AI

MDPI

Review

Multimodal Classification Algorithms for Emotional Stress Analysis with an ECG-Centered Framework: A Comprehensive Review

Xinyang Zhang, Haimin Zhang and Min Xu *

Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia; xinyang.zhang-2@student.uts.edu.au (X.Z.); haimin.zhang@uts.edu.au (H.Z.)
* Correspondence: min.xu@uts.edu.au

Abstract

Emotional stress plays a critical role in mental health conditions such as anxiety, depression, and cognitive decline, yet its assessment remains challenging due to the subjective and episodic nature of conventional self-report methods. Multimodal physiological approaches, integrating signals such as electrocardiogram (ECG), electrodermal activity (EDA), and electromyography (EMG), offer a promising alternative by enabling objective, continuous, and complementary characterization of autonomic stress responses. Recent advances in machine learning and artificial intelligence (ML/AI) have become central to this paradigm, as they provide the capacity to model nonlinear dynamics, inter-modality dependencies, and individual variability that cannot be effectively captured by rule-based or single-modality methods. This paper reviews multimodal physiological stress recognition with an emphasis on ECG-centered systems and their integration with EDA and EMG. We summarize stress-related physiological mechanisms, catalog public and self-collected databases, and analyze their ecological validity, synchronization, and annotation practices. We then examine preprocessing pipelines, feature extraction methods, and multimodal fusion strategies across different stages of model design, highlighting how ML/AI techniques address modality heterogeneity and temporal misalignment. Comparative analysis shows that while deep learning models often improve within-dataset performance, their generalization across subjects and datasets remains limited. Finally, we discuss open challenges and future directions, including self-supervised learning, domain adaptation, and standardized evaluation protocols. This review provides practical insights for developing robust, generalizable, and scalable multimodal stress recognition systems for mental health monitoring.

Keywords: emotional stress; multimodal physiological signals; electrocardiogram; electrodermal activity; multimodal fusion; machine learning

Check for updates

Academic Editor: Steven R. Livingstone

Received: 1 December 2025
Revised: 19 January 2026
Accepted: 3 February 2026
Published: 9 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Emotional stress has become a global public health challenge and is closely associated with anxiety, depression, and cognitive decline, exerting negative impacts on both daily life and work performance [1]. Recent epidemiological reports indicate a sustained rise in mental health disorders worldwide, with anxiety and depressive disorders affecting more than 300 million people globally, and accounting for a substantial proportion of years lived with disability [2,3]. Moreover, chronic stress exposure has been increasingly linked to accelerated cognitive decline and elevated risk of neurodegenerative disorders, particularly in aging populations.

Current stress assessment approaches predominantly rely on self-report questionnaires, which are highly subjective and insufficient for objective, continuous, and real-time monitoring [4,5]. In this context, physiological signals such as electrocardiogram (ECG), electromyography (EMG), and electrodermal activity (EDA) have gained increasing attention for their ability to reflect dynamic changes in the autonomic nervous system [6]. Multimodal physiological fusion techniques integrate complementary information across different signal channels, and combining these modalities with subjective scales is of great importance for enhancing psychological health monitoring in daily life.

In this review, emotional stress is adopted as an operational concept referring to stress-related physiological responses elicited or modulated by emotionally salient stimuli. Unlike general psychological stress, which often involves prolonged exposure to external demands and coping processes, emotional stress is characterized by short-term affective arousal accompanied by measurable autonomic nervous system responses. From a physiological perspective, emotional stress shares substantial overlap with high-arousal emotional states, particularly along the arousal dimension, which explains why datasets originally developed for emotion recognition are frequently utilized in stress-related research. Accordingly, this review focuses on studies in which stress labels are defined through emotionally driven tasks or stimuli and validated via physiological responses, thereby bounding the scope of the survey to emotional stress rather than chronic or purely cognitive stress paradigms.

In recent years, several surveys and review studies have summarized progress in physiological signal-based emotion and stress recognition. Existing reviews have examined stress detection from wearable sensors [7], affective computing with multimodal physiological signals [8], and machine learning or deep learning methods for emotion analysis. These studies provide valuable overviews of signal characteristics, modeling techniques, and application scenarios. However, most prior reviews either focus on a single modality or treat multimodal fusion at a high level, without systematically analyzing fusion strategies across different integration stages or explicitly addressing challenges such as modality heterogeneity, temporal misalignment, and cross-dataset generalization. In addition, limited attention has been paid to the interaction between physiological mechanisms, data characteristics, and algorithmic design choices. These gaps motivate the need for a more structured and mechanism-aware review that connects physiological foundations with multimodal fusion strategies and learning-based classification methods.

Despite significant progress has been made in emotional stress analysis, major gaps remain at both data and algorithmic levels. Public databases are limited by small sample sizes, narrow demographic coverage, inconsistent synchronization protocols, and coarse annotation standards [9,10]. Existing classification models often suffer from overfitting and poor cross-dataset generalization, as many fusion strategies fail to account for modality heterogeneity and temporal misalignment [11,12]. Therefore, a comprehensive review is needed to organize current knowledge, compare methodologies across different stages of multimodal fusion, and clarify emerging challenges and research trends. This work provides a systematic analysis of physiological mechanisms, multimodal databases, preprocessing pipelines, feature extraction techniques, fusion strategies, and classification algorithms, offering insights to guide the development of accurate, interpretable, and generalizable emotional stress recognition systems.

2. Physiological and Psychometric Basis of Emotional Stress

2.1. Physiological Foundations of Multimodal Stress Assessment

In this review, emotional stress is used as an operational concept referring to stress-related physiological responses that are elicited or modulated by emotional stimuli [13]. Unlike general psychological stress, which often involves prolonged exposure to external

demands and coping processes, emotions typically describe short-term affective states characterized by dimensions such as valence and arousal [14]. Emotional stress lies at the intersection of these constructs, where emotionally driven arousal acts as a controlled stressor and produces measurable autonomic responses. From a physiological perspective, emotional stress shares substantial overlap with high-arousal emotional states, which explains why datasets originally designed for emotion classification are frequently adopted in stress-related research [15].

Emotional stress activates the autonomic nervous system, leading to measurable changes in cardiovascular activity, muscle tension, and sweat gland secretion. These physiological pathways can be captured through ECG, EMG, and EDA, which represent complementary modalities reflecting sympathetic and parasympathetic responses. Unlike subjective questionnaires, these signals provide continuous, real-time, and objective indicators of stress-related changes, forming the physiological foundation for multimodal emotional stress detection. Understanding the mechanisms and characteristics of these signals is essential for selecting appropriate features and designing effective classification models for stress recognition.

ECG is a fundamental physiological modality for emotional stress detection due to its sensitivity to autonomic nervous system regulation [16]. A standard ECG waveform is composed of the P wave, QRS complex, and T wave, collectively referred to as the PQRST morphology, which reflects successive phases of cardiac electrical activity. Specifically, the P wave corresponds to atrial depolarization, the QRS complex represents rapid ventricular depolarization, and the T wave reflects ventricular repolarization.

Under stress, increased sympathetic activation and reduced parasympathetic modulation can alter this morphology, most notably through changes in QRS duration and QT interval. Prolonged or shortened QRS duration indicates modified ventricular conduction dynamics, while QT interval variation reflects stress-induced changes in ventricular repolarization timing [17,18]. These alterations are associated with autonomic imbalance and have been consistently observed under both acute and chronic stress conditions [19]. Although indices such as the LF/HF ratio remain debated due to sensitivity to respiration and methodological variance [20], they are still widely used as non-invasive indicators of stress-related autonomic shifts [21]. The characteristic structure of the PQRST complex and its stress-related variations are illustrated in Figure 1.

EMG records the electrical activity of skeletal muscles and reflects neuromuscular responses regulated by both voluntary control and autonomic modulation during emotional stress [22,23]. At the physiological level, EMG signals arise from the superposition of motor unit action potentials (MUAPs), where each motor unit consists of a motor neuron and the muscle fibers it innervates. Stress-induced sympathetic activation increases muscle tension and alters motor unit firing patterns, leading to elevated EMG amplitude and shifts in spectral content, particularly in facial and upper limb muscles [24].

Typical EMG waveforms may exhibit biphasic or triphasic MUAP shapes, which reflect the spatial and temporal propagation of action potentials along muscle fibers and are commonly used to characterize neuromuscular activity under both normal and stress-related conditions. These waveform characteristics are illustrated in Figure 2, demonstrating EMG's capability to capture subtle stress-related muscle activation. However, EMG signals are susceptible to motion artifacts, electrode placement variability, and inter-individual differences in muscle physiology, which may reduce reliability when EMG is used as a single modality [25]. Consequently, EMG is most effective when integrated with complementary physiological signals such as ECG and EDA in multimodal stress detection systems.

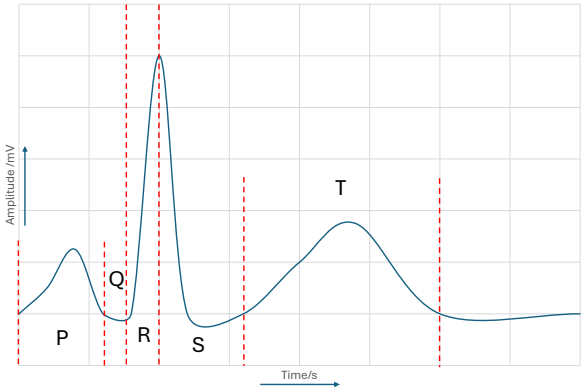

Figure 1. Schematic illustration of a standard ECG waveform showing the P wave, QRS complex, and T wave (PQRST morphology). The P wave represents atrial depolarization, the QRS complex corresponds to ventricular depolarization, and the T wave reflects ventricular repolarization. QRS duration and QT interval are commonly analyzed features that are sensitive to stress-induced autonomic modulation. The red dashed lines indicate the temporal boundaries used to delineate the characteristic waveform components and interval measurements.

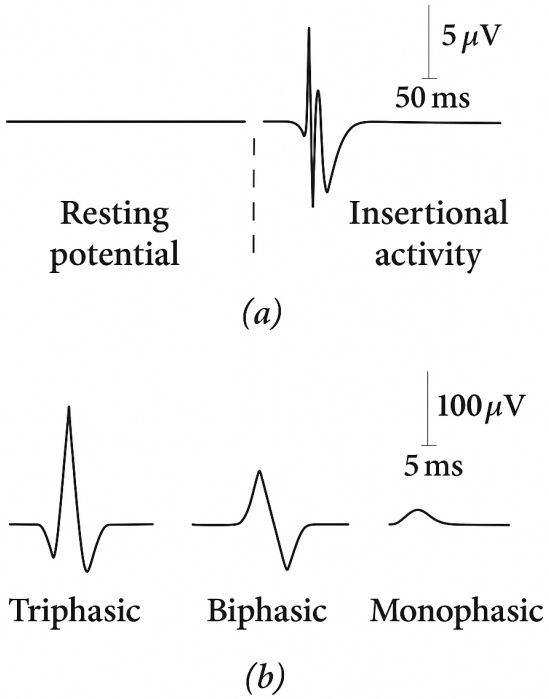

Figure 2. Typical intramuscular EMG waveforms: (**a**) resting potential and insertional activity observed during needle electrode insertion; (**b**) examples of triphasic, biphasic, and monophasic motor unit action potentials (MUAPs), which reflect different patterns of muscle fiber activation and are commonly analyzed to characterize neuromuscular responses under stress and pathological conditions.

EDA reflects sweat gland activity regulated by the sympathetic nervous system and is highly sensitive to emotional and cognitive stress [26,27]. When individuals experience psychological arousal, increased sympathetic activation leads to elevated skin conductance, characterized by tonic skin conductance level (SCL) and phasic skin conductance responses (SCR) [28]. These dynamic features, including rise time, amplitude, and recovery rate, are illustrated in Figure 3 and are widely used as indicators of short-term stress reactivity [29]. However, EDA is easily influenced by environmental factors and individual variations in sweat gland activity, which may reduce its reliability when used alone [30]. Consequently,

EDA is often integrated with ECG and EMG to leverage its high sympathetic specificity and improve the robustness of multimodal stress recognition systems [31].

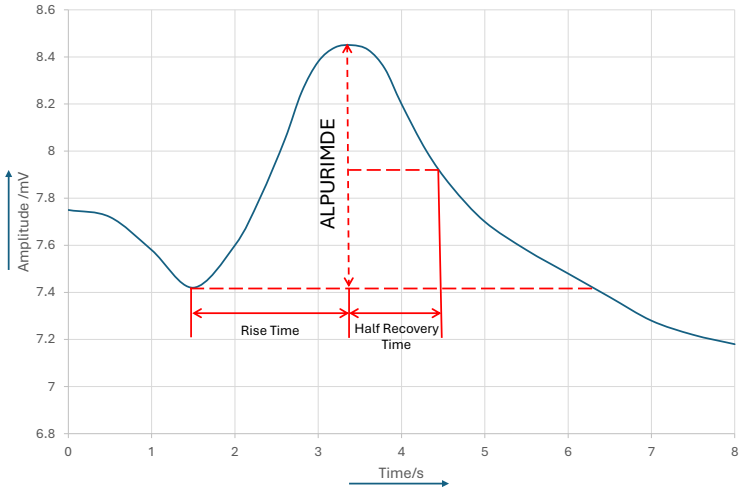

Figure 3. Example of a Skin Conductance Response (SCR) curve. The graph illustrates three typical features of phasic EDA: rise time, amplitude, and half recovery time, which are commonly used to evaluate transient sympathetic nervous activity.

