attention for visual representation learning and demonstrated the use of Mamba for high-resolution image processing [50]. This broader adoption is relevant to physiological time–frequency analysis because spectrograms and scalograms can likewise be converted into sequences requiring efficient contextual modeling.

Mamba-2 further developed the connection between SSMs and attention through the structured state-space duality framework [14]. This work identified close mathematical relationships between the two paradigms and introduced an improved SSM formulation with higher computational efficiency. Together, S4, S5, Mamba, and Mamba-2 illustrate the progression from structured long-range state modeling toward selective and efficient sequence architectures.

These developments are relevant to physiological signal analysis because EEG and ECG contain temporally dependent observations, while their time–frequency representations may produce long sequences in which informative patterns occur over multiple temporal scales. Physiological recordings may also include transient artifacts and degraded segments. Selective state propagation therefore offers a possible means of modeling long-range dependencies while regulating the contribution of individual sequence elements.

Despite the rapid development of SSMs in language modeling, long-range sequence benchmarks, and computer vision, their use in multimodal physiological affect recognition remains comparatively underexplored. In particular, limited work has combined selective SSMs with heterogeneous multi-view physiological representations and quality-aware cross-modal fusion. This gap motivates examining Mamba as part of an integrated fusion architecture rather than only as a direct replacement for a Transformer encoder.

Trial-grouped evaluation addresses within-trial dependence by keeping all windows from a given trial in the same partition. Subject-independent evaluation imposes a stronger constraint by separating participants across training and testing, while leave-one-subject-out (LOSO) evaluation repeats this process for each held-out participant. These protocols assess different forms of generalization and should not be treated as interchangeable with a

random window-level split.

Evaluation practices nevertheless remain heterogeneous across physiological emotion-recognition studies. Reported results may be based on window-level random splits, trial-grouped validation, subject-dependent evaluation, or subject-independent protocols, which complicates direct comparison between models. In addition, the performance difference between leakage-permitting and leakage-safe configurations is not always quantified using the same dataset, architecture, and preprocessing pipeline.

Accordingly, the present study treats trial-grouped, subject-independent, and LOSO protocols as the primary evaluation settings. A window-level random split is used only as an exploratory methodological analysis to quantify the effect of trial-level leakage. This separation prevents performance obtained under correlated train–test sampling from being interpreted as evidence of genuine generalization.

2.5. Research Gap

The literature shows substantial progress in EEG representation learning, multimodal physiological fusion, multi-view learning, reliability-aware integration, and efficient sequence modeling. However, these strands have generally developed in parallel. EEG studies have explored increasingly sophisticated spatial, temporal, graph, attention, and time–frequency representations; multimodal studies have focused on combining central and peripheral physiological sources; multi-view studies have addressed complementary or incomplete representations; and SSM research has focused on efficient long-context sequence processing. Their combined use in a single physiological affect-recognition framework remains limited.

In particular, an open gap concerns the joint integration of (i) heterogeneous EEG and ECG views, (ii) view- and modality-level reliability estimation, (iii) cross-modal interaction modeling, (iv) efficient long-range temporal aggregation, and (v) evaluation protocols that prevent trial-level leakage. The proposed MRF-CMFT framework is designed to address this combined gap by bringing these components together and assessing them through ablation, cross-dataset, controlled-degradation, and leakage-aware experiments. The resulting analysis focuses not only on predictive performance but also on how representation quality and evaluation design affect multimodal physiological fusion.

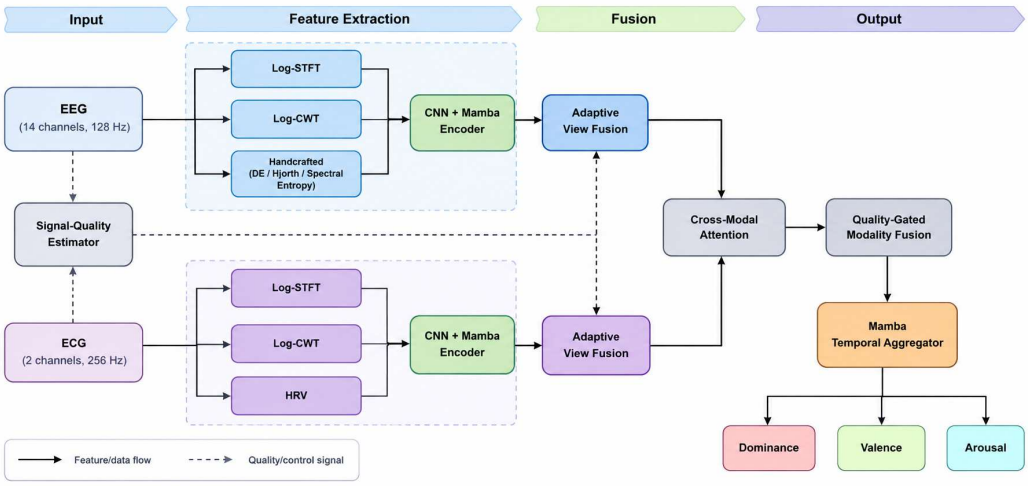


Figure 1: Overview of the proposed MRF-CMFT architecture.

3. Methodology

3.1. Proposed Framework and Problem Formulation

Figure 1 illustrates the overall architecture of the Multi-Representation Fusion with Cross-Modal Mamba Transformer (MRF-CMFT) framework. The proposed framework consists of five sequential stages: (1) signal preprocessing, (2) multi-view representation generation, (3) modality-specific feature encoding, (4) quality-aware adaptive cross-modal fusion, and (5) trial-level emotion classification.

First, synchronized EEG and ECG recordings are preprocessed to reduce noise and physiological artifacts before being segmented into fixed-length analysis windows. Each window is subsequently transformed into complementary feature representations, including handcrafted statistical descriptors, Short-Time Fourier Transform (STFT) spectrograms, and Morlet Continuous Wavelet Transform (CWT) scalograms. These heterogeneous representations are independently encoded through lightweight convolutional feature extractors followed by Mamba-based sequence modeling networks to capture both local spatial patterns and long-range temporal dependencies. Finally, an AdaptiveViewSelector module dynamically estimates the reliability of each modality and representation, producing a quality-aware fused

representation that is used to predict the emotional state in terms of valence, arousal, and dominance.

Compared with conventional multimodal fusion approaches that treat all modalities equally, the proposed framework explicitly models the reliability of physiological signals under realistic acquisition conditions. Through quality-calibrated adaptive view fusion and a quality-aware modality gate, MRF-CMFT dynamically regulates the contribution of each representation and modality according to its estimated trustworthiness. As a result, the framework maintains robust affect recognition performance even when EEG or ECG signals are affected by noise, motion artifacts, missing segments, or temporal misalignment.

3.1.1. Problem Formulation

Let

$$X^{EEG} = \{x_1^{EEG}, x_2^{EEG}, \ldots, x_T^{EEG}\} \tag{1}$$

and

$$X^{ECG} = \{x_1^{ECG}, x_2^{ECG}, \ldots, x_T^{ECG}\} \tag{2}$$

denote the synchronized EEG and ECG recordings corresponding to a single emotional trial, where T represents the number of temporal samples. The objective of the proposed framework is to learn a mapping

$$f_\theta : (X^{EEG}, X^{ECG}) \rightarrow y, \tag{3}$$

where θ denotes the learnable model parameters and

$$y = (y_v, y_a, y_d) \tag{4}$$

contains the binary labels for the three affective dimensions of valence, arousal, and dominance, respectively.