In emotional stress research, single-modality physiological signals rarely capture the complex relation between psychological states and bodily responses [32]. Among non-invasive indicators, ECG, EMG, and EDA, respectively, index cardiovascular regulation, muscle activation, and sympathetic arousal. ECG reflects autonomic balance via heart-rate variability and waveform morphology; EMG quantifies facial and limb muscle tension; EDA provides rapid phasic responses to short-term arousal. Each modality, however, has characteristic limitations, including motion artifacts, electrode-placement sensitivity, and environmental dependence, which together constrain standalone reliability [33,34]. A side-by-side appraisal (Table 1) clarifies comparative strengths and weaknesses and motivates modality-complementary designs. Accordingly, this synthesis provides the rationale for multimodal feature fusion and classification frameworks [35].

While ECG, EMG, and EDA provide the physiological basis for stress recognition, signal-only models are constrained by noise, inter-individual variability, and construct validity. Incorporating psychometric scales as labels and auxiliary features anchors physiological patterns to perceived stress, improving both accuracy and interpretability. The next subsection introduces commonly used questionnaires and how they complement physiological modalities in multimodal assessment.

Although EEG is included in Table 1 for completeness, it is not treated as a primary modality in this review. The purpose of including EEG is to provide a comparative reference that highlights fundamental trade-offs between central and peripheral physiological signals in emotional stress research. As summarized in Table 1, EEG offers high temporal resolution and direct access to central nervous system activity, making it valuable in controlled laboratory settings. However, its practical deployment in stress monitoring is constrained by susceptibility to noise, complex preprocessing requirements, and limited portability, which restrict its applicability in wearable and real-world scenarios.

Table 1. Comparative summary of ECG, EEG, EMG, and EDA in emotional stress detection.

Dimension	ECG	EEG	EMG	EDA
Key Feature Indicators	HRV (SDNN, LF/HF), QTc, PQRST morphology	Band power (delta, theta, alpha, beta), frontal asymmetry, coherence, entropy	RMS, mean frequency, facial/limb MUAP waveforms	SCL (baseline), SCR (amplitude, rise time, recovery time)
Stress-Related Changes	HRV reduction, LF/HF increase, QTc fluctuation, PQRST waveform changes	Increased frontal theta and beta activity, reduced alpha power, altered connectivity	Muscle tension increase, higher amplitude and frequency, irregular contractions	SCR frequency and amplitude increase, SCL elevation
Advantages	Clinically mature, well established medical basis	High temporal resolution, direct reflection of central nervous system activity	Directly reflects muscle tension and facial expression changes	Highly sensitive to sympathetic activity, effective for short-term stress detection
Limitations	Strongly affected by inter-individual variability and motion artifacts [33], limited generalizability [36]	Sensitive to noise and artifacts, requires careful preprocessing and calibration, limited portability	Sensitive to electrode placement, prone to noise and artifacts [34], non-specific muscle tension changes	Strongly influenced by temperature [37] and humidity and skin properties [38], nonlinear and time-varying [39]
Applicable Scenarios	Clinical stress monitoring, cardiovascular stress studies, wearable devices for health	Laboratory-based stress experiments, cognitive workload assessment, affective computing	Facial EMG in emotion recognition, workplace stress detection, muscle fatigue studies	Cognitive load tasks, affective computing, psychophysiological stress research

By contrast, ECG, EMG, and EDA provide more feasible and scalable solutions for continuous stress assessment, particularly under ambulatory or long-term monitoring conditions. Therefore, EEG is discussed here primarily to contextualize modality selection and to justify the scope of this review, rather than to advocate its inclusion in ECG-centered multimodal frameworks. This synthesis clarifies that the focus on ECG, complemented by EMG and EDA, reflects a deliberate balance between physiological relevance and deployment feasibility.

2.2. Common Stress-Related Questionnaires in Stress Assessment

Psychometric assessment provides standardized and quantifiable tools for evaluating psychological stress, offering greater objectivity than clinical interviews [40]. These questionnaires allow for systematic evaluation of mental states and serve as valuable complements to physiological monitoring. They are also used to evaluate the effectiveness of psychological interventions, providing evidence-based guidance for mental health management. Commonly employed instruments are summarized in Table 2, which highlights their assessment dimensions, scoring systems, and application scenarios.

Table 2. Summary of psychological questionnaires for stress assessment.

Type	Measurement Dimensions	No. of Items	Scoring	Evaluation Objective	Application Scenarios	Advantages
PSS [41]	Perceived stress	10/14 (PSS-10/PSS-14)	5-point Likert (0–4)	Assess subjective stress perception	Mental health research; stress management	Easy to implement; suitable for general population
STAI [42]	State and Trait Anxiety (SA and TA)	20 each (40 total)	4-point Likert (1–4)	Differentiate SA and TA	Clinical anxiety assessment; screening	Separately assesses SA and TA
PANAS [43]	Positive or Negative affect	10 each (20 total)	5-point Likert (1–5)	Assess affective states	Emotion research; therapy evaluation	Distinguishes affect valence
SRQ [44]	Physiological, psychological, and behavioral responses to stress	24	5-point Likert (1–5)	Assess comprehensive stress responses	Stress research; psychophysiological studies	Suited for multimodal physiological signal studies

Questionnaires such as the PSS and SRQ provide structured frameworks for quantifying perceived and physiological stress responses [45]. However, their reliance on self-reports introduces potential biases related to recall accuracy, emotion regulation, and situational context. As a result, their diagnostic precision benefits from integration with objective physiological indicators such as ECG, EMG, and EDA, which capture continuous autonomic responses.

Integrating questionnaire-derived scores with physiological signals enhances interpretability and generalization of stress detection models [46]. When employed as both labels and auxiliary features, psychometric data bridge subjective perception and physiological measurement, enabling multimodal frameworks to more accurately capture emotional stress. This integration establishes a conceptual foundation for developing comprehensive, data-driven approaches to emotional stress assessment.

3. Multimodal Stress Databases

3.1. Public and Self-Collected Stress Databases

Emotional stress databases play a crucial role in stress research by providing standardized multimodal data for developing and validating recognition algorithms [47]. Currently, multimodal signals are primarily obtained from two types of databases: public and self-collected datasets [48]. Representative public databases are summarized in Table 3, which include WESAD, AMIGOS, MAHNOB-HCI, DREAMER, and PhysioNet. These datasets provide opportunities to explore physiological and behavioral responses under controlled emotional or stress-inducing conditions [49].

Table 3. Comparison of widely used multimodal stress databases.

Database	WESAD [50]	AMIGOS [51]	MAHNOB-HCI [52]	DREAMER [53]	PhysioNet (Stress Rec.) [54]
Physiological Modalities	ECG, EDA, Respiration, Motion	ECG, EDA, EEG, Facial expressions	ECG, EEG, EDA, Respiration	ECG, EEG	ECG, EDA, Respiration
No. of Subjects	15	40	27	23	40
Stress Induction Method	Wearable sensor-based protocol	Video-induced emotions	Video-induced emotions	Video-induced emotions	Driving simulation
Primary Application Domain	Wearable stress or affect detection	Affective computing; multimodal analysis	Human–computer interaction	VR-based emotion recognition	Driver stress evaluation
Annotation Type	Task phases; self-reports	Self-reports (arousal/ valence)	Self-reports; event markers	Self-reports (arousal/ valence)	Driving states; coarse labels
Recording Duration/ Protocol	Baseline task recovery blocks (minutes)	Multiple short video clips (minutes/ clip)	Video clips with synchronization	Video sessions (multi-trial)	Long continuous driving sessions
Acquisition Device/ Environment	Wearable sensors (lab/semi-free)	Lab; biosensors + video	Lab; medical grade devices	Lab; biosensors	Driving simulator; semi-realistic
Synchronization Alignment	Provided; multi-sensor timestamps	Requires alignment across modalities	Provided across channels	Provided for physiological signals	Provided; per-segment labels
Key Limitations	Limited sample size	Modality alignment required	Lack of high-stress tasks	Restricted modality coverage	High inter-individual variability

As shown in Table 3, most public multimodal databases focus on core physiological signals such as ECG, EDA, and EEG, while few include complementary modalities like respiration, motion, or facial features [55,56]. Video-induced paradigms dominate experimental designs due to their convenience and reproducibility, though driving simulation and wearable scenarios are also emerging [57]. Annotations typically rely on subjective self-reports (e.g., valence/arousal) or coarse task phases, providing basic temporal structure but limited event-level precision.

Despite their wide use, several common challenges remain. First, modality imbalance continues to be a key limitation because behavioral, EMG, and environmental signals are insufficiently represented, thereby restricting the thorough assessment of robust fusion techniques [58]. Second, ecological validity is constrained, as most stressors involve short, video-based tasks with moderate intensity, which restrict generalization to real-world conditions. Thirdly, annotation inconsistency and lack of unified synchronization protocols make cross-database comparisons difficult. Furthermore, many datasets include small, homogeneous cohorts, reducing their representativeness and fairness in modeling population diversity.

To mitigate data scarcity and annotation imbalance in public stress databases, data augmentation and oversampling strategies have been increasingly explored. Conventional approaches include random oversampling, class-balanced reweighting, and noise-based perturbations such as time warping, amplitude scaling [59], and frequency domain augmentation [60], which aim to improve class balance without altering label distributions. However, these techniques often fail to capture the complex temporal and cross-modal

dependencies inherent in physiological stress responses, particularly under extreme low-sample regimes.

More recently, generative modeling has emerged as a promising direction for synthesizing realistic physiological data. Generative adversarial networks (GANs) and variational autoencoders (VAEs) have been applied to generate synthetic ECG, EDA, and multimodal samples that preserve marginal distributions and temporal characteristics [61]. In scenarios of extreme data scarcity, Generative Adversarial Network Synthesis and Markov Random Fields for Oversampling (GANSO) has been proposed to jointly model local temporal structure and global statistical dependencies, enabling more stable oversampling when training data are very limited [62]. Despite their potential, generative augmentation methods require careful validation to avoid distributional drift, mode collapse, or amplification of annotation noise, and synthesized data should therefore be treated as complementary rather than replacement resources for real recordings [63].

Consequently, public stress databases are best positioned as benchmarks and baselines for algorithm evaluation rather than as comprehensive training resources. Their standardized formats facilitate reproducibility and algorithmic comparison, yet they cannot fully capture the complexity of naturalistic stress responses [64]. Researchers are increasingly developing their own datasets that broaden the range of signal modalities, incorporate ecologically valid paradigms, and enhance annotation granularity [47]. This complementary strategy, which leverages public datasets for benchmarking and self-collected datasets for contextual extension, provides a balanced foundation for advancing multimodal stress recognition research and enhancing model robustness across diverse environments [48].

3.2. Experimental Paradigms and Methodological Insights

Existing public stress databases suffer from limited modality coverage, small sample sizes, and inconsistent annotations, restricting their ability to model real-world stress dynamics [65]. To address these shortcomings, researchers increasingly construct self-collected databases with customized task designs and flexible labeling schemes [66]. Such databases enable higher ecological validity by allowing diverse experimental settings and multimodal signal acquisition [67]. They have thus become a vital data source for advancing multimodal emotional stress recognition research.

Self-collected databases often employ carefully designed paradigms to elicit measurable stress responses across cognitive, emotional, social, and ecological dimensions [68]. Cognitive tasks such as the Stroop or MIST tests emphasize standardized control but induce limited stress intensity. Emotional induction via videos or music provides reproducibility yet suffers from individual variability [69]. Social paradigms like the Trier Social Stress Test (TSST) and immersive VR simulations offer strong ecological realism but involve ethical and technical challenges [70]. Table 4 summarizes their mechanisms, measured signals, and trade-offs.

Table 4. Comparison of typical experimental tasks in self-collected stress databases.

Task Type	Specific Task	Stress Induction Mechanism	Main Measured Signals	Advantages	Limitations
Cognitive	Stroop [71]	Color–word conflict and cognitive load	ECG, EDA, EEG	Simple and highly standardized	Limited stress intensity
	MIST [72]	Time pressure and negative feedback	ECG, EDA, EEG	Strong induction and widely used	Artificial setting and low ecological validity
Emotional Induction	Video stimuli	Emotionally evocative film clips	ECG, EDA, EEG, facial expression	Easy to implement and well controlled	Large interindividual variation
	Audio stimuli	Music and affective sounds	ECG, EDA	Simple and noninvasive	Narrow stimulus range
	IAPS image set [73]	Standardized affective pictures	ECG, EDA, EEG	High standardization and replicable	Monotony and weak persistence

Table 4. *Cont.*

Task Type	Specific Task	Stress Induction Mechanism	Main Measured Signals	Advantages	Limitations
Social Stress	TSST [74]	Public speaking and interview causing social evaluation pressure	ECG, EDA, cortisol, EMG	High ecological validity and strong stress	Hard to standardize and ethical concerns
Real-World Simulation	Driving simulation [75]	Traffic complexity and performance pressure	ECG, EDA, respiration, motion	High ecological validity	High cost and strong individual variability
	Immersive VR tasks [76]	Virtual environments that elicit stress	ECG, EDA, EEG, motion	High immersion and realism	Technology dependence and expense

Compared with public datasets such as WESAD or AMIGOS, self-collected databases offer greater flexibility in modality selection, annotation granularity, and task customization [77]. While public databases ensure accessibility and benchmarking consistency, they often rely on low-stress or single-task conditions that limit generalizability [50]. In contrast, self-collected datasets enable richer multimodal acquisition under controlled yet realistic scenarios [78]. However, the cost of equipment, time, and standardization remains a barrier to large-scale replication and open sharing.