Rather than processing an entire physiological recording directly, each trial is partitioned into a sequence of fixed-length, optionally overlapping analysis windows. Window-level feature representations are extracted independently from each modality and subsequently aggregated to produce a single trial-level prediction. This hierarchical strategy enables the model to exploit fine-grained temporal dynamics while ensuring that evaluation is performed at the trial level.

To prevent optimistic performance estimates arising from temporal leakage, all windows extracted from the same trial remain within the same partition throughout training, validation, and testing. Consequently, no information from a trial appears simultaneously in multiple data splits under the primary evaluation protocols. An alternative window-level random split is investigated separately for exploratory analysis and is explicitly reported as a leakage-permitting protocol rather than a primary evaluation setting.

3.2. Data Preprocessing and Multi-View Representation

Reliable physiological emotion recognition requires careful preprocessing to reduce recording artifacts while preserving affect-related information. Accordingly, EEG and ECG recordings undergo modality-specific preprocessing before feature extraction.

3.2.1. EEG Preprocessing

Raw EEG recordings are first examined for invalid or non-finite samples, which are removed or replaced using standard interpolation procedures where appropriate. Each channel is subsequently band-pass filtered to suppress low-frequency baseline drift and high-frequency noise while retaining frequency components associated with cognitive and emotional processing. Baseline correction is performed independently for every channel using the pre-stimulus reference segment to compensate for subject-specific resting activity.

Following the strategy adopted in previous DREAMER studies, only the final 60s of each trial are retained. This emphasizes sustained affective activity and reduces the influence of transient adaptation responses. Finally, recordings are segmented into fixed-length windows for subsequent multi-view representation learning.

3.2.2. ECG Preprocessing

The ECG pipeline consists of artifact removal, band-pass filtering, and baseline correction. Following noise suppression, cardiac activity is characterized through heartbeat detection and RR-interval estimation. From the resulting heartbeat sequence, heart-rate (HR) and heart-rate variability (HRV) descriptors are extracted. HR captures the overall cardiac response to arousal, while HRV reflects autonomic nervous system (ANS) regulation. By jointly exploiting central nervous system (CNS) responses from EEG and autonomic responses from ECG, the framework captures complementary aspects of human emotional processing.

3.2.3. Multi-View Representations

To exploit the complementary information contained in EEG and ECG signals, each analysis window is transformed into three heterogeneous views: (1) handcrafted statistical features, (2) short-time Fourier transform (STFT) spectrograms, and (3) Morlet continuous wavelet transform (CWT) scalograms. As illustrated in Fig. 2, these representations capture complementary characteristics of the same physiological window, combining compact domain-specific descriptors with localized time–frequency information. Unlike single-representation approaches, this multi-view strategy enables the network to exploit handcrafted domain knowledge alongside automatically derived time–frequency patterns.

3.2.4. Handcrafted Feature Representation

For EEG, three feature families are computed per channel and frequency band:

- **Differential Entropy (DE):** Defined as

$$DE = \frac{1}{2}\ln(2\pi e\sigma^2), \tag{5}$$

  assuming a Gaussian signal distribution with variance σ^2. Five band-dynamics rows (differential-entropy bands referenced to the pre-stimulus baseline) are formed.

- **Hjorth Parameters:** Activity, Mobility, and Complexity are computed to characterize signal power, dominant frequency, and waveform complexity.

- **Spectral Entropy:** Spectral entropy is used to quantify the irregularity and complexity of the EEG power spectrum.

These descriptors are stacked into a compact 14-row feature map per window (5 band-dynamics rows + 9 summary rows).

For ECG, a 5-dimensional HRV delta vector is extracted from the RR-interval sequence: heart rate, HRV coefficient of variation, RMSSD, SDNN, and pNN50 [32]. All descriptors are computed as stimulus minus baseline and concatenated into a compact low-dimensional representation.

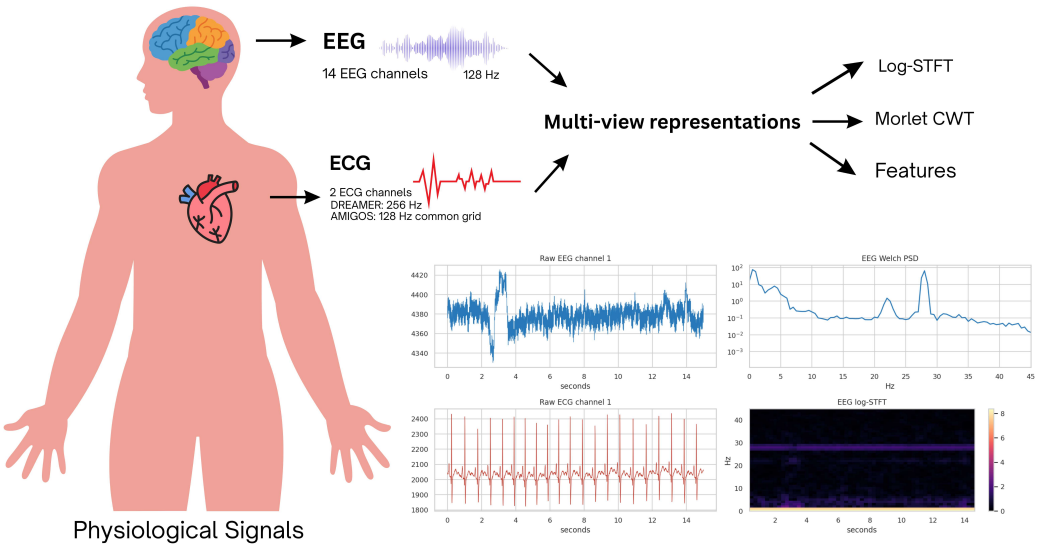


Figure 2: EEG and ECG data overview.

3.2.5. Short-Time Fourier Transform Representation

To preserve the evolution of frequency over time, each window is transformed into a log-STFT power map:

$$STFT(t, f) = \sum_{n=-\infty}^{\infty} x[n]w[n-t]e^{-j2\pi fn}, \tag{6}$$

where $w(\cdot)$ denotes a Hamming window. A 128-sample window with 96-sample overlap is used for EEG and a 64-sample overlap for ECG. The resulting spectrogram provides a two-dimensional time–frequency representation (frequency $\times$ time) in which intensity corresponds to log spectral power.

3.2.6. Morlet Continuous Wavelet Transform Representation

To capture transient emotional responses at different temporal scales, each window is represented using a 12-scale Morlet continuous wavelet transform (CWT):

$$W(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-b}{a} \right) dt, \tag{7}$$

where a and b denote the scale and translation parameters, respectively,

and $\psi(\cdot)$ represents the complex Morlet mother wavelet. The frequency range is set to 2–45 Hz for EEG and 0.8–40 Hz for ECG. The resulting scalogram is resampled to 24 time bins, yielding a compact log-CWT representation with adaptive time–frequency resolution.

STFT spectrograms and CWT scalograms are generated offline and cached prior to training. This strategy reduces computational overhead during model optimization by decoupling signal transformation from gradient-based learning and ensures consistent representations across different evaluation protocols.

3.3. MRF-CMFT Encoder Architecture

The MRF-CMFT employs a unified encoder architecture with independent branches for each representation. Each branch uses a lightweight convolutional feature extractor followed by a Mamba-based sequence modeling network.