Recent self-collected databases exhibit diverse paradigms and signal modalities, integrating physiological, behavioral, and contextual data [79]. Representative examples include Stress-ID and ADA Base, which provide comprehensive multimodal coverage, and WEMAC, which introduces gender-focused and VR-based designs [80]. Others, such as EmpathicSchool and Multi-PENG, emphasize naturalistic tasks and fine-grained psychological labeling [81]. Despite these advances, key challenges remain unresolved, such as small sample sizes, subjective labeling, and inconsistent data quality across modalities [82]. Table 5 summarize their main features and limitations.

Table 5. Summary of self-collected stress databases.

Name	Sample Size	Task Type	Specific Tasks	Modalities	Annotation Scheme	Advantages	Limitations
Stress-ID [79]	65 participants (18F and 47M, age 21–55)	Cognitive, Emotional, Social, Relaxation Control	Breathing baseline; emotional videos; seven interactive stress tasks; public speaking; relaxation	ECG, EDA, Respiration, Video, Audio	Self-assessments (0 to 10 stress); SAM; binary and three-class labels	Comprehensive multimodal coverage; large dataset; detailed annotations; baseline models provided and publicly available	Controlled lab setup; possible sensor-induced stress; subjective bias; gender imbalance; partial data loss
Muse [83]	28 college students (during and after final exams)	Emotional induction and monologue elicitation	Baseline + monologues + emotional videos; stress level assessed via PSS scale	Video, Thermal camera, Audio, Heart rate, Skin conductance and temperature	Self-reports (PSS, SAM); external annotations via AMT	Naturalistic stress context; multimodal coverage; synchronized baseline and task data	Small sample and student-only cohort; contextual rather than controlled stress elicitation

Table 5. *Cont.*

Name	Sample Size	Task Type	Specific Tasks	Modalities	Annotation Scheme	Advantages	Limitations
Empathic School [84]	20 students (aged 21 to 35)	Cognitive, Emotional, Social, Relaxation	Magazine reading, presentation prep, IQ and Stroop, music, funny video, breathing, rest sessions	Facial video, EDA, HR, BVP, IBI, Skin temperature, ACC	NASA-TLX scores; video-based expression labels	Comprehensive wearable and facial coverage; multiple task types; public code	Lab-only environment; coarse temporal resolution; no chronic stress analysis
VERRD [85]	34 participants (final 26 used)	Real-world simulation	360° VR video environment (12 clips)	Eye tracking, ECG, GSR, Self-reports	SAM and VAS scales with circumplex valence–arousal model	High immersion and ecological validity; multimodal integration; publicly available features	Small sample; no EEG data; raw video stimuli unavailable
ADA Base [86]	51 participants	Cognitive + Driving Simulation	n-back (1–3 back, single and dual) and semi-autonomous driving tasks with secondary infotainment load	ECG, EDA, EMG, PPG, Respiration, Skin Temp, Eye tracking, Facial video, Cortisol	Baseline and load levels (low, medium, high); subjective questionnaires (NASA-TLX, PSS, PANAS)	Broad multimodal coverage; realistic simulation; continuous load labels with synchronization	Missing EEG; limited sample; privacy restrictions in video data; artifact-prone sessions
WEMAC [80]	100 women (20–77 years)	Emotional induction (VR fear vs neutral)	Immersive VR videos eliciting fear and neutral states	BVP, GSR, Skin Temp, Resp, EMG, Motion, Speech features	Discrete (12 emotions) and dimensional (VAD) ratings; speech annotations	High ecological validity; large female cohort; VR-based emotion elicitation	Order effects; limited stimuli range; gender-specific sample
Multi-PENG [81]	39 participants (30M and 9F, mean age 24.3)	Video game tasks with graded difficulty	Sports and fighting games (FIFA'23, Street Fighter V) with round-level surveys and pauses	EEG, Eye tracking, Heart rate, Controller inputs, Facial video, Gameplay footage, Surveys	Self-reports; third-party annotations subset	Rich multimodal data; precise temporal alignment; public availability on Kaggle	Limited games; motion artifacts; partial missing modalities; small sample
ForDigit-Stress [82]	40 participants (57.5% F, 40% M, 2.5% diverse; mean age 22.7 ± 3.2)	Social stress task (digital mock interview)	Remote job interviews with 14 stages (self-intro, motivation, logic and math questions, etc.)	PPG, EDA, Cortisol, Facial AUs, Eye tracking, Body skeleton, Speech, HD video, Audio	Frame-by-frame annotations (2 psychologists + self-reports + cortisol validation)	Ecological interview scenario; comprehensive modalities; continuous labels and baseline features	Limited sample; missing eye tracking; EDA delay vs labels; single scenario context

Future database development should balance standardization with ecological realism, combining traditional paradigms with immersive VR or AR environments [87]. Expanding demographic diversity through multi-institutional collaboration will improve generalizability and fairness. Annotation frameworks should integrate self-reports with behavioral and physiological cues for more objective, multi-layered labeling [88]. Technical improvements in synchronization, noise control, and long-term stability will enhance data reliability [89]. Ultimately, standardized acquisition protocols and open data sharing will enable reproducibility and advance multimodal stress research [90].

4. Preprocessing and Feature Extraction

This section summarizes the theoretical foundations and methodological considerations of preprocessing and feature extraction, thereby establishing a unified basis for reproducible, comparable, and physiologically meaningful analyses in emotional stress research. Because ECG, EMG, and EDA differ markedly in sampling rates, dynamic ranges, latencies, and noise characteristics, insufficient preprocessing can bias core indices, inflate model variance, and ultimately undermine the robustness and generalizability of multimodal fusion and downstream classification. Accordingly, we first outline a modality-aware preprocessing workflow and parameter-selection principles for multimodal settings, with emphasis on reporting standards and the avoidance of data leakage. We then summarize representative features across the time domain, frequency domain, and time–frequency domain, as well as nonlinear descriptors for each modality, and further extend the discussion to cross-modal coupling measures, feature selection, and dimensionality reduction, highlighting the physiological interpretability and task relevance of the resulting feature representations.

4.1. Preprocessing Overview

Preprocessing plays a foundational role in multimodal emotional stress analysis, ensuring that ECG, EMG, and EDA signals collected from heterogeneous devices are temporally aligned, denoised, and comparable before feature extraction. Due to differences in sampling rates, dynamic ranges, and noise characteristics, unprocessed signals can distort physiological indices and reduce model robustness. Recent studies emphasize that standardized pipelines integrating synchronization, denoising, and normalization substantially improve data reliability and model generalization, particularly in cross-device and real-world environments [91]. Proper preprocessing thus bridges the gap between raw physiological recordings and reproducible, interpretable multimodal analytics. Figure 4 illustrates the complete signal-processing workflow adopted in multimodal emotional stress analysis, highlighting the sequential steps from data acquisition to classification.

In addition to the standard preprocessing steps illustrated in Figure 4, data augmentation techniques are frequently incorporated in the literature to mitigate limited sample sizes and class imbalance in stress datasets. Common augmentation strategies for physiological signals include time-domain perturbations such as window cropping, time shifting, amplitude scaling, and additive noise injection, which preserve label semantics while increasing data diversity. Frequency-domain augmentation, including spectral masking or random frequency scaling, has also been applied to enhance robustness against device-specific noise and recording variability. For multimodal settings, modality-consistent augmentation and random modality dropout are sometimes adopted to improve fusion robustness under partial signal loss. These augmentation techniques are typically applied after segmentation and normalization, and prior to model training, serving to enhance generalization without altering the underlying physiological interpretation [59,92].

The typical workflow includes four steps: time synchronization and resampling, which correct clock drift and align modalities to a unified time base, often 250–500 Hz for ECG, 100–200 Hz for EMG envelopes, and 10–32 Hz for EDA [93]; denoising and filtering, using zero-phase bandpass and notch filters with adaptive refinement to suppress motion and power-line interference [94]; segmentation and label alignment, partitioning recordings into fixed or event-locked windows while compensating physiological latency and avoiding cross-trial leakage [95]; and normalization and quality control, applying z-scoring or baseline correction to reduce individual variability without eroding physiological meaning [96].

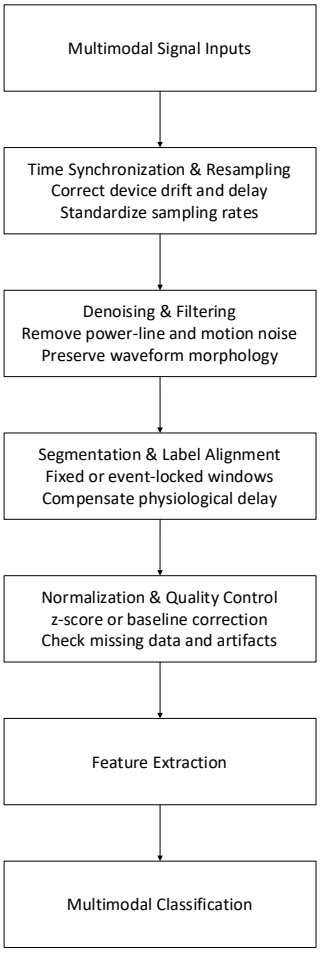

Figure 4. Overview of preprocessing and feature extraction workflow for multimodal emotional stress analysis.

Effective preprocessing follows several key principles. First, physiological interpretability must be preserved; filters should suppress structured noise without distorting waveform morphology or event timing. Second, statistical transparency and reproducibility are critical: studies should report sampling rates, filter bands, missing data handling, and synchronization error statistics to enable replication [97]. Third, intersubject and interdevice harmonization mitigates nonphysiological variance using calibration or batch-correction methods such as ComBat and mixed-effects residualization [98]. Finally, data augmentation can expand training variability through controlled perturbations, including baseline wander, power-line residue, and minor time shifts, provided that physiological plausibility and label alignment are maintained [99].

In summary, preprocessing in emotional stress studies aims to create a reproducible and physiologically coherent foundation for multimodal fusion. By combining time alignment, adaptive denoising, normalization, and careful documentation, researchers can minimize bias from heterogeneous acquisition and improve cross-study comparability. Despite advances in adaptive filtering and automated pipelines, several challenges still exist in managing real-time drift, uncertain latencies, and device variability. Future work should prioritize adaptive and self-supervised preprocessing frameworks that integrate quality indices and uncertainty estimation to deliver more robust multimodal emotional stress analysis [100].

4.2. Feature Extraction for Emotional-Stress Classification

Feature extraction is a critical step that bridges raw physiological recordings and interpretable stress-related patterns. Its goal is to extract compact and meaningful descriptors that reflect autonomic and muscular responses under emotional stress. Recent studies emphasize that time, frequency, time–frequency, and nonlinear domains capture complementary physiological dynamics, each revealing distinct aspects of autonomic modulation and psychophysiological arousal [10]. Properly engineered features not only enhance model accuracy but also improve generalization across subjects and contexts. With the proliferation of wearable and multimodal sensors, robust feature representations play an increasingly central role in achieving high-precision emotional stress detection [101].

ECG, EMG, and EDA remain the principal physiological modalities for stress recognition, each reflecting a different facet of autonomic or somatic activation. ECG features focus on heart rate variability (HRV), encompassing time-domain indices such as SDNN and RMSSD, spectral metrics like LF and HF power, and nonlinear measures including entropy and Poincaré geometry [102,103]. Recent works combine multiple feature scales through graph neural networks or wavelet scattering to better capture morphological and frequency correlations [104,105]. EMG features quantify muscle activation using root mean square, waveform length, and median frequency; hybrid time–frequency and attention-based representations have improved robustness under motion artifacts and intersubject variability [106,107]. EDA features describe sympathetic arousal through skin conductance level (SCL) and response (SCR) characteristics, with energy- and entropy-based descriptors providing additional discrimination. Transformer and CNN–LSTM frameworks have recently automated EDA feature learning and artifact detection for more reliable classification [108].

Subjective scales such as the Perceived Stress Scale (PSS) and Stress Response Questionnaire (SRQ) provide complementary self-reported dimensions of stress that can be quantitatively integrated with physiological features. Correlation and regression analyses reveal consistent relationships between questionnaire scores and HRV, SCR, or EMG-derived activity, confirming their physiological validity [109]. In multimodal learning, feature-level fusion directly concatenates questionnaire and physiological descriptors for unified modeling [110], while decision-level fusion aggregates independent model outputs through weighted voting or probabilistic rules to enhance robustness under heterogeneous noise [111]. Hybrid or cross-modal attention approaches further align temporal structures between discrete questionnaire scores and continuous biosignals, improving interpretability and robustness to subjective bias.

Overall, physiological and subjective features jointly underpin multimodal stress recognition. Carefully selected descriptors across ECG, EMG, and EDA capture distinct but complementary responses of the autonomic nervous system and muscle activation, while their fusion with psychological measures enhances contextual understanding. Future research should emphasize unified feature embeddings and adaptive multimodal fusion mechanisms that integrate cross-modal correlations and uncertainty awareness to achieve more interpretable and generalizable emotional stress classification.

5. Multimodal Feature Fusion and Cross-Modal Representation

Multimodal fusion and cross-modal representation are essential because no single modality can capture the asynchronous and complementary responses of ECG, EMG, and EDA. We compare fusion paradigms and introduce cross-modal modeling that aligns and shares representations to cope with temporal asynchrony, physiological lags, domain shift, noise, and missing data. The section then addresses robustness and external

generalization in real-world use, using rigorous splits, external validation, ablation, and significance testing to balance physiological interpretability with reproducible modeling.