3.3.1. CNN-Based Local Feature Extraction

For STFT and CWT images, the convolutional stem consists of stacked layers incorporating batch normalization, ReLU activation, and spatial pooling. Handcrafted features are projected into the same embedding dimension through a fully connected layer. Each representation is subsequently reshaped into a token sequence

$$Z = \{z_1, z_2, \ldots, z_N\}, \tag{8}$$

where N denotes the sequence length and $z_i \in R^d$ represents the i-th token.

3.3.2. Mamba-Based Sequence Modeling

The proposed framework employs the Mamba selective state-space model (SSM) as its primary sequence-modeling backbone. Given an input token sequence Z, the Mamba encoder updates its internal hidden state according to

$$h_t = f(h_{t-1}, z_t), \tag{9}$$

where h_t denotes the hidden state at step t. The selective state-space formulation enables efficient sequence processing with linear scaling with respect to sequence length, in contrast to the quadratic complexity of standard

self-attention. The final representation F is obtained through global average pooling:

$$F = \frac{1}{N} \sum_{i=1}^{N} h_i. \tag{10}$$

3.3.3. Unified Multi-View Encoding

Each modality $m \in \{EEG, ECG\}$ produces a set of encoded representations:

$$F_m = \left\{ F_m^{HF}, F_m^{STFT}, F_m^{CWT} \right\}, \tag{11}$$

where HF denotes the handcrafted-feature representation. Maintaining independent branches preserves the distinct statistical and structural characteristics of each representation before adaptive fusion.

3.4. Quality-Aware Adaptive Cross-Modal Fusion

The Quality-Aware Adaptive Cross-Modal Fusion (QACMF) module dynamically estimates modality reliability and adjusts the contribution of each modality to the final prediction.

3.4.1. Signal Quality Estimation

For EEG, the quality representation $q_{EEG} \in R^{d_q}$ is estimated from signal-to-noise ratio (SNR), artifact intensity, and channel consistency. For ECG, $q_{ECG} \in R^{d_q}$ is derived from RR-interval consistency and HRV stability. The modality-specific quality representations are concatenated as

$$q = [q_{EEG}; q_{ECG}]. \tag{12}$$

3.4.2. Adaptive View Selection

An AdaptiveViewSelector learns attention coefficients conditioned on both representation content and estimated signal quality. The quality representation is passed through a gating network:

$$g = \sigma(W_q q + b_q), \tag{13}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The resulting gating vector modulates each encoded representation through element-wise multiplication:

$$\widetilde{F}_i = g_i \odot F_i. \tag{14}$$

Consequently, representations associated with higher estimated quality receive stronger contributions, whereas artifact-corrupted inputs are attenuated.

3.4.3. Cross-Modal Feature Fusion

The quality-weighted modality representations are integrated into a fused representation:

$$F_{\text{fusion}} = \alpha_{EEG}\widetilde{F}_{EEG} + \alpha_{ECG}\widetilde{F}_{ECG}, \tag{15}$$

where the modality weights are obtained through

$$[\alpha_{EEG}, \alpha_{ECG}] = Softmax(W_f q + b_f), \tag{16}$$

such that

$$\alpha_{EEG} + \alpha_{ECG} = 1. \tag{17}$$

The resulting fused representation is subsequently passed to the final classification head.

3.5. Objective Function and Training Strategy

Training follows a two-stage procedure designed to learn robust unimodal representations before joint multimodal optimization.

3.5.1. Objective Function

The task is formulated as multi-task binary classification over the three affective dimensions: valence (v), arousal (a), and dominance (d). The binary cross-entropy (BCE) loss for an individual affective dimension is defined as

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{18}$$

where N is the number of training samples, y_i is the ground-truth binary label, and $\hat{y}_i$ is the predicted probability.

The overall training objective is obtained by averaging the losses across the three affective dimensions:

$$\mathcal{L} = \frac{\mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_d}{3}. \tag{19}$$

3.5.2. Two-Stage Training Strategy

The proposed framework is optimized in two stages:

- **Stage I – Unimodal Pretraining:** The EEG and ECG branches are pretrained independently for 20 epochs to learn modality-specific, task-relevant representations before multimodal interaction is introduced.
- **Stage II – Joint Multimodal Fine-Tuning:** The pretrained encoders are connected through the QACMF module. The pretrained encoders remain frozen during the first five fusion warm-up epochs, allowing the newly initialized fusion parameters to stabilize. The encoders are then unfrozen and the complete network is jointly optimized for up to 40 epochs.

The complete training and inference procedure of the proposed MRF-CMFT framework is summarized in Algorithm 1.

Algorithm 1: Training Procedure of the Proposed MRF-CMFT Framework

Input: EEG recordings X_{EEG}, ECG recordings X_{ECG}
Output: Trained MRF-CMFT model

Preprocess the EEG and ECG recordings
Segment each trial into fixed-length analysis windows
Generate handcrafted features, STFT spectrograms, and CWT scalograms for each window
Compute modality-specific signal-quality metrics
Initialize the EEG encoder, ECG encoder, and AdaptiveViewSelector
for *epoch* $= 1$ *to* 20 **do**
    Pretrain the EEG encoder
for *epoch* $= 1$ *to* 20 **do**
    Pretrain the ECG encoder
Initialize the multimodal fusion stage using the pretrained encoders
Freeze the pretrained encoders
for *epoch* $= 1$ *to* 5 **do**
    Train the AdaptiveViewSelector and fusion layers
Unfreeze all network parameters
for *epoch* $= 1$ *to* 40 **do**
    Train the complete MRF-CMFT framework using AMP
    Update the EMA model weights
    Adjust the learning rate using cosine annealing with warm-up
    Evaluate on the validation set
    if *Validation Macro-F1 improves* **then**
        Save the best model checkpoint
    if *Early stopping criterion is satisfied* **then**
        break
return Trained MRF-CMFT model

4. Experimental Setup

4.1. Datasets

The MRF-CMFT framework was evaluated on two benchmark datasets: **DREAMER** and **AMIGOS**. These datasets were selected to assess multimodal performance across varying recording qualities and participant scales.

- **DREAMER**: The DREAMER dataset includes 23 participants exposed to 18 audiovisual stimuli. EEG recordings consist of 14 channels sampled at 128 Hz, accompanied by ECG recordings sampled at the same frequency. Following the preprocessing strategy described in Section 3, only the final 60 seconds of each trial were retained. The self-reported valence, arousal, and dominance ratings were binarized using the midpoint of the original 1–5 rating scale.

- **AMIGOS**: The AMIGOS dataset comprises 40 participants and provides a more challenging multimodal physiological benchmark. Compared with DREAMER, the recordings exhibited a higher prevalence of non-finite EEG samples and other signal-quality issues, requiring additional sanitization before feature extraction. The same representation and evaluation pipeline was subsequently applied to maintain methodological consistency across datasets.

Preprocessing discipline. All filtering, baseline correction, and window normalization are trial-local feature and metadata scalers are fitted on training indices only no statistic from held-out subjects is used at any point. The pipeline follows the conventions of standardized EEG preprocessing [51] adapted to the trial structure of DREAMER and AMIGOS.

4.2. Evaluation Protocols

To reduce overly optimistic performance estimates caused by temporal correlation between analysis windows, three leakage-safe evaluation protocols were employed. All protocols are subject- or trial-disjoint by construction.