5.1. Multimodal Feature Fusion Strategies

From a methodological perspective, multimodal physiological fusion is motivated by the fact that emotional stress manifests through multiple, partially independent physiological pathways, including cardiovascular, electrodermal, and neuromuscular responses. No single modality can fully capture these dynamics due to differences in temporal resolution, noise characteristics, and physiological sensitivity. Prior studies have shown that fusing complementary modalities can improve robustness and discriminability by reducing uncertainty and compensating for modality-specific limitations, especially under real-world conditions with noise or partial signal loss [112]. However, fusion is not universally beneficial, as performance gains depend on modality complementarity, data quality, synchronization accuracy, and sample size, and improper fusion may introduce redundancy or exacerbate overfitting [113]. These observations motivate the need for principled fusion strategies and provide the rationale for the systematic comparison of feature-level, hidden-layer, and decision-level fusion paradigms in this section.

Figure 5 and Table 6 provide a structured overview of the three mainstream multimodal fusion strategies commonly used in emotional stress recognition, explicitly illustrating their fusion stages, interaction mechanisms, and practical trade-offs.

Multimodal feature fusion integrates complementary sources such as ECG, EMG, EDA, and subjective questionnaire data to enhance emotional stress classification. Differences in temporal resolution, data type, and physiological linkage across modalities motivate three mainstream strategies: feature-level, hidden-layer, and decision-level fusion. Each presents distinct trade-offs—feature-level fusion requires strict temporal and statistical alignment, hidden-layer fusion strengthens cross-modal dependency learning at higher computational cost, and decision-level fusion offers robustness to missing modalities at the expense of deep inter-modal interactions.

Feature-level fusion concatenates ECG, EMG, and EDA features with questionnaire data into a unified vector for classification. It captures interactions between physiological and psychological responses but is sensitive to redundancy and asynchrony. Techniques such as Z-score normalization, PCA, and Autoencoder-based mapping improve alignment and compactness [114,115]. Advanced variants employ Deep CCA and modality dropout to enhance robustness to missing data [116]. Despite simplicity and interpretability, early fusion may suffer from overfitting and modality imbalance; therefore, reporting should include feature dimensionality, missing-data proportion, and cross-session robustness to ensure reproducibility.

Hidden-layer fusion operates at intermediate network stages, enabling shared representation learning through mechanisms such as cross-modal attention, latent-space projection, and mixture-of-experts gating [117,118]. CNN-, LSTM-, and Transformer-based encoders independently process each modality before interaction, facilitating complementary feature extraction under asynchronous or noisy conditions [119,120]. To address physiological delays and sampling disparities, learnable offset encoding, visibility masks, and modality dropout can be adopted. While more computationally intensive, this strategy enhances semantic correlation and cross-device generalization when sufficient data and tuning are available.

Decision-level fusion independently trains classifiers for each modality and aggregates their outputs through weighted voting, probability averaging, or attention-weighted ensembles [121,122]. It excels in handling partial or missing modalities and maintains high interpretability by preserving unimodal transparency [123]. However, its limited

ability to capture deep inter-modal dependencies constrains accuracy in strongly coupled tasks. Transformer-based decision weighting has recently improved adaptability and interpretability under real-world variability [124].

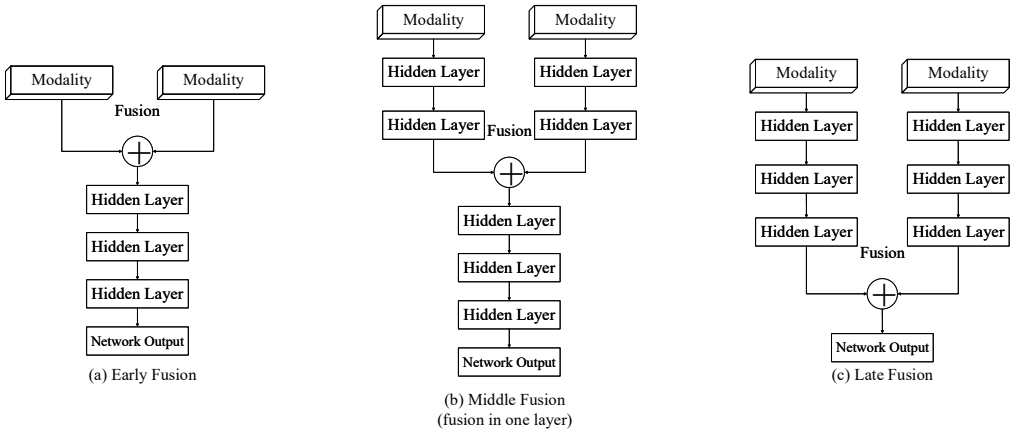

Figure 5. Comparison of representative multimodal fusion strategies for emotional stress recognition, including early (feature-level), mid-level (hidden-layer), and late (decision-level) fusion.

From a theoretical perspective, whether multimodal fusion improves over single-modality inference depends on information complementarity and sample efficiency. Although complementary modalities can reduce uncertainty and lower achievable error bounds, fusion is not guaranteed to be beneficial in practice because it also increases feature dimensionality, noise sources, and estimation burden. Pereira et al. theoretically compared early and late fusion under Gaussian assumptions by analyzing error probabilities with perfect model knowledge and finite training data [125]. Their results indicate that early fusion is optimal with sufficient data, whereas under finite-sample regimes, its performance degrades more severely due to dimensionality expansion, making late fusion preferable in data-limited settings. This theoretical insight explains why empirical gains from fusion in stress recognition vary across datasets: fusion helps when modalities are complementary and well aligned, but may degrade performance when additional modalities are noisy, weakly informative, or poorly synchronized.

Figure 5 and Table 6 jointly illustrate the architectural and algorithmic trade-offs among the three fusion paradigms. Feature-level fusion emphasizes simplicity and interpretability; hidden-layer fusion captures deep semantic complementarity across asynchronous modalities; decision-level fusion prioritizes flexibility and robustness. Future directions include unified fusion architectures integrating early–mid–late interactions, self-supervised and contrastive learning for cross-modal consistency, modality-aware completion and lightweight deployment for wearable devices, and standardized evaluation with interpretable metrics to enhance reproducibility and practical adoption [100].

Table 6. Comparative summary of mainstream multimodal fusion strategies and representative studies for emotional stress recognition.

Strategy	Key Properties	Typical Methods	Strengths	Limitations	When to Use
Early fusion (feature-level)	Fusion at input feature stage; shallow interaction; sensitive to misalignment; low compute; high interpretability.	PCA [126], Z-score [127], CCA/DCCA [128], autoencoder mapping [129].	Simple and deployable; transparent feature attribution; effective when signals are well synchronized.	Feature redundancy and modality imbalance; overfitting risk under high dimensionality; weak under asynchrony or missing signals.	Well-aligned datasets, stable acquisition, and sufficient sample size with mature feature engineering.
Mid fusion (hidden-layer)	Fusion at intermediate layers; deep interaction via shared latent space; moderate missing-modality tolerance; high compute; moderate interpretability.	Cross-modal attention [130], latent projection [131], MoE or gating [118,132], Transformer fusion [133].	Learns explicit cross-modal dependencies; handles heterogeneous feature spaces; improved robustness under noise with proper regularization.	Data-hungry; tuning-sensitive; possible training instability (modality dominance or gradient imbalance); heavier deployment cost.	Asynchronous or heterogeneous signals where learned interactions are required and sufficient training data or resources are available.
Late fusion (decision-level)	Fusion at output stage; minimal interaction; strong robustness to missing modalities; moderate compute; high interpretability per modality.	Weighted voting [134], probability averaging [135], stacking ensembles [136], attention-weighted decision fusion [137].	Fault-tolerant and modular; easy to add or remove modalities; works under partial signal loss; strong reproducibility with unimodal baselines.	Limited deep inter-modal semantics; performance depends on unimodal model quality; may underperform when strong cross-modal coupling exists.	Real-world deployment with missing data, sensor dropouts, or when modularity and reliability are priorities.

5.2. Cross-Modal Representation and Consistency Modeling

In emotional stress recognition, ECG, EDA, and EMG capture complementary but heterogeneous facets of autonomic responses, differing in physiological origin, temporal dynamics, and noise characteristics. Simple concatenation often fails to align, transform, or compensate information across modalities. We therefore organize cross-modal representation learning around three objectives: temporal alignment, distributional alignment, and semantic alignment to obtain modality-invariant yet physiologically meaningful features [138].

Cross-modal temporal alignment aims to mitigate inconsistencies in representation caused by differences in sampling frequency and physiological response latency across modalities. In emotional stress scenarios, ECG signals typically respond in near real time, EDA signals exhibit a delay of several seconds due to the slow dynamics of sweat gland activation, and EMG signals are highly sensitive to rapid changes in muscle activity [139]. To address these disparities, learnable temporal shift parameters and delay-gating mechanisms can be incorporated during the encoding stage to establish trainable phase relationships across modalities [140]. This can be complemented with temporal positional encoding or phase-aware embedding to capture rhythmic characteristics at multiple time scales [141]. For scenarios requiring fine-grained segment-level alignment, dynamic time warping or its differentiable variants may be employed to accommodate local temporal distortions [142], while anti-aliasing resampling and causal convolution can be used to avoid artificial correlations and future information leakage induced by interpolation [127]. Collectively, these mechanisms enable the construction of a comparable temporal baseline across modalities without disrupting intrinsic physiological timing patterns.

Statistical alignment targets discrepancies in feature distributions and domain shifts that arise from modality characteristics or device heterogeneity [143]. Its goal is to reduce modality discriminability in the representation space and promote domain invariance. Representative approaches include kernel-based statistical matching methods such as maximum mean discrepancy to constrain higher-order moment convergence [144]; adversarial domain alignment to suppress modality-specific information through a discriminator network [145]; and correlation or covariance alignment techniques to achieve second-order

distribution matching without additional adversarial training [146]. In cases involving batch effects or device-specific biases, recalibrated normalization, whitening transformations, or hierarchical normalization schemes may be combined to attenuate session- and hardware-specific variability [147]. When used in combination with temporal alignment, statistical alignment reinforces the assumption that identical emotional states should exhibit similar distributional structures across modalities, thereby improving external generalization across devices and recording sessions.

Semantic alignment seeks to ensure that different modalities express the same emotional construct within a unified high-level representation space [148]. It operates by guiding the model to cluster cross-modal samples of the same emotional category into a shared semantic region. Typical strategies include supervised or semi-supervised contrastive learning and prototype-based metric learning [149], which enforce intra-class compactness and inter-class separability to enhance discriminability and modality interchangeability [150]; mutual information maximization [151] and deep canonical correlation analysis [152], which strengthen cross-modal correlations and facilitate the formation of a common subspace representation; and curriculum-based weighting or confidence-aware learning in scenarios with noisy or incomplete labels [153]. Compared with statistical alignment alone, semantic alignment is more directly task-driven and is particularly critical for cross-task transfer and cross-subject generalization. It also provides a strong constraint for the development of shared latent space representations, ensuring consistent and interpretable downstream inference across heterogeneous modalities.

Table 7 provides a comparison of temporal, distributional, and semantic alignment strategies with respect to their objectives, underlying mechanisms, and applicability. The analysis indicates that although these strategies offer complementary advantages for addressing modality heterogeneity, they remain constrained by localized alignment effects, susceptibility to noise, and limited capacity to model real-world dynamics. Building on shared latent-space modeling, future research should prioritize: developing scalable latent representations that adapt to time-varying modality characteristics; introducing semantics-preserving and invariance constraints that generalize across devices and scenarios; and strengthening adaptive alignment and robust inference under low signal-to-noise ratios, nonstationarity, and partial-modality conditions. Progress along these directions is critical for enhancing the stability and deployability of multimodal physiological systems in complex real-world settings and provides the theoretical and technical basis for robustness-oriented modeling in such environments.

Table 7. Comparative analysis of cross-modal alignment strategies in multimodal physiological stress recognition.

Aspect	Objective	Representative Algorithms	Theoretical Assumption	Advantages	Limitations	Scenarios
Temporal Alignment	To correct variations in sampling frequency and physiological response latency, enabling temporal correspondence across modalities.	Learnable time-shift modules [154]; temporal position encoding [155]; dynamic time warping (DTW) and differentiable variants [156]; causal temporal convolution [157]; delay-aware attention mechanisms [158].	Modality-specific discrepancies can be modeled through learnable temporal shifts and local temporal scaling.	Directly accounts for physiological latency and asynchronous sampling, improving segment-level alignment and temporal consistency.	Excessive alignment may introduce artificial temporal distortions; global alignment is difficult under causal or real-time constraints.	Suitable when modalities exhibit distinct temporal dynamics (e.g., near-instant ECG response, delayed EDA activation, rapid EMG fluctuations).

Table 7. *Cont.*

Aspect	Objective	Representative Algorithms	Theoretical Assumption	Advantages	Limitations	Scenarios
Distribution Alignment	To reduce inter-modality divergence in feature distributions and mitigate domain shift caused by device, session, or subject variability.	Maximum Mean Discrepancy (MMD) [159]; Domain-Adversarial Neural Networks (DANN) [160]; Correlation Alignment (CORAL) [161]; whitening transformation [162]; layer-wise normalization recalibration [163].	Modalities share alignable statistical moments that can be matched through explicit distributional constraints.	Mitigates domain shift without requiring strong supervision, enhancing robustness to inter-subject or inter-device variability.	Statistical alignment may overlook semantic structure; adversarial optimization can be unstable.	Preferred in scenarios with device heterogeneity, batch effects, or cross-session variability.
Semantic Alignment	To map heterogeneous modalities into a unified semantic space in which representations of the same emotional state converge.	Contrastive learning frameworks (InfoNCE, SupCon) [164]; triplet and prototypical networks [165]; mutual information maximization [166]; Deep Canonical Correlation Analysis (DCCA) [167].	Different modalities approximate a shared semantic manifold and can be aligned through discriminative or correlation-based objectives.	Enhances discriminability and modality-invariant representation, facilitating external generalization and cross-modal transfer.	Requires reliable labels or pseudo-labels; risk of semantic collapse or dominance of a single modality.	Most effective when semantic consistency across modalities is critical for modality-agnostic inference or transfer learning.