1. **Subject-Dependent (SD):** A K-fold trial-grouped cross-validation protocol was employed (10 folds for DREAMER, 5 folds for AMIGOS). Complete trials were assigned to individual folds grouped by `(subject, trial)`, ensuring that windows originating from the same trial could not appear simultaneously in the training and test partitions. The

same subject may appear in both training and test folds, which is the intended within-subject regime.

2. **Subject-Independent (SI):** Approximately 20% of subjects were held out exclusively for testing, with 2–4 subjects used for validation and the remainder for training. Five random seeds (42, 52, 62, 72, and 82) were used. Assertions enforce pairwise subject-disjointness between training, validation, and test partitions, providing an evaluation of generalization to previously unseen subjects.
3. **Leave-One-Subject-Out (LOSO):** Each participant was iteratively withheld as the test subject, with a dedicated validation subject excluded from training. This protocol provides a stringent assessment of cross-subject generalization and enables the characterization of subject-level performance variability.

In addition to the primary evaluation protocols, an exploratory **window-level random split** was performed to quantify the effect of trial-level information leakage. In this setting, individual analysis windows were randomly shuffled and partitioned without enforcing trial boundaries. Consequently, highly correlated windows originating from the same trial could appear in both training and test partitions. This protocol is not considered a valid measure of generalization and is therefore reported separately from the leakage-safe results.

4.3. Implementation Details

All models were implemented using PyTorch and trained on a single NVIDIA Tesla T4 GPU. The training procedure followed the two-stage strategy described in Section 3, consisting of 20 epochs of independent unimodal pretraining followed by up to 40 epochs of joint multimodal fine-tuning.

The AdamW optimizer was employed for parameter optimization. During the fusion stage, a reduced learning rate of 6×10^{-5} was applied to the pretrained encoders, while newly initialized fusion parameters were optimized using a higher learning rate. A cosine annealing learning-rate schedule with linear warm-up was employed to improve optimization stability.

Automatic Mixed Precision (AMP) was used to reduce computational overhead, while an Exponential Moving Average (EMA) of the model parameters was maintained to improve training stability. Early stopping was performed using validation Macro-F1 with a patience of five epochs.

For reproducibility, subject-independent experiments were repeated using five random seeds: 42, 52, 62, 72, and 82. All normalization statistics were computed exclusively from the corresponding training partition within each evaluation fold to prevent information leakage.

4.4. Evaluation Metrics

Performance was evaluated using per-task accuracy, balanced accuracy, precision, recall, F1-score, Macro-F1, and the area under the receiver operating characteristic curve (AUC).

Macro-F1 is computed across the three affective dimensions—valence, arousal, and dominance—and is used as the primary aggregate classification metric. For each evaluation fold, decision thresholds are selected using the corresponding validation partition. Aggregate results are reported as mean $\pm$ standard deviation where repeated folds or subjects are available.

5. Results and Discussion

This section presents a comprehensive evaluation of the proposed MRF-CMFT framework on the DREAMER and AMIGOS datasets. The evaluation examines predictive performance under leakage-controlled subject-dependent, subject-independent, and leave-one-subject-out protocols, followed by robustness analysis under controlled signal degradations. the results are compared with representative state-of-the-art methods under comparable evaluation settings.

5.1. Main Results on DREAMER

Table 1 and Fig. 3 summarize the trial-level performance of MRF-CMFT on the DREAMER dataset under the three leakage-safe evaluation protocols. Overall, the model achieves predictive performance above chance for most protocol–label combinations, although absolute scores remain moderate and substantially lower than those obtained under leakage-permitting window-level evaluation.

Under the **subject-dependent (SD)** trial-grouped 10-fold protocol, the model obtains an overall mean AUC of 0.614, with all three affective dimensions performing near or above 0.60 (valence 0.600, arousal 0.614, dominance 0.628). The corresponding Macro-F1 scores are 0.554, 0.587, and 0.567, respectively. These results indicate that, when subject identity is allowed to

overlap between training and test folds, the physiological signals contain usable affective information, although the three dimensions are not equally predictable.

Performance becomes more challenging under the stricter **subject-independent (SI)** protocol (five random seeds). Mean AUCs are 0.640 / 0.534 / 0.581 for valence / arousal / dominance. Valence is the most stable task (SD = 0.036), while arousal exhibits the highest variability (SD = 0.076). The corresponding Macro-F1 scores drop to 0.595, 0.479, and 0.510. This degradation relative to the SD setting highlights the difficulty of generalizing to completely unseen subjects.

The most stringent evaluation is provided by **leave-one-subject-out (LOSO)** cross-validation. Overall mean AUC falls to 0.573, with substantial inter-subject dispersion (valence 0.629±0.153, arousal 0.499±0.168, dominance 0.590±0.208). Medians (0.613 / 0.507 / 0.625) are higher than the means for valence and dominance, indicating that a small number of particularly difficult held-out subjects pull the averages downward. Macro-F1 values under LOSO are 0.528, 0.443, and 0.472. The large standard deviations across subjects confirm that cross-subject generalization remains a major challenge in physiological emotion recognition.

Taken together, the results show a clear performance hierarchy (SD > SI > LOSO) and reveal that valence is generally the most robust dimension under subject-independent conditions, while arousal is the most sensitive to inter-subject variability.

Table 1: Trial-level results of MRF-CMFT on DREAMER (23 subjects, 414 trials). Values are mean ± SD across SI seeds (5), SD folds (10), and LOSO folds (23) for LOSO, medians are shown in parentheses.

Protocol	Metric	Valence	Arousal	Dominance	Overall
SI (5 seeds)	Accuracy	0.604 ± 0.030	0.507 ± 0.069	0.531 ± 0.038	0.547
	Balanced acc.	0.614 ± 0.021	0.491 ± 0.050	0.532 ± 0.050	0.546
	Macro-F1	0.595 ± 0.025	0.479 ± 0.052	0.510 ± 0.038	0.528
	AUC	**0.640 ± 0.036**	0.534 ± 0.076	0.581 ± 0.057	0.585
SD trial-grouped	Accuracy	0.568 ± 0.095	0.601 ± 0.042	0.582 ± 0.097	0.584
	Balanced acc.	0.573 ± 0.096	0.594 ± 0.042	0.589 ± 0.083	0.585
	Macro-F1	0.554 ± 0.097	0.587 ± 0.044	0.567 ± 0.104	0.569
	AUC	0.600 ± 0.119	0.614 ± 0.061	0.628 ± 0.078	0.614
LOSO (23 folds)	Accuracy	0.614 ± 0.161	0.522 ± 0.140	0.544 ± 0.158	0.560
	Balanced acc.	0.570 ± 0.144	0.503 ± 0.119	0.533 ± 0.139	0.535
	Macro-F1	0.528 ± 0.173	0.443 ± 0.134	0.472 ± 0.163	0.481
	AUC	0.629 ± 0.153	0.499 ± 0.168	0.590 ± 0.208	0.573

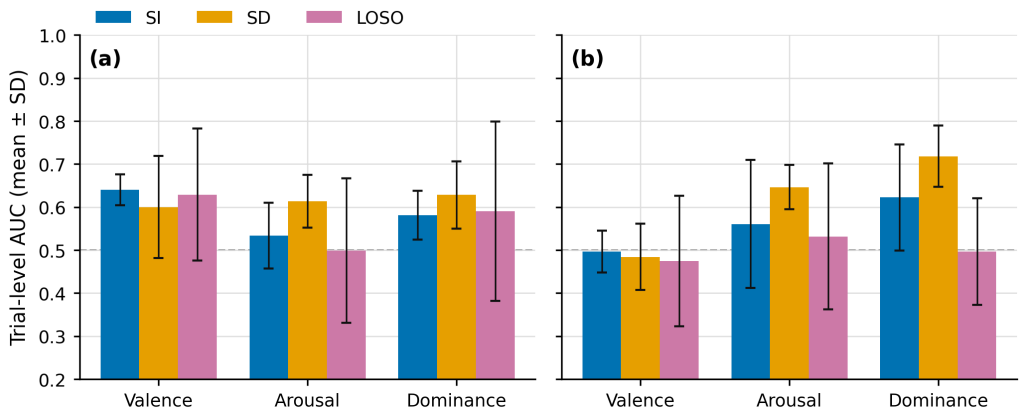


Figure 3: Trial-level AUC (mean ± SD) of MRF-CMFT under the three leakage-controlled protocols, per task, on (a) DREAMER and (b) AMIGOS. SI: five-seed subject-independent SD: subject-dependent trial-grouped CV (10 folds DREAMER, 5 folds AMIGOS) LOSO: leave-one-subject-out.

5.2. Main Results on AMIGOS

Table 2 reports the performance of the same MRF-CMFT pipeline on the 15-subject AMIGOS mirror (288 trials). Overall, results are substantially lower and more variable than those observed on DREAMER, reflecting both the smaller sample size and the higher prevalence of signal-quality issues in this subset.

Under the **subject-independent (SI)** protocol, mean AUCs are 0.496 / 0.561 / 0.623 for valence / arousal / dominance. Dominance is clearly the strongest dimension across all protocols, while valence remains close to chance level under both SI and SD settings. Arousal again exhibits the largest seed-to-seed variability (SI AUC ranging from 0.301 to 0.677).

In the **subject-dependent (SD)** 5-fold trial-grouped evaluation, dominance reaches a mean AUC of 0.718 ± 0.071, confirming that it is the most reliably detectable affective dimension when subject identity can overlap between folds. Valence continues to perform near chance.

The most challenging setting is **leave-one-subject-out (LOSO)**. The overall mean AUC drops to approximately 0.501 (i.e., near chance on average). Individual subject-level AUCs range widely from 0.19 to 0.95 depending on the task, indicating extreme inter-subject heterogeneity. This high variability is consistent with the known recording-quality problems of several participants (particularly subjects 22–24) and with the limited number of subjects available in the AMIGOS mirror.

Collectively, the AMIGOS results highlight two important points: (1) dominance is the most robust affective dimension across protocols, and (2) cross-subject generalization remains extremely difficult on this smaller and noisier dataset.

Table 2: Trial-level results of MRF-CMFT on the AMIGOS mirror (15 subjects, 288 trials). Values are mean ± SD across SI seeds (5), SD folds (5), and LOSO folds (15) for LOSO, medians are shown in parentheses.

Protocol	Metric	Valence	Arousal	Dominance	Overall
SI (5 seeds)	Accuracy	0.485 ± 0.050	0.554 ± 0.161	0.616 ± 0.106	0.552
	Balanced acc.	0.489 ± 0.056	0.579 ± 0.058	0.598 ± 0.108	0.556
	Macro-F1	0.439 ± 0.056	0.510 ± 0.159	0.585 ± 0.112	0.512
	AUC	0.496 ± 0.049	0.561 ± 0.149	0.623 ± 0.123	0.560
SD trial-grouped	Accuracy	0.483 ± 0.042	0.615 ± 0.068	0.698 ± 0.119	0.598
	Balanced acc.	0.471 ± 0.046	0.622 ± 0.062	0.670 ± 0.097	0.588
	Macro-F1	0.457 ± 0.045	0.600 ± 0.072	0.674 ± 0.108	0.577
	AUC	0.484 ± 0.077	0.646 ± 0.051	0.718 ± 0.071	0.616
LOSO (15 folds)	Accuracy	0.498 ± 0.114	0.456 ± 0.160	0.518 ± 0.160	0.490
	Balanced acc.	0.488 ± 0.093	0.479 ± 0.119	0.481 ± 0.073	0.482
	Macro-F1	0.418 ± 0.115	0.387 ± 0.139	0.399 ± 0.113	0.401
	AUC	0.474 ± 0.151	0.532 ± 0.170	0.496 ± 0.124	0.501

5.3. Robustness to Low-Quality Data

Fig. 4 and Fig. 5, together with Tables 3–4, present the controlled low-quality robustness evaluation of MRF-CMFT on DREAMER and AMIGOS. The analysis includes clean inputs and seven controlled degradation conditions, covering modality-specific noise, simultaneous noise, missing modalities, temporal misalignment, and severe corruption. For each scenario, accuracy, balanced accuracy, Macro-F1, and AUC are reported at the trial level. Considering these metrics together provides complementary information about both discrete classification behavior and ranking quality.

On DREAMER, MRF-CMFT demonstrates a clear degree of robustness across the evaluated corruption conditions, although the effect of degradation depends on the affective dimension and the corrupted input. The most pronounced deterioration occurs when the EEG modality is missing, particularly for valence, followed by smaller degradations for arousal and dominance. By comparison, ECG noise, missing ECG, and temporal misalignment produce comparatively limited changes in the reported metrics. The consistency of the four metrics across scenarios indicates that the observed robustness is not confined to AUC; it is also reflected in final classification performance, as measured by accuracy, balanced accuracy, and Macro-F1 (Fig. 5).

The AMIGOS evaluation reveals a different task-dependent robustness profile (Fig. 4). MRF-CMFT maintains relatively stable performance across most corruption scenarios, with dominance exhibiting the strongest and most consistent behavior among the three affective dimensions. In particular, the dominance AUC remains above 0.83 across the evaluated scenarios, while the corresponding accuracy, balanced accuracy, and Macro-F1 values show substantially less variation than those of the weaker affective dimensions. Valence remains comparatively difficult, whereas arousal is more sensitive to EEG degradation. This pattern is consistent with the dataset-level variability observed in the subject-independent evaluation.

The different robustness profiles of DREAMER and AMIGOS demonstrate that low-quality degradation does not affect all datasets or affective dimensions uniformly. The stronger sensitivity to missing EEG on DREAMER, in particular, suggests that the contribution of a modality depends on both dataset characteristics and task-specific signal content. These results also indicate why a single global fusion rule may be inadequate for heterogeneous physiological recordings. Instead, the quality-aware gates dynamically modulate the contribution of individual views and modalities according to their estimated signal quality. Together with corruption-aware training, this mechanism allows degraded inputs to be down-weighted while informative modalities and views remain available for prediction.

Overall, the results support the intended role of MRF-CMFT as a quality-aware fusion framework rather than merely a clean-data classifier. The scenario-level analysis, combined with four complementary evaluation metrics, shows that the framework can preserve useful predictive information across a range of controlled low-quality conditions. At the same time, the observed task- and dataset-specific failure modes identify limitations that remain relevant for practical deployment and motivate further investigation of naturally occurring artifacts, missing inputs, and cross-dataset robustness.

Table 3: DREAMER robustness: MRF-CMFT trial-level AUC per scenario (SI, seed 42). The largest clean-to-scenario drop per task is bold.