5.3. Robustness and Generalization in Real-World Scenarios

Although multimodal emotion and stress recognition has achieved steady progress under controlled laboratory settings, real-world deployment remains limited. Physiological signals in natural conditions often suffer from unstable quality, incomplete modalities, and device or subject variability, making simple fusion strategies insufficient for robust inference and cross-domain generalization [168]. Ensuring robustness and external generalization has therefore become essential for translating experimental success into practical systems.

Real-world motion, posture shifts, and sensor slippage inject nonstationary artifacts that overlap true physiology, e.g., ECG R-peak distortions and EDA spikes or drifts from contact changes [169], complicating denoising and alignment [170]. Task interaction further couples motion and autonomic dynamics, confounding classification boundaries [171]. Low SNR during everyday activities degrades synchrony across modalities [172]; deep models then overfit spurious cues and destabilize under modality dropout or interruptions [173,174]. Beyond noise, strong inter and intra subject variability and context dependence (fatigue, attention) impede a stable decision boundary and harm cross-task generalization [175–177]. Distribution shifts across users [178], devices [179], and stress paradigms further erode transferability.

To evaluate external generalization, a standardized benchmark suite should be established, incorporating leave-one-subject-out validation for person-level transfer and additional cross-device, cross-session, and cross-dataset evaluations to reveal distribution and label shifts [180,181]. Define splits that prevent temporal and trial leakage by disallowing adjacent windows across splits, and document modality availability, including the percentage of windows with dropouts and the rate of alignment failures. Beyond accuracy and AUROC, report class imbalance aware metrics such as balanced accuracy, Matthews correlation coefficient, and area under the precision–recall curve, as well as runtime, latency,

and memory footprint to reflect deployability. Quantify uncertainty and calibration using measures such as predictive entropy, Brier score, and expected calibration error, and pair these with statistical confidence through bootstrap confidence intervals, permutation tests, and effect sizes for key comparisons. Conduct robustness stress tests, including synthetic noise corruption, modality ablation, and inference under missing modalities, and include ablations that remove alignment losses or fusion blocks to attribute gains. Finally, release random seeds, preprocessing and alignment parameters, code versions, and device metadata to ensure full reproducibility.

Prioritize invariance and adaptation at three layers. Signal level: adaptive filtering, blind source separation, multi-sensor artifact decoupling, plus baseline normalization and dynamic recalibration to reduce drift. Representation level: domain invariant learning [182], contrastive or self-supervised pretraining for label-efficient stability [183], and prototype alignment or normalization to tighten class structure across hardware and sessions [184]. Learning level: meta and few-shot personalization for rapid subject adaptation [185], adversarial or noise injection training with uncertainty modeling for robustness [186], and test time adaptation to track input shifts online [187].

In summary, instability, modality disruption, and domain shifts remain the core barriers to real-world stress recognition. Robustness and generalization should thus be treated as core design principles rather than afterthoughts or post-hoc refinements. The next chapter classifies current algorithms by architectural paradigm and generalization capability, providing a structured framework for developing deployable and scalable multimodal systems.

6. Classification Algorithms for Emotional Stress Recognition

After cross-modal representation and fusion, the goal is classification, which involves mapping integrated features to stable and generalizable emotion and stress labels. Current work follows two tracks: traditional pipelines built on handcrafted features and shallow classifiers, and deep models that learn spatiotemporal representations and align modalities end-to-end. This section reviews representative advances and limits of both paradigms, comparing feature dependence, generalization, and deployability, and highlighting practice-oriented improvements for real-world use.

6.1. Traditional Machine Learning-Based Approaches

Traditional machine learning provides the methodological foundation for multimodal emotion and stress recognition, combining manually designed features with supervised classifiers in a standardized analysis framework. As shown in Figure 6, physiological signals such as ECG, EDA, and EMG are preprocessed, transformed into discriminative features, and refined through feature selection or dimensionality reduction to reduce redundancy and improve generalization. The resulting feature set is then used to train conventional classifiers and validated under subject-dependent and subject-independent settings to assess robustness and generalization. This framework, widely adopted in early research, established a reproducible benchmark and revealed key limitations in robustness, fusion efficiency, and transferability, forming the foundation for subsequent deep learning developments.

Traditional machine learning-based approaches rely on handcrafted feature representations that encode domain-specific physiological knowledge prior to classification [188]. For ECG signals, commonly used features are derived from heart rate variability and waveform morphology, including time-domain indices such as SDNN and RMSSD, frequency-domain measures such as LF and HF power, and nonlinear descriptors such as entropy and Poincaré plot statistics [189]. These features are designed to capture autonomic nervous system regulation under stress conditions.

For EMG signals, handcrafted features typically quantify muscle activation intensity and spectral characteristics, including root mean square, waveform length, zero-crossing rate, and median or mean frequency, reflecting stress-related changes in muscle tension and neuromuscular activity [190]. EDA features mainly characterize sympathetic arousal and are often extracted from both tonic and phasic components, such as skin conductance level, skin conductance response amplitude, rise time, and recovery metrics.

Feature selection plays a key role in traditional machine learning frameworks, serving as an intermediate stage between feature extraction and classification to identify informative subsets from high-dimensional multimodal physiological data. Common strategies include statistical filtering, wrapper-based evaluation, and embedded importance estimation during model training. Based on the refined features, studies typically apply classifiers such as support vector machines, k-nearest neighbors, decision trees, random forests, logistic regression, and naive Bayes, while ensemble learning methods have been introduced to enhance robustness. For sequential physiological data, probabilistic models such as hidden Markov models and conditional random fields have been explored to capture temporal dependencies. Overall, this framework is valued for its interpretability and engineering practicality but remains sensitive to feature design, data quality, and parameter selection, which limits generalization under asynchronous or noisy multimodal conditions. Table 8 summarizes representative studies employing traditional machine learning approaches, outlining datasets, signal modalities, classifier configurations, and key limitations to provide a comparative overview of their performance and applicability.

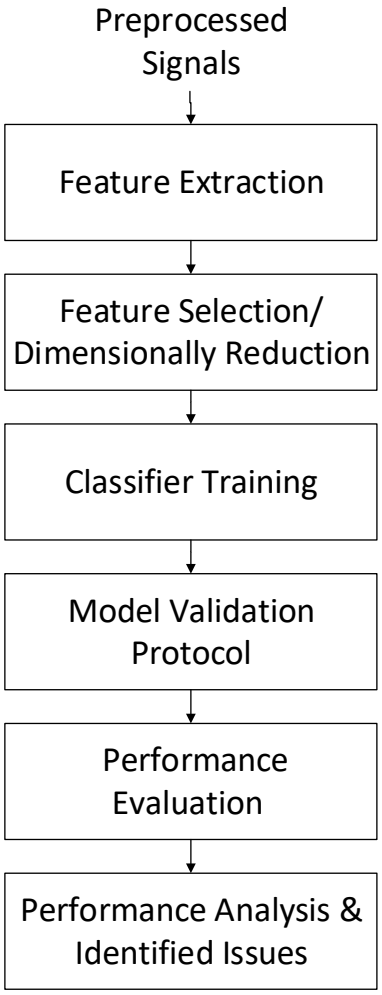

Figure 6. Traditional machine learning-based multimodal stress recognition framework.

Table 8. Traditional multimodal stress classification methods with preprocessing and feature engineering details.

Author	Datasets	Signals	Extracted Features	Feature Selection	Classifier	Accuracy	F1 Score	Limitations
Torres-Valencia et al. [191]	DEAP, MAHNOB-HCI	EEG (main), GSR, HR, Resp., Temp., EMG, EOG	EEG spectral features, HRV indices, statistical descriptors	Not specified	SVM	75.17%	79.25%	Limited to SVM; binary classification only; shallow fusion; no temporal modeling; no real-time validation
Hao et al. [192]	Self-collected	EEG, PPG, EOG	Time–frequency and statistical features	Random Forest Selection (RFS)	1D-CNN + RFS	90.67%	91.47%	Small sample size; all-male cohort; limited diversity of feature selection methods
Gunawan et al. [193]	Self-built	ECG, EMG, EEG	Statistical and time-domain features	–	KNN	73.33%	–	Small sample size; unbalanced data; no F1/recall reported; no model comparison
Patil et al. [194]	BioVid Heat Pain	ECG, EDA, EMG	HRV features, EDA statistical features, EMG amplitude descriptors	Not specified	Logistic Regression	83.20%	–	F1 and recall not reported; sensitive to feature dimensionality; limited nonlinear modeling ability
Dutsinma et al. [195]	Self-built	Heart Rate, Blood Pressure	Heart rate and blood pressure statistics	–	Decision Tree	95%	–	Small sample size; limited modalities; restricted applicability
Abadi et al. [196]	Self-built	Finger pressure, hand motion, facial expression	Pressure and motion statistical descriptors	Late fusion-based aggregation	Semi-Naive Bayesian (Late Fusion)	93%	93%	Facial data loss; limited physiological integration; robustness to sensor failure needs improvement

As shown in Table 8, reported performance varies markedly across datasets and signal combinations, which reflects instability introduced by heterogeneous feature engineering and fusion choices as well as inconsistent data splits and preprocessing [197]. Early fusion that directly merges multimodal features can inflate or depress results through redundancy and misalignment, and many studies omit key metrics such as F1 score and recall, further hindering fair comparison [198]. To improve reliability, future work should adopt standardized protocols that prevent temporal and trial leakage, use a common set of metrics, and fully disclose preprocessing pipelines, feature inventories, fusion settings, and random seeds, with code and data made accessible for replication [92].

Traditional machine learning methods rely on handcrafted descriptors such as heart rate variability and power spectral density that often transfer poorly across datasets and tasks, and most pipelines use short or static windows that miss long term dynamics and delayed physiological responses. Fusion remains sensitive to sampling heterogeneity and timing mismatches, which introduce redundancy and noise that obscure cross-modal relations. Under subject independent evaluation, such as leave one subject out, performance typically declines, revealing weak generalization to inter-individual variability and domain shifts [175,180]. Robustness is further limited by motion artifacts and intermittent modalities in real-world recordings, which degrade stability during deployment [172]. Promising directions include more physiologically grounded feature construction with redundancy control, multiscale temporal modeling, alignment-aware and missing-aware fusion, and rigorous validation under the standardized protocols outlined above.

In summary, improvements to traditional pipelines converge on three directions: constructing physiologically meaningful features with careful selection to curb redundancy and improve transferability, introducing temporal modeling with rigorous evaluation to capture nonstationary and delayed responses, and adopting adaptive fusion to handle modality heterogeneity and missing data. These steps yield incremental gains but do not overcome the core dependence on handcrafted features and shallow architectures, which limits learning of high-level nonlinear relations across modalities. As a result, deep learning has become the leading paradigm by enabling end-to-end feature learning, explicit temporal modeling, and integrated multimodal representation within a single framework.

6.2. Deep Learning-Based Approaches

Deep learning has become the primary direction for multimodal emotion and stress recognition because traditional pipelines based on handcrafted features, limited temporal modeling, and weak adaptability to modality heterogeneity cannot meet real-world requirements. Compared with manual features and shallow classifiers, end-to-end models learn temporal dynamics, spatial structure, and cross-modal interactions within a single trainable system, which improves robustness and generalization under inter-individual variability. Recent progress concentrates on attention and Transformer architectures, adaptive cross-modal alignment, self-supervised pretraining with transfer learning, and compression for edge deployment. As shown in Figure 7, the workflow integrates signal representation, hierarchical feature extraction, multimodal fusion, model training, and evaluation in one pipeline, and shared encoders with cross-modal attention enable the integration of heterogeneous physiological signals.

Deep learning approaches for multimodal emotion and stress recognition are commonly categorized into four major families: convolutional neural networks, recurrent neural networks, hybrid convolutional–recurrent models, and attention-based Transformer architectures. Emerging directions include state space models exemplified by Mamba [199], mixture of experts with dynamic routing [132], generative models such as GAN and VAE [200], self-supervised and contrastive learning, knowledge guided networks including

Kolmogorov Arnold Networks [201], and graph neural networks. Convolutional models specialize in capturing local morphological patterns and are computationally efficient on short temporal windows. Recurrent architectures capture temporal dependencies across sequences. Hybrid convolutional–recurrent models integrate both local and long-range dynamics. Transformer-based architectures learn long-distance relationships and adaptively weight cross-modal interactions, although they typically require larger datasets and greater computational resources. Taken together, these model classes present distinct trade-offs with respect to accuracy, robustness, data requirements, computational cost, and practical deployability.

With respect to optimization and convergence, the reviewed classification models are commonly trained by minimizing an empirical risk objective, yet their optimization properties differ markedly across paradigms. Traditional machine learning pipelines built on handcrafted features often employ convex or quasi-convex objectives, where optimality can be more explicitly characterized and convergence is typically stable under standard solvers. In contrast, deep multimodal models are optimized using stochastic gradient-based methods on highly non-convex objectives. As a result, optimality in deep learning-based stress recognition is generally understood in a practical sense, referring to convergence toward stable solutions with good generalization performance rather than guarantees of global optima [202].