Scenario	Valence	Arousal	Dominance
Clean	0.622	0.658	0.651
EEG noisy	0.595	0.640	0.641
ECG noisy	0.628	0.674	0.657
Missing EEG	0.503	0.568	0.576
Missing ECG	0.622	0.662	0.628
Both noisy	0.611	0.644	0.640
Misaligned	0.623	0.658	0.655
Corrupted	0.597	0.655	0.661

Table 4: AMIGOS robustness: MRF-CMFT trial-level AUC per scenario (SI, seed 42; test subjects 15, 22, and 24). Dominance remains above 0.83 under all eight scenarios.

Scenario	Valence	Arousal	Dominance
Clean	0.539	0.660	0.858
EEG noisy	0.540	0.605	0.863
ECG noisy	0.539	0.668	0.855
Missing EEG	0.554	0.668	0.848
Missing ECG	0.558	0.659	0.832
Both noisy	0.569	0.612	0.858
Misaligned	0.537	0.658	0.858
Corrupted	0.527	0.633	0.855

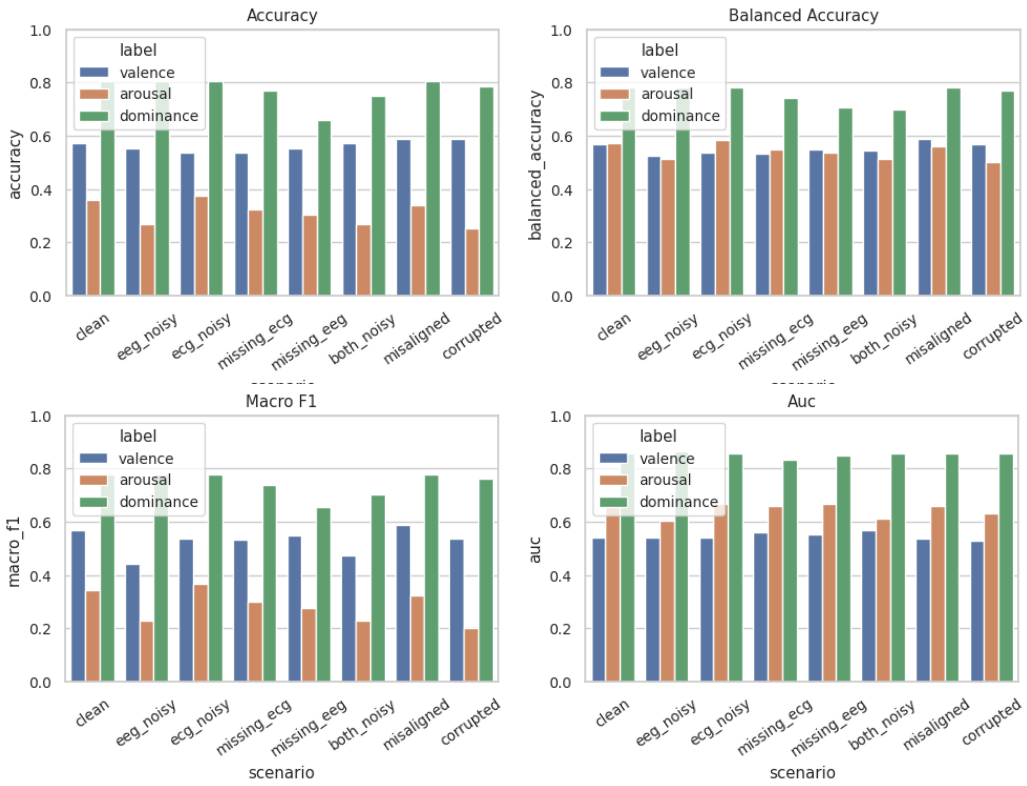


Figure 4: Low-quality robustness evaluation of MRF-CMFT on AMIGOS. Trial-level accuracy, balanced accuracy, macro-F1, and AUC are shown for clean inputs and the seven controlled corruption conditions across valence, arousal, and dominance.

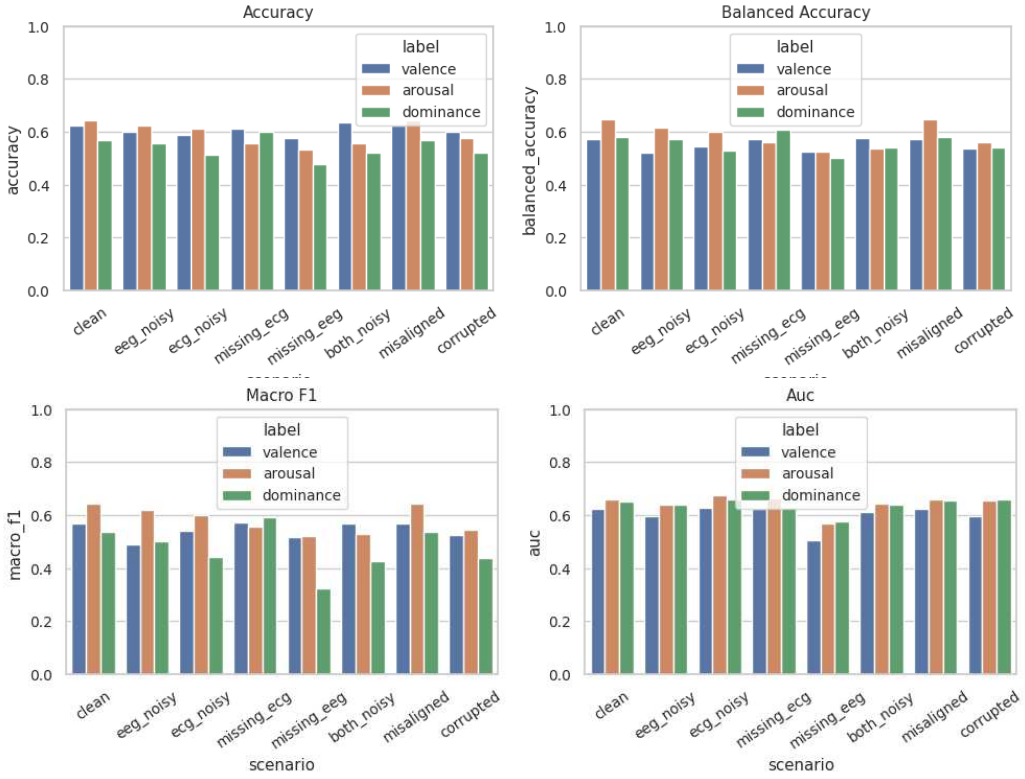


Figure 5: Low-quality robustness evaluation of MRF-CMFT on DREAMER. Trial-level accuracy, balanced accuracy, macro-F1, and AUC are shown for clean inputs and the seven controlled corruption conditions across valence, arousal, and dominance.

5.4. The Evaluation-Protocol effect

Fig. 6 contrasts the valid trial-level results with an *exploratory* random window-level split included solely to reproduce the high numbers that appear in the literature. When windows of the same trial are allowed in both training and test, window-level AUC reaches 0.987–0.995 with the same DREAMER data, representations, and preprocessing, whereas the leakage-controlled trial-level protocols yield 0.51–0.64—an inflation of roughly 0.35–0.49 AUC driven by near-duplicate windows of the same trial appearing in both training and test. This exploratory split is intentionally separate from the valid protocol suite and must never be interpreted as generalization performance it quantifies the inflation that motivates our strict protocol design.