Convergence behavior in multimodal deep learning is further affected by modality heterogeneity. Differences in signal scale, sampling frequency, physiological latency, and data availability across modalities can induce gradient imbalance and modality dominance, leading to unstable training or suboptimal fusion [203]. In practice, convergence is commonly evaluated through smooth loss reduction, bounded generalization gaps on validation sets, robustness across random seeds, and the absence of gradient explosion or collapse. To improve convergence stability, existing studies adopt strategies such as modality-wise normalization, balanced or curriculum sampling, alignment-aware fusion, regularization and early stopping, gradient clipping, and uncertainty- or attention-weighted training to prevent a single modality from overwhelming the fused representation [204]. Transparent reporting of these optimization settings and convergence diagnostics is essential for reproducibility and fair comparison across multimodal stress recognition studies.

Table 9a,b indicate that recent multimodal stress recognition studies include a wide variety of datasets, signal combinations, and architectural paradigms. Overall, deep learning models demonstrate clear advantages in automatically learning hierarchical and cross-modal representations without manual feature engineering. Convolutional neural networks effectively extract local morphological features and are suitable for short-window and edge-oriented analysis, while recurrent and hybrid CNN–RNN architectures capture temporal dependencies in non-stationary physiological signals. Attention-based and Transformer models achieve adaptive weighting and long-range dependency learning, facilitating dynamic cross-modal alignment and feature interaction. State-space models such as Mamba enable efficient long-sequence modeling with reduced memory consumption, and mixture-of-experts or dynamic gating structures enhance robustness under modality imbalance and individual variability. Self-supervised and contrastive learning approaches improve data efficiency and cross-device generalization, whereas graph neural networks and generative models contribute to structural modeling and data augmentation, addressing label scarcity and class imbalance in complex multimodal scenarios.

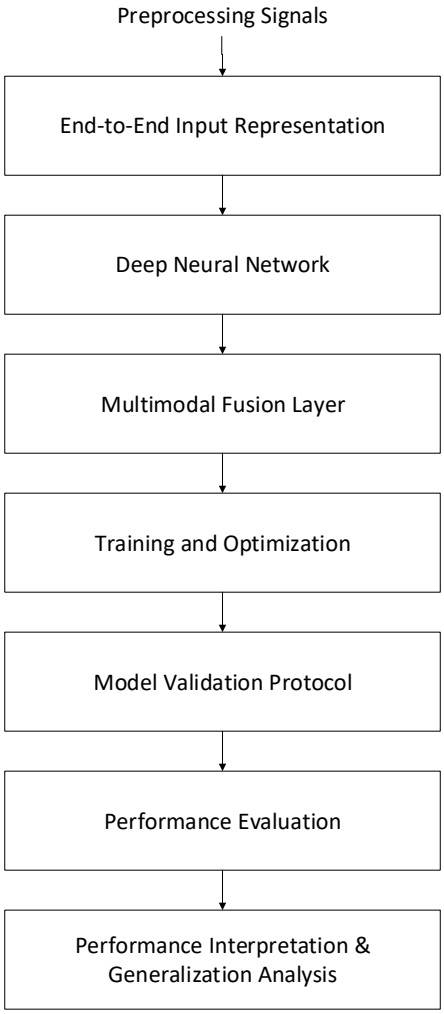

Figure 7. Deep learning-based multimodal stress recognition framework.

Despite these advances, deep learning models still face challenges in data dependence, generalization, and practical deployment. Their performance often relies on large, high-quality datasets and extensive computational resources, which limit reproducibility and accessibility across research environments. Cross-subject and cross-device variability remains a key obstacle to robust generalization, and overfitting to specific experimental settings is still prevalent. Moreover, most frameworks require complete and synchronized modalities, reducing their reliability under signal loss or asynchronous acquisition. Future development should therefore emphasize unified cross-modal representation learning, adaptive model calibration, and efficiency-oriented architectures that balance accuracy with resource constraints. Additionally, incorporating explainability and uncertainty estimation will be critical to improving transparency and trustworthiness in real-world applications.

In summary, deep learning has become the dominant paradigm for multimodal emotion and stress recognition, offering superior feature abstraction and adaptive fusion capabilities compared with traditional methods. However, its benefits are not absolute, as data requirements, computational cost, and interpretability remain open issues. The following section systematically compares traditional and deep learning paradigms in terms of performance, generalization, complexity, and practicality, providing a structured foundation for subsequent model selection and optimization.

Table 9. (a) Comparison of multimodal stress classification algorithms based on deep learning (Part A). (b) Comparison of multimodal stress classification algorithms based on deep learning (Part B, continued from (a)).

(a)

Author	Datasets	Signals	Classifier	Accuracy	F1 Score	Limitations
Bahar et al. [205]	KMED, DEAP	EEG + facial image	AAT + SIFT + LBP + FLF + SVM	KMED: 89.95%, DEAP: 92.44%	–	Signal-to-image processing required; only binary classification is used; sensitive to data synchronization and frame selection
Chouinard et al. [206]	RECOLA, LM-TSST	Facial video, ECG, EDA	Decoupled Mamba Network + Time Relaxation Reconstruction Mechanism + Transformer	Only Concordance correlation coefficient: RECOLA: 0.3921, LM-TSST: 0.3774	–	Sensitive to modal differences; Limited handling of modal imbalance; Generalization depends on training distribution; Lack of multi-scenario validation
Xingchao Wang et al. [207]	DEAP, Self-built, Million Song Dataset	EEG, ECG, facial video	Multimodal LSTM + Emotional Markov Chain	Valence: 86.75%, Arousal: 83.24%	–	The granularity of emotion modeling is limited, the computational complexity is high, there are large differences between individuals, and personalized adjustments are required. Real-time performance is affected by hardware
Le Fang et al. [208]	Emo-MG, IEMOCAP, EMOTIC	EEG + micro-gesture, voice + video, image	multimodal model based on Kolmogorov v-A Arnold Network, Transformer variant with KAN attention mechanism	Emo-MG: 83.54%, IEMOCAP: 72.16%, EMOTIC (Valence dim.): 73.41%	–	The test speed is slightly slower; the IEMOCAP data is still lower than some SOTA models (such as CORRECT); the Transformer version is more complicated to calculate
Sathi-shkumar Moorthy et al. [209]	AffWild2, AFEW-VA, IEMOCAP	Speech audio, spectrogram, MFCC, facial images video	Hybrid Multi-Attention Network, Contains CSSA + HASPCM modules	IEMOCAP: 75.39%	–	The model structure is complex; it consumes a lot of resources; it takes a long time to train; it depends on the quality of annotations
Erdem et al. [210]	DEAP	EEG+ Facial expression video	GRU(main), LSTM, Transformer	Single modality GRU: 91.8%,Multimodality: 97.8%	GRU: 97%	The Transformer model is complex and requires a large amount of data; facial expression data is affected by the quality of the recording, some of it needs to be synthesized and supplemented
Ao Li et al. [131]	DEAP, WESAD	EEG, ECG, EMG, EOG, GSR	CovNet + Attention Fusion Model + Conditional Self-Attention GAN + CovNet Classifier	DEAP: 96.06%, WESAD: 95.70%	DEAP: 95.80%, WESAD: 95.45%	High reliance on generative models; requires high-computing equipment; lacks verification in real complex environments; scalability and cross-domain adaptability need further research

Table 9. *Cont.*

(b)

Author	Datasets	Signals	Classifier	Accuracy	F1 Score	Limitations
Kaveti Pavan et al. [211]	IIT Hyderabad in-house controlled driving dataset	ECG, EDA	1D-CNN with cross squeeze-and-excitation attention; LOSOCV	76.54% ECG→EDA attention; 76.37% LOSOCV mean	73.02% (best)	Small, single-site, all-male cohort; two-class setup; short windows; device- and scenario-specific; no benchmarking on public datasets; limited evidence for real-world generalization and edge deployment latency/energy.
Axel Gedeon et al. [118]	ASCERTAIN and KEMDy20	ECG, EEG, EDA, TEMP, IBI, EMO, Audio signal, Text	Cross-Attention Gated Mixture of Experts.	Arousal: 99.71%, Valence: 99.71%	Arousal: 99.83%, Valence: 99.76%	The model is highly complex, has a large number of parameters, and requires high hardware resources; it requires multimodal and complete input; its generalization ability has not been verified in real environments; and its model interpretability is limited.
Yekta et al. [212]	SWEET, DAPPER, LabToDaily	EDA, PPG, TEMP, ACC	CNN-complex	–	SWEET: 97.80%	The signal is severely affected by motion artifacts, the model is insensitive to motion state, the Transformer module is not effective when the data scale is insufficient, requires a long time window, and is not suitable for real-time emotion detection. Self-supervision requires a large amount of unlabeled data, and the training cost is high.
Thaduri et al. [213]	Stress-Lysis	ECG, EDA, Behavioral signals, Environmental signals	GNN-T-GCN	92.3%	–	The generalization ability is not verified. It requires a lot of computing resources. The mixing of environmental signals and physiological signals may introduce noise. It does not deal with motion artifacts and lost data in real wearable devices. It has not been verified under real-time monitoring conditions.

6.3. Comparison and Analysis

In order to deepen the comparison of classification methods for multimodal stress recognition, this section juxtaposes traditional machine learning with deep learning based models. Building on the representative studies summarized in Tables 8 and 9a,b, we evaluate both paradigms across data efficiency, interpretability, fusion depth, alignment capability and deployability. Table 10 distills the key contrasts and offers practical guidance for model selection and future research.

Table 10. Comparison between traditional and deep learning-based methods for multimodal stress recognition.

Modals	Feature Method	Data Scale	Interpre-Tability	Fusion Type	Alignment	Deployment Cost	Generalization	Sensitivity
Traditional Machine Learning	Manual feature extraction	Can work with small datasets	High	Mostly early or decision-level fusion	Requires manual synchronization or preprocessing	Low computational resources required; suitable for edge devices	Relatively robust in low-variance tasks	More reliant on individual feature quality
Deep Learning	Automatic feature learning via deep networks	Requires large-scale, labeled data for optimal performance	Low; needs XAI tools such as SHAP or Grad-CAM	Supports mid-level fusion via attention or KAN mechanisms	Can model inter-modal dynamics and delays	High computational cost; lightweight models needed for deployment	Sensitive to overfitting without regularization or transfer learning	Better at capturing high-order correlations between modalities

Recent studies increasingly reveal a widening gap between traditional machine learning and deep learning for multimodal emotion and stress recognition. Niu et al. demonstrated an interpretable machine learning framework that integrates EEG, ECG, and clinical variables and highlights the value of transparent physiological modeling [214]. On consumer devices, Khuntia et al. achieved high accuracy by fusing wearable ECG with speech using a deep model [215]. For cross modal correlation learning, Zhao and collaborators introduced a contrastive autoencoder-based fusion approach that strengthens inter-modality representation consistency [129]. In parallel, Feng and colleagues explored deep learning with Internet of Things pipelines for real-time monitoring and personalized ECG analysis [216]. Methodological guidance further suggests accounting for redundancy and mutual correlation when designing multimodal fusion strategies [217]. Together, these studies underscore the advantages of deep models in cross-modal modeling and fusion.

Deep learning offers clear benefits in representation learning and cross modal integration, yet important gaps remain in practice. Generalization is limited under small samples and inter subject variability, robustness is fragile under missing or asynchronous modalities and motion artefacts, interpretability and uncertainty are weak, and latency, energy, and memory costs challenge edge deployment. Contributing factors include scarce labels, heterogeneous preprocessing and evaluation protocols, and barriers to reproducibility. Addressing these issues points to several directions: self supervised pretraining to mitigate data scarcity, domain adaptation and personalization for cross subject transfer, robust objectives with explicit missing modality learning for reliability, model compression and scheduling for resource constrained deployment, and physiology informed modeling with explainability and calibrated uncertainty under standardized and reproducible protocols.

7. Key Challenges in Multimodal Deep Learning

Although multimodal deep learning enhances emotion and stress recognition by exploiting complementary physiological signals, it still faces fundamental challenges in robustness, generalization, and real-time applicability. These arise from intrinsic modality

differences, asynchronous dynamics, limited data availability, and opaque model behavior. Understanding these challenges is essential for guiding future improvements in multimodal affective computing.

Despite remarkable advances in multimodal deep learning, emotion and stress recognition systems still face persistent challenges that restrict their robustness and practical scalability. First, modal heterogeneity and structural incompatibility among physiological signals lead to information redundancy, feature misalignment, and negative transfer across modalities. Second, temporal asynchrony between fast-response signals and delayed modalities causes semantic mismatches and unstable fusion, further aggravated by device-level sampling discrepancies. Third, the field remains highly dependent on large-scale labeled datasets, yet physiological data collection and annotation are costly and subjective, resulting in small-sample bias, label imbalance, and limited cross-dataset transferability. Moreover, poor generalization and inter-subject variability expose models to distribution shifts across individuals, devices, and experimental contexts, while fusion dependency makes model performance fragile under modality dropout, noise, or domain drift. In addition, the lack of interpretability in deep multimodal architectures hinders transparency and clinical trust, and real-time deployment is constrained by computational complexity, latency, and energy limitations in wearable environments. Collectively, these challenges highlight the gap between laboratory accuracy and real-world applicability, emphasizing the need for adaptive, explainable, and resource-efficient multimodal frameworks. Table 11 summarizes these prominent challenges, their manifestations, impacts, and representative solution directions.