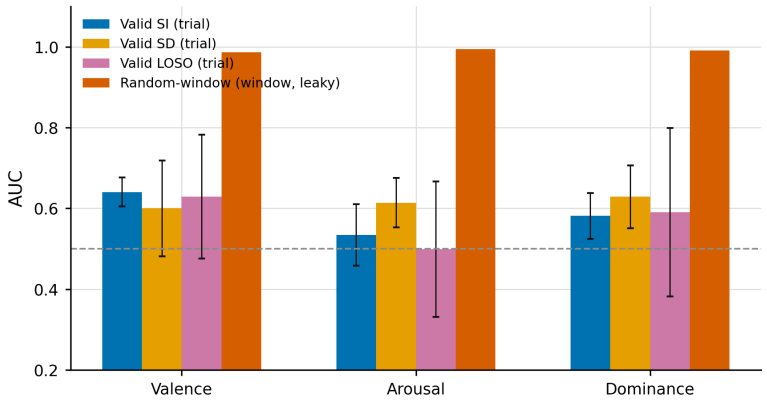


Figure 6: Valid trial-level AUC of MRF-CMFT on DREAMER (SI five-seed, SD trial-grouped 10-fold, LOSO 23-fold) versus an exploratory random window-level split of the same data (leaky by design, included only to reproduce an exploratory window-level split representative of commonly used evaluation practices). Random-window evaluation inflates AUC to ≈0.99 versus 0.51–0.64 under leakage-controlled protocols.

5.5. Comparison with State-of-the-Art

Table 5 compares MRF-CMFT with representative subject-independent methods on the DREAMER dataset. To ensure a fair comparison, only methods evaluated under leave-one-subject-out (LOSO) or equivalent subject-independent protocols are considered. The comparison focuses exclusively on valence classification accuracy.

Table 5: Valence accuracy comparison between MRF-CMFT and representative subject-independent methods on the DREAMER dataset. Bold indicates the best accuracy among the reported methods.

Study / Method	**Protocol**	**Valence Accuracy (%)**	**Ours Gain (pp)**
SVM [5]	LOSO	56.57	+4.83
DBN [17]	LOSO	56.43	+4.97
DGCNN [25]	LOSO	59.64	+1.76
FCN-LA [52]	LOSO	60.63	+0.77
JDA-NN [53]	LOSO	60.55	+0.85
IIR-formulated SNN [54]	LOSO	54.03	+7.37
NeuroSense [55]	LOSO	52.90	+8.50
BiDANN [56]	LOSO	59.98	+1.42
RGNN [57]	LOSO	54.68	+6.72
MRF-CMFT (Proposed)	**LOSO**	**61.40**	–

The proposed MRF-CMFT achieves a valence accuracy of 61.40% under

LOSO evaluation, outperforming all selected subject-independent baselines included in the comparison. The improvement ranges from 0.77 percentage points over FCN-LA to 8.50 percentage points over NeuroSense. These results demonstrate that MRF-CMFT provides competitive valence recognition under a strict subject-independent evaluation protocol. "'

5.6. Limitations

Several limitations should be considered when interpreting the results. First, as shown in Table 2, the AMIGOS experiments used a 15-subject mirror of the full 40-subject release. Consequently, the reported values are subset-specific and are not directly comparable with results obtained from the complete AMIGOS dataset. The approximately chance-level mean observed under AMIGOS LOSO evaluation may reflect both the limited size of the mirror and the extensive handling of non-finite data for participants 22–24; it should therefore not be generalized to the full release. Because the held-out splits are small, subject-level dispersion is used as the primary estimate of uncertainty. Paired statistical tests are not treated as the principal basis for inference because they would be underpowered in this setting.

Second, some robustness experiments injected noise without a fixed random seed. These results should therefore be interpreted as controlled robustness observations rather than averaged estimates of noise-related variability. Future studies should repeat each scenario across multiple noise seeds and report mean performance with confidence intervals.

Third, the component ablations of cross-attention, gating, quality estimation, and individual modalities used a reduced training budget. They consequently do not provide definitive evidence of the independent contribution of each component. Matched full-budget ablations with identical optimization settings are needed for more reliable attribution of performance gains. In addition, the gradient-based explanations generated using Grad-CAM and Integrated Gradients were not fully validated quantitatively. Future work should evaluate explanation faithfulness, stability, and relevance using quantitative measures.

Finally, the binarization thresholds differ because the self-report scales of the two datasets are different: values greater than 3.0 are used for DREAMER, whereas values greater than 5.0 are used for AMIGOS. These thresholds are reported explicitly to support reproducibility and to prevent direct comparison of label distributions without accounting for the underlying annotation scales.

6. Conclusion and Future Work

This study proposed MRF-CMFT, a quality-aware multimodal framework for EEG-ECG emotion recognition that integrates complementary handcrafted and time-frequency representations through CNN-Mamba encoders and adaptive cross-modal fusion. Experiments on DREAMER and AMIGOS demonstrate the potential of combining neural and autonomic physiological information while highlighting substantial inter-subject variability under strict cross-subject evaluation. Importantly, the controlled leakage analysis revealed that window-level random splitting can produce substantial and misleading performance inflation, emphasizing the necessity of trial- and subject-aware evaluation protocols. The results further support quality-aware fusion as a promising strategy for improving the robustness of multimodal physiological recognition under degraded signal conditions. Future work will focus on stronger subject-independent representation learning, uncertainty-aware fusion, and evaluation on larger and naturally degraded physiological datasets.

Declarations

Ethics approval and consent to participate: This article does not contain any studies with human participants or animals performed by any of the authors.
Consent for publication: All authors have read and approved the final version of the manuscript and give their consent for its publication in this journal.
Availability of data and material: Data is available from the authors upon request.
Competing interests: The authors declare that there is no conflict of interest regarding the publication of this paper.
Funding: Not Applicable.
Authors' contributions: H. Mostafad: Acquisition of data, analysis, and interpretation of data; conception and design of methodology and model; software design. **N. M.Elsherbinid**: Review, editing, validation, and verification of experimental results. **M. Samy**: Interpretation of data, conception and design of methodology, drafting the manuscript. **A. Anter**: Conception and design of study, formal analysis and visualisation, writing and revising the manuscript for important intellectual content, investigation and supervision.

References

[1] R. W. Picard, Affective Computing, MIT Press, Cambridge, MA, USA, 2000.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine 18 (1) (2001) 32–80.

[3] S. M. Alarcão, M. J. Fonseca, Emotions recognition using EEG signals: A survey, IEEE Transactions on Affective Computing 10 (3) (2019) 374–393.

[4] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (1) (2012) 18–31. doi:10.1109/T-AFFC.2011.15.

[5] S. Katsigiannis, N. Ramzan, DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices, IEEE Journal of Biomedical and Health Informatics 22 (1) (2018) 98–107. doi:10.1109/JBHI.2017.2688239.

[6] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, I. Patras, AMIGOS: A dataset for affect, personality and mood research on individuals and groups, IEEE Transactions on Affective Computing 12 (2) (2021) 479–493. doi:10.1109/TAFFC.2018.2884461.

[7] Z. Ahmad, N. Khan, A survey on physiological signal-based emotion recognition, Bioengineering 9 (11) (2022) 688.

[8] A. Apicella, P. Arpaia, G. D'Errico, D. Marocco, G. Mastrati, N. Moccaldi, R. Prevete, Toward cross-subject and cross-session generalization in eeg-based emotion recognition: Systematic review, taxonomy, and methods, Neurocomputing 604 (2024) 128354.