Building on Table 11, recent research converges on structure-aware fusion, which combines modality-specific encoders with shared and private embeddings, cross-modal attention, and adaptive gating to address heterogeneity, as well as on adaptive temporal alignment to accommodate latency mismatches. Data-efficient learning and cross-domain generalization approaches target small samples and distribution shift. To curb fusion dependency, uncertainty-aware reliability weighting, modality dropout, and robustness- or causal-oriented regularization help stabilize decisions under noise and missing inputs. Interpretability is enhanced through embedded explanatory modules and multimodal causal analysis, while real-time deployment leverages lightweight architectures alongside pruning, quantization, knowledge distillation, and edge–cloud co-inference to meet latency and energy budgets. Collectively, these directions prioritize adaptive, explainable, and resource-aware design, narrowing the gap between laboratory performance and real-world application.

In summary, the effectiveness of multimodal deep learning for emotion and stress recognition depends not only on achieving high accuracy but also on ensuring robustness, interpretability, and efficiency under real-world conditions. Future research should prioritize adaptive fusion architectures, data-efficient learning, and hardware-aware deployment to bridge the gap between laboratory performance and practical applications. Establishing standardized benchmarks across devices, subjects, and environments will be essential for the sustainable advancement of multimodal affective computing.

Table 11. Summary of key challenges in multimodal deep learning for emotion and stress recognition.

Challenge	Manifestation	Impact	Key Points	Solutions
Modal heterogeneity and structural incompatibility	Inconsistent sampling rates, temporal–spectral characteristics, and dynamic ranges across modalities cause misaligned features and redundant information.	Reduces discrimination and generalization; leads to negative transfer and unstable convergence.	Heterogeneous signals differ in temporal response and noise; shared encoders fail to capture modality-specific semantics.	Modality-specific encoders with shared–private embedding spaces; adaptive normalization and attention weighting; confidence-aware fusion; calibration-aware models [218].
Temporal asynchrony between modalities	Physiological and device-level delays cause unsynchronized data streams.	Leads to semantic mismatches, feature redundancy, and cross-modal interference; affects temporal generalization.	Static alignment cannot adapt to dynamic tasks; physiological lags and rating delays persist.	Dynamic and Soft-DTW alignment; asynchronous attention; adaptive delay modeling; cross-modal temporal calibration [219].
Dependence on large-scale labeled datasets	Physiological data collection and annotation are costly and subjective, leading to small-sample imbalance and missing modalities.	Overfitting and unstable training; poor scalability under real-world data scarcity.	Annotation subjectivity and limited public datasets restrict model capacity; multimodal imbalance worsens with missing data.	Self-supervised and contrastive learning; few-shot learning; incremental and federated adaptation; multi-expert Transformer with sparse gating [118].
Poor generalization and small-sample adaptability	Strong inter-subject variability and inconsistent experimental paradigms cause domain shift and poor cross-dataset transfer.	Degraded performance under unseen subjects, devices, or contexts.	Differences in physiological baselines and sensor configurations hinder subject-independent modeling.	Domain adaptation; invariant representation learning; meta-learning for personalization; contrastive reinforcement transfer learning [220].
Fusion dependency	Model performance heavily depends on complete, synchronized, high-quality multimodal inputs.	Missing or noisy modalities cause cascading attention mismatches and unstable predictions.	Deep fusion assumes modality complementarity and shared latent space; lacks physiological priors.	Uncertainty-aware fusion; modality dropout; reliability weighting; self-supervised multimodal contrastive learning [221].
Lack of interpretability	Deep models act as black boxes; decisions lack physiological or psychological explainability.	Limits trust, clinical adoption, and ethical approval.	SHAP and LIME improve transparency but remain locally inconsistent and computationally costly.	Embedded explainable modules; multimodal causal analysis; model-agnostic and differentiable interpretability frameworks [222].
Real-time deployment and efficiency	High latency, energy consumption, and limited compute capacity on edge devices hinder real-time response.	Prevents large-scale and wearable implementation; increases delay and power usage.	Multi-branch architectures and cross-modal attention raise complexity; quantization may reduce precision.	Model pruning, quantization, knowledge distillation; dynamic inference; edge–cloud co-inference; lightweight CNN or Transformer architectures [223].

8. Discussion and Future Directions

This review has systematically examined recent advances in multimodal emotion and stress recognition, with a particular focus on physiological signals such as ECG, EMG, and EDA and their integration through machine learning and deep learning-based fusion strategies. By organizing existing studies across physiological foundations, public and self-collected datasets, preprocessing pipelines, feature extraction, fusion paradigms, and classification models, we highlighted both the strengths and limitations of current approaches. Despite notable progress, persistent challenges—including modality heterogeneity, temporal asynchrony, limited data scale, weak cross-subject generalization, and deployment constraints—continue to restrict real-world applicability and reproducibility.

Overall, the reviewed evidence indicates that multimodal fusion can substantially enhance stress recognition performance when modalities are complementary and well aligned, but may degrade performance under noisy, weakly informative, or data-limited conditions. Deep learning-based models offer superior representation learning and adaptive fusion capabilities compared with traditional machine learning pipelines, yet their benefits are often offset by high data dependence, computational cost, and limited interpretability. These

findings underscore the importance of principled fusion design, transparent evaluation, and robustness-oriented modeling in physiological stress analysis.

Looking forward, future research should prioritize a small number of key directions to strengthen translational impact: the development of standardized and ecologically valid multimodal datasets; robust and lightweight fusion architectures that balance accuracy with efficiency; alignment-aware and uncertainty-aware learning to handle modality heterogeneity and missing data; and evaluation protocols that explicitly consider generalization, robustness, and deployability.

In conclusion, multimodal emotion and stress recognition remains a promising yet challenging field at the intersection of affective computing and intelligent healthcare. By consolidating current knowledge and identifying critical methodological gaps, this review provides a structured foundation for developing reliable, interpretable, and scalable stress recognition systems that can move beyond laboratory settings toward real-world deployment.

Author Contributions: Conceptualization, X.Z. and M.X.; methodology, X.Z.; formal analysis, X.Z.; investigation, X.Z.; resources, M.X.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z. and H.Z.; visualization, X.Z.; supervision, M.X.; project administration, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

ECG	Electrocardiogram
EDA	Electrodermal Activity
EMG	Electromyography
EEG	Electroencephalography
HRV	Heart Rate Variability
SCL	Skin Conductance Level
SCR	Skin Conductance Response
MUAP	Motor Unit Action Potential
PSS	Perceived Stress Scale
SRQ	Stress Response Questionnaire
STAI	State–Trait Anxiety Inventory
PANAS	Positive and Negative Affect Schedule
SVM	Support Vector Machine
KNN	k-Nearest Neighbors
HMM	Hidden Markov Model
CRF	Conditional Random Field
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
KAN	Kolmogorov–Arnold Network

MoE	Mixture of Experts
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
DTW	Dynamic Time Warping
MMD	Maximum Mean Discrepancy
DANN	Domain-Adversarial Neural Network
CORAL	Correlation Alignment
LOSOCV	Leave-One-Subject-Out Cross Validation
AUROC	Area Under the Receiver Operating Characteristic Curve
BVP	Blood Volume Pulse
IBI	Inter-Beat Interval
PPG	Photoplethysmography
VR	Virtual Reality
AR	Augmented Reality

References

1. Christiansen, J.; Qualter, P.; Friis, K.; Pedersen, S.; Lund, R.; Andersen, C.; Bekker-Jeppesen, M.; Lasgaard, M. Associations of loneliness and social isolation with physical and mental health among adolescents and young adults. *Perspect. Public Health* **2021**, *141*, 226–236. [CrossRef] [PubMed]
2. Baxter, A.J.; Scott, K.M.; Ferrari, A.J.; Norman, R.E.; Vos, T.; Whiteford, H.A. Challenging the myth of an "epidemic" of common mental disorders: Trends in the global prevalence of anxiety and depression between 1990 and 2010. *Depress. Anxiety* **2014**, *31*, 506–516. [CrossRef]
3. Franks, K.H.; Rowsthorn, E.; Bransby, L.; Lim, Y.Y.; Chong, T.T.J.; Pase, M.P. Association of self-reported psychological stress with cognitive decline: A systematic review. *Neuropsychol. Rev.* **2023**, *33*, 856–870. [CrossRef]
4. Merten, T. The self-report fallacy: When diagnosis predominantly relies on subjective symptom report. *Curr. Opin. Psychol.* **2025**, *65*, 102096. [CrossRef]
5. Sara, J.D.S.; Toya, T.; Ahmad, A.; Clark, M.M.; Gilliam, W.P.; Lerman, L.O.; Lerman, A. Mental stress and its effects on vascular health. *Mayo Clin. Proc.* **2022**, *97*, 951–990. [CrossRef]
6. Janse, P.D.; van Sonsbeek, M.A.; Bovendeerd, B.; de Jong, K. Progress feedback in psychotherapy: Advantages, challenges, and future directions. *Curr. Opin. Psychol.* **2025**, *66*, 102110. [CrossRef]
7. Naeem, M.; Fawzi, S.A.; Anwar, H.; Malek, A.S. Wearable ECG systems for accurate mental stress detection: A scoping review. *J. Public Health* **2025**, *33*, 1181–1197. [CrossRef]
8. Kapase·, A.B.; Uke, N. A comprehensive review in affective computing: An exploration of artificial intelligence in unimodal and multimodal emotion recognition systems. *Int. J. Speech Technol.* **2025**, *28*, 541–563. [CrossRef]
9. Ometov, A.; Mezina, A.; Lin, H.C.; Arponen, O.; Burget, R.; Nurmi, J. Stress and emotion open access data: A review on datasets, modalities, methods, challenges, and future research perspectives. *J. Healthc. Inform. Res.* **2025**, *9*, 247–279. [CrossRef]
10. Haque, Y.; Zawad, R.S.; Rony, C.S.A.; Al Banna, H.; Ghosh, T.; Kaiser, M.S.; Mahmud, M. State-of-the-art of stress prediction from heart rate variability using artificial intelligence. *Cogn. Comput.* **2024**, *16*, 455–481. [CrossRef]
11. Zapf, H.; Boettcher, J.; Haukeland, Y.; Orm, S.; Coslar, S.; Fjermestad, K. A systematic review of the association between parent-child communication and adolescent mental health. *JCPP Adv.* **2024**, *4*, e12205. [CrossRef]
12. Madaan, A.; Singh, A. From sensors to insight: A review of physiological signal processing for stress prediction. In *Proceedings of the AI-Driven Smart Healthcare for Society 5.0, Kolkata, India, 14–15 February 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 182–187. [CrossRef]
13. Kotlęga, D.; Gołąb-Janowska, M.; Masztalewicz, M.; Ciećwież, S.; Nowacki, P. The emotional stress and risk of ischemic stroke. *Neurol. Neurochir. Pol.* **2016**, *50*, 265–270. [CrossRef]
14. Lazarus, R.S. Psychological stress and coping in adaptation and illness. *Int. J. Psychiatry Med.* **1974**, *5*, 321–333. [CrossRef]
15. Levenson, R.W. The autonomic nervous system and emotion. *Emot. Rev.* **2014**, *6*, 100–112. [CrossRef]
16. Cipresso, P.; Colombo, D.; Riva, G. Computational psychometrics using psychophysiological measures for the assessment of acute mental stress. *Sensors* **2019**, *19*, 781. [CrossRef] [PubMed]
17. Giannakakis, G.; Grigoriadis, D.; Giannakaki, K.; Simantiraki, O.; Roniotis, A.; Tsiknakis, M. Review on psychological stress detection using biosignals. *IEEE Trans. Affect. Comput.* **2022**, *13*, 440–460. [CrossRef]
18. Hammad, M.; Maher, A.; Wang, K.; Jiang, F.; Amrani, M. Detection of abnormal heart conditions based on characteristics of ECG signals. *Measurement* **2018**, *125*, 634–644. [CrossRef]