[9] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2) (2019) 423–443. doi:10.1109/TPAMI.2018.2798607.

[10] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6558–6569.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[13] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023). doi:10.48550/arXiv.2312.00752.

[14] T. Dao, A. Gu, Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality, arXiv preprint arXiv:2405.21060 (2024).

[15] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Transactions on Image Processing 24 (12) (2015) 5812–5825.

[16] P. C. Petrantonakis, L. J. Hadjileontiadis, Emotion recognition from EEG using higher order crossings, IEEE Transactions on Information Technology in Biomedicine 14 (2) (2010) 186–197. doi:10.1109/TITB.2009.2034649.

[17] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Transactions on Autonomous Mental Development 7 (3) (2015) 162–175.

[18] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, B. Hu, Exploring eeg features in cross-subject emotion recognition, Frontiers in Neuroscience 12 (2018) 162.

[19] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, WIREs Data Mining and Knowledge Discovery 8 (4) (2018) e1253. doi:10.1002/widm.1253.

[20] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces, Journal of Neural Engineering 15 (5) (2018) 056013.

[21] T. Papalexopoulos, C. Tzolov, B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. De Freitas, Constrained discrete black-box optimization using mixed-integer programming, in: Proceedings of the 39th International Conference on Machine Learning (ICML), Vol. 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 17382–17402.

[22] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, M. Faisal, Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review, Neural Computing and Applications 35 (20) (2023) 14681–14722. doi:10.1007/s00521-021-06352-5.

[23] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[24] H. Chao, L. Dong, Y. Liu, B. Lu, Improved deep feature learning by synchronization measurements for multi-channel EEG emotion recognition, Complexity 2020 (2020) 6816502. doi:10.1155/2020/6816502.

[25] T. Song, W. Zheng, P. Song, Z. Cui, Emotion recognition using dynamical graph convolutional neural networks, IEEE Transactions on Affective Computing 11 (3) (2020) 532–541.

[26] S. Asadzadeh, T. Yousefi Rezaii, S. Beheshti, S. Meshgini, Accurate emotion recognition using bayesian model based eeg sources as dynamic graph convolutional neural network nodes, Scientific Reports 12 (1) (2022) 10282. doi:10.1038/s41598-022-14217-7.

[27] W. Chen, Y. Liao, R. Dai, Y. Dong, L. Huang, Eeg-based emotion recognition using graph convolutional neural network with dual attention mechanism, Frontiers in Computational Neuroscience 18 (2024) 1416494. doi:10.3389/fncom.2024.1416494.

[28] X. Cheng, et al., Regional-to-global spatiotemporal graph transformer for EEG-based emotion recognition (2024).

[29] Y. Ding, C. Tong, J. Park, W.-L. Zheng, B.-L. Lu, C. Guan, EmT: A novel transformer for generalized cross-subject EEG emotion recognition, IEEE Transactions on Neural Networks and Learning SystemsArXiv:2406.18345 (2025).

[30] Y. Yin, X. Zheng, B. Hu, Y. Zhang, X. Cui, EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM, Applied Soft Computing 100 (2021) 106954. doi:10.1016/j.asoc.2020.106954.

[31] Y. Zhou, et al., Dual-scale temporal graph convolutional network for EEG signal representation (2025).

[32] F. Shaffer, J. P. Ginsberg, An overview of heart rate variability metrics and norms, Frontiers in Public Health 5 (2017) 258.

[33] A. Chowdhury, et al., A CNN-LSTM architecture for multimodal emotion recognition using physiological signals, IEEE Transactions on Affective Computing 11 (3) (2020) 421–433.

[34] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE Journal of Selected Topics in Signal Processing 11 (8) (2017) 1301–1309. doi:10.1109/JSTSP.2017.2764438.

[35] Y. Wang, et al., Intra-modality and inter-modality attention network for multimodal emotion recognition, IEEE Transactions on Cognitive and Developmental Systems 16 (2) (2024) 612–624.

[36] Z. Wan, et al., Low-rank tensor fusion with cross-modal attention for subject-independent emotion recognition, Biomedical Signal Processing and Control 88 (2024) 105612.

[37] H. Li, et al., Dynamic graph convolution and temporal self-attention for joint EEG-ECG functional connectivity analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 33 (2025) 115–126.

[38] J. Wang, et al., Attention-enhanced CNN-GRU architecture for multimodal EEG-ECG emotion recognition, Computers in Biology and Medicine 170 (2025) 108012.

[39] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634 (2013).

[40] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE Transactions on Knowledge and Data Engineering 31 (10) (2019) 1863–1883.

[41] R. Wu, H. Wang, H.-T. Chen, G. Carneiro, Deep multimodal learning with missing modality: A survey, Transactions on Machine Learning Research (2024).

[42] J. Tang, Q. Yi, S. Fu, Y. Tian, Incomplete multi-view learning: Review, analysis, and prospects, Applied Soft Computing 153 (2024) 111278.

[43] X. Meng, S. Wei, General incomplete multimodal learning via dynamic quality perception, arXiv preprint arXiv:2607.06943 (2026).

[44] Y. Xie, et al., Adaptive uncertainty-aware alignment and fusion network for multimodal EEG emotion recognition, IEEE Transactions on Affective Computing 16 (1) (2025) 88–101.

[45] X. Zhu, et al., Uncertainty-guided multimodal fusion for robust physiological emotion recognition, Information Fusion 115 (2025) 102714.

[46] A. Gu, T. Dao, S. Ermon, A. Rudra, C. Ré, HiPPO: Recurrent memory with optimal polynomial projections, in: Advances in Neural Information Processing Systems, 2020.

[47] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, C. Ré, Combining recurrent, convolutional, and continuous-time models with linear state space layers, in: Advances in Neural Information Processing Systems, 2021.

[48] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, in: International Conference on Learning Representations, 2022.

[49] J. T. H. Smith, A. Warrington, S. W. Linderman, Simplified state space layers for sequence modeling, in: International Conference on Learning Representations, 2023.

[50] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision Mamba: Efficient visual representation learning with bidirectional state space model, arXiv preprint arXiv:2401.09417 (2024).

[51] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, K. A. Robbins, The prep pipeline: Standardized preprocessing for large-scale eeg analysis, Frontiers in Neuroinformatics 9 (2015) 16.

[52] P. Zhong, D. Wang, C. Miao, Eeg-based emotion recognition using regularized graph neural networks, IEEE Transactions on Affective Computing 13 (3) (2020) 1290–1301.

[53] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, X. Zhou, A bi-hemisphere domain adversarial neural network model for eeg emotion recognition, IEEE Transactions on Affective Computing 12 (2) (2021) 494–504.

[54] C. Tan, M. Sarlija, N. Kasabov, Neurosense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal eeg patterns, Neurocomputing 434 (2021) 137–148.

[55] P. Gong, P. Wang, Y. Zhou, et al., Fractal spiking neural network scheme for eeg-based emotion recognition, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023).

[56] Y. Li, W. Zheng, Z. Cui, T. Zhang, Y. Zong, A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI) (2018) 1561–1567.

[57] W.-L. Zheng, J.-Y. Zhu, Y. Peng, B.-L. Lu, Eeg-based emotion classification using deep belief networks, 2014 IEEE International Conference on Multimedia and Expo (ICME) (2014) 1–6.