19. Kim, H.G.; Cheon, E.J.; Bai, D.S.; Lee, Y.H.; Koo, B.H. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investig.* **2018**, *15*, 235–245. [CrossRef] [PubMed]
20. Billman, G.E. The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Front. Physiol.* **2013**, *4*, 26. [CrossRef]
21. Castaldo, R.; Melillo, P.; Bracale, U.; Caserta, M.; Triassi, M.; Pecchia, L. Acute mental stress assessment via short-term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomed. Signal Process. Control* **2015**, *18*, 370–377. [CrossRef]
22. Moritani, T.; Stegeman, D.; Merletti, R. Basic physiology and biophysics of EMG signal generation. In *Electromyography: Physiology, Engineering, and Noninvasive Applications*; Wiley: Hoboken, NJ, USA, 2004; pp. 1–25.
23. Farina, D.; Stegeman, D.; Merletti, R. Biophysics of the generation of EMG signals. In *Surface Electromyography: Physiology, Engineering, and Applications*; Wiley: Hoboken, NJ, USA, 2016; pp. 1–24.
24. Kret, M.; Stekelenburg, J.; Roelofs, K.; De Gelder, B. Perception of Face and Body Expressions Using Electromyography, Pupillometry and Gaze Measures. *Front. Psychol.* **2013**, *4*, 28. [CrossRef] [PubMed]
25. Ahmed, M.; Grillo, M.; Taebi, A.; Kaya, M.; Thibbotuwawa Gamage, P. A Comprehensive Analysis of Trapezius Muscle EMG Activity in Relation to Stress and Meditation. *BioMedInformatics* **2024**, *4*, 1047–1058. [CrossRef]
26. Rissanen, J. Collecting Biosignals: Data Experiments with EDA and EEG. Master's Thesis, Tampere University of Applied Sciences, Tampere, Finland, 2024.
27. Braithwaite, J.J.; Watson, D.G.; Jones, R.; Rowe, M. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* **2013**, *49*, 1017–1034.
28. Rahma, O.N.; Putra, A.P.; Rahmatillah, A.; Putri, Y.S.K.A.; Fajriaty, N.D.; Ain, K.; Chai, R. Electrodermal activity for measuring cognitive and emotional stress level. *J. Med. Signals Sens.* **2022**, *12*, 155–162. [CrossRef]
29. Liu, Y.; Du, S. Psychological stress level detection based on electrodermal activity. *Behav. Brain Res.* **2018**, *341*, 50–53. [CrossRef] [PubMed]
30. Bari, D.S. Gender differences in tonic and phasic electrodermal activity components. *Sci. J. Univ. Zakho* **2020**, *8*, 29–33. [CrossRef]
31. Greco, A.; Valenza, G.; Scilingo, E.P. Modeling for the Analysis of the EDA. In *Advances in Electrodermal Activity Processing with Applications for Mental Health: From Heuristic Methods to Convex Optimization*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 19–33.
32. Ernst, H.; Scherpf, M.; Pannasch, S.; Helmert, J.R.; Malberg, H.; Schmidt, M. Assessment of the human response to acute mental stress—An overview and a multimodal study. *PLoS ONE* **2023**, *18*, e0294069. [CrossRef]
33. Khalili, M.; GholamHosseini, H.; Lowe, A.; Kuo, M.M. Motion artifacts in capacitive ECG monitoring systems: A review of existing models and reduction techniques. *Med. Biol. Eng. Comput.* **2024**, *62*, 3599–3622. [CrossRef]
34. Boyer, M.; Bouyer, L.; Roy, J.S.; Campeau-Lecours, A. Reducing Noise, Artifacts and Interference in Single-Channel EMG Signals: A Review. *Sensors* **2023**, *23*, 2927. [CrossRef]
35. Yang, S.; Gao, Y.; Zhu, Y.; Zhang, L.; Xie, Q.; Lu, X.; Wang, F.; Zhang, Z. A deep learning approach to stress recognition through multimodal physiological signal image transformation. *Sci. Rep.* **2025**, *15*, 22258. [CrossRef]
36. Li, J.; Li, J.; Wang, X.; Zhan, X.; Zeng, Z. A Domain Generalization and Residual Network-Based Emotion Recognition from Physiological Signals. *Cyborg Bionic Syst.* **2024**, *5*, 74. [CrossRef] [PubMed]
37. Qasim, M.S.; Bari, D.S.; Martinsen, Ø.G. Influence of ambient temperature on tonic and phasic electrodermal activity components. *Physiol. Meas.* **2022**, *43*, 065001. [CrossRef] [PubMed]
38. Khan, T.H.; Villanueva, I.; Vicioso, P.; Husman, J. Exploring relationships between electrodermal activity, skin temperature, and performance during. In *Proceedings of the 2019 IEEE Frontiers in Education Conference (FIE), Covington, KY, USA, 16–19 October 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5. [CrossRef]
39. Posada-Quintero, H.F.; Reljin, N.; Mills, C.; Mills, I.; Florian, J.P.; VanHeest, J.L.; Chon, K.H. Time-varying analysis of electrodermal activity during exercise. *PLoS ONE* **2018**, *13*, e0198328. [CrossRef]
40. Bjaastad, J.F.; Jensen-Doss, A.; Moltu, C.; Jakobsen, P.; Hagenberg, H.; Joa, I. Attitudes toward standardized assessment tools and their use among clinicians in a public mental health service. *Nord. J. Psychiatry* **2019**, *73*, 387–396. [CrossRef]
41. Lee, E.H. Review of the psychometric evidence of the perceived stress scale. *Asian Nurs. Res.* **2012**, *6*, 121–127. [CrossRef]
42. Spielberger, C.D.; Gorsuch, R.L.; Lushene, R.; Vagg, P.R.; Jacobs, G.A. *Manual for the State-Trait Anxiety Inventory (STAI)*; Consulting Psychologists Press: Palo Alto, CA, USA, 1983; ISBN 0-87120-197-6.
43. Watson, D.; Clark, L.A.; Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personal. Soc. Psychol.* **1988**, *54*, 1063. [CrossRef]
44. Beusenberg, M.; Orley, J.H.; WHO. *A User's Guide to the Self Reporting Questionnaire (SRQ/Compiled by M. Beusenberg and J. Orley*; WHO: Geneva, Switzerland, 1994.
45. Ringgold, V.; Burkhardt, F.; Abel, L.; Kurz, M.; Müller, V.; Richer, R.; Eskofier, B.M.; Shields, G.S.; Rohleder, N. Multimodal stress assessment: Connecting task-related changes in self-reported stress, salivary biomarkers, heart rate, and facial expressions in the context of the stress response to the Trier Social Stress Test. *Psychoneuroendocrinology* **2025**, *180*, 107560. [CrossRef]

46. Wuensch, M.; Frenzel, A.C.; Pekrun, R.; Sun, L. Enjoyable for some, stressful for others? Physiological and subjective indicators of achievement emotions during adaptive versus fixed-item testing. *Contemp. Educ. Psychol.* **2025**, *82*, 102388. [CrossRef]
47. Kalateh, S.; Estrada-Jimenez, L.A.; Nikghadam-Hojjati, S.; Barata, J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access* **2024**, *12*, 103976–104019. [CrossRef]
48. Ladakis, I.; Fotopoulos, D.; Chouvarda, I. Integrative Analysis of Open Datasets for Stress Prediction. *J. Med. Biol. Eng.* **2025**, *45*, 385–399. [CrossRef]
49. Zhang, X.; Wei, X.; Zhou, Z.; Zhao, Q.; Zhang, S.; Yang, Y.; Li, R.; Hu, B. Dynamic Alignment and Fusion of Multimodal Physiological Patterns for Stress Recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 685–696. [CrossRef]
50. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction, Paris, France, 9–13 October 2018*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 400–408.
51. Miranda-Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* **2018**, *12*, 479–493. [CrossRef]
52. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2011**, *3*, 42–55. [CrossRef]
53. Katsigiannis, S.; Ramzan, N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [CrossRef]
54. Healey, J.A.; Picard, R.W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 156–166. [CrossRef]
55. Jiang, Z.; Seyedi, S.; Griner, E.; Abbasi, A.; Rad, A.B.; Kwon, H.; Cotes, R.O.; Clifford, G.D. Evaluating and mitigating unfairness in multimodal remote mental health assessments. *PLoS Digit. Health* **2024**, *3*, e0000413. [CrossRef]
56. Mordacq, J.; Milecki, L.; Vakalopoulou, M.; Oudot, S.; Kalogeiton, V. Multimodal Learning for Detecting Stress under Missing Modalities. In *Proceedings of the WiCV 2024—Women in Computer Vision Workshop in Conjunction with CVPR, Seattle, WA, USA, 18 June 2024*; IEEE: Piscataway, NJ, USA, 2024
57. Rodrigues, S.; Kaiseler, M.; Queirós, C. Psychophysiological Assessment of Stress Under Ecological Settings. *Eur. Psychol.* **2015**, *20*, 204–226. [CrossRef]
58. Zhu, X.; Guo, C.; Feng, H.; Huang, Y.; Feng, Y.; Wang, X.; Wang, R. A review of key technologies for emotion analysis using multimodal information. *Cogn. Comput.* **2024**, *16*, 1504–1530. [CrossRef]
59. Carvalho, M.; Pinho, A.J.; Brás, S. Resampling approaches to handle class imbalance: A review from a data perspective. *J. Big Data* **2025**, *12*, 71. [CrossRef]
60. Huang, Y.; Zhang, Z.; Yang, Y.; Mo, P.C.; Zhang, Z.; He, J.; Hu, S.; Wang, X.; Li, Y. Exploring Skin Potential Signals in Electrodermal Activity: Identifying Key Features for Attention State Differentiation. *IEEE Access* **2024**, *12*, 100832–100847. [CrossRef]
61. Akpinar, M.H.; Sengur, A.; Salvi, M.; Seoni, S.; Faust, O.; Mir, H.; Molinari, F.; Acharya, U.R. Synthetic Data Generation via Generative Adversarial Networks in Healthcare: A Systematic Review of Image- and Signal-Based Studies. *IEEE Open J. Eng. Med. Biol.* **2025**, *6*, 183–192. [CrossRef]
62. Salazar, A.; Vergara, L.; Safont, G. Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets. *Expert Syst. Appl.* **2021**, *163*, 113819. [CrossRef]
63. Fajardo, V.A.; Findlay, D.; Jaiswal, C.; Yin, X.; Houmanfar, R.; Xie, H.; Liang, J.; She, X.; Emerson, D.B. On oversampling imbalanced data with deep conditional generative models. *Expert Syst. Appl.* **2021**, *169*, 114463. [CrossRef]
64. Wu, Y.; Mi, Q.; Gao, T. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics* **2025**, *10*, 418. [CrossRef]
65. Dominguez-Catena, I.; Paternain, D.; Galar, M. Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5209–5226. [CrossRef] [PubMed]
66. Xu, H.; Zeng, M.; Liu, H.; Xie, X.; Tian, L.; Yan, J.; Chen, C. A dynamic transfer network for cross-database atrial fibrillation detection. *Biomed. Signal Process. Control* **2024**, *90*, 105799. [CrossRef]
67. Al-Azani, S.; El-Alfy, E.S.M. A review and critical analysis of multimodal datasets for emotional AI. *Artif. Intell. Rev.* **2025**, *58*, 334. [CrossRef]
68. Zhang, B.; Morère, Y.; Sieler, L.; Langlet, C.; Bolmont, B.; Bourhis, G. Reaction time and physiological signals for stress recognition. *Biomed. Signal Process. Control* **2017**, *38*, 100–107. [CrossRef]
69. Romeo, Z.; Fusina, F.; Semenzato, L.; Bonato, M.; Angrilli, A.; Spironelli, C. Comparison of Slides and Video Clips as Different Methods for Inducing Emotions: An Electroencephalographic Alpha Modulation Study. *Front. Hum. Neurosci.* **2022**, *16*, 901422. [CrossRef]
70. Parsons, T.D. Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences. *Front. Hum. Neurosci.* **2015**, *9*, e00660. [CrossRef]

71. Williams, J.M.G.; Mathews, A.; MacLeod, C. The emotional Stroop task and psychopathology. *Psychol. Bull.* **1996**, *120*, 3. [CrossRef] [PubMed]
72. Galy, E.; Mélan, C. Effects of cognitive appraisal and mental workload factors on performance in an arithmetic task. *Appl. Psychophysiol. Biofeedback* **2015**, *40*, 313–325. [CrossRef]
73. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. International affective picture system (IAPS): Technical manual and affective ratings. *Nimh Cent. Study Emot. Atten.* **1997**, *1*, 3.
74. Allen, A.P.; Kennedy, P.J.; Dockray, S.; Cryan, J.F.; Dinan, T.G.; Clarke, G. The trier social stress test: Principles and practice. *Neurobiol. Stress* **2017**, *6*, 113–126. [CrossRef]
75. Razzak, R.; Li, Y.; Sokhadze, E.; He, S. Stress and Driving Performance Evaluation through VR and Physiological Metrics: A Pilot Study. *JISARA* **2025**, *18*, 30. [CrossRef]
76. Nasri, M. Towards Intelligent VR Training: A Physiological Adaptation Framework for Cognitive Load and Stress Detection. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA, 16–19 June 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 419–423.
77. Mahesh, B.; Weber, D.; Garbas, J.; Foltyn, A.; Oppelt, M.; Becker, L.; Rohleder, N.; Lang, N. Setup for Multimodal Human Stress Dataset Collection. In Proceedings of the 12th International Conference on Methods and Techniques in Behavioral Research, and 6th Seminar on Behavioral Methods, Virtual, 18–20 May 2022.
78. Ferreira, S.O. Emotional activation in human beings: Procedures for experimental stress induction. *Psicol. USP* **2019**, *30*, e180176. [CrossRef]
79. Chaptoukaev, H.; Strizhkova, V.; Panariello, M.; D'Alpaos, B.; Reka, A.; Manera, V.; Thümmler, S.; Ismailova, E.; Evans, N.; Bremond, F.; et al. StressID: A Multimodal Dataset for Stress Identification. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 29798–29811.
80. Miranda Calero, J.A.; Gutiérrez-Martín, L.; Rituerto-González, E.; Romero-Perales, E.; Lanza-Gutiérrez, J.M.; Peláez-Moreno, C.; López-Ongil, C. Wemac: Women and emotion multi-modal affective computing dataset. *Sci. Data* **2024**, *11*, 1182. [CrossRef]
81. Rashed, A.; Shirmohammadi, S.; Hefeeda, M. Descriptor: Multimodal Dataset for Player Engagement Analysis in Video Games (MultiPENG). *IEEE Data Descr.* **2025**, *2*, 17–25. [CrossRef]
82. Heimerl, A.; Prajod, P.; Mertes, S.; Baur, T.; Kraus, M.; Liu, A.; Risack, H.; Rohleder, N.; André, E.; Becker, L. The ForDigitStress Dataset: A Multi-Modal Dataset for Automatic Stress Recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1219–1234. [CrossRef]
83. Jaiswal, M.; Bara, C.-P.; Luo, Y.; Burzo, M.; Mihalcea, R.; Provost, E.M. MuSE: A Multimodal Dataset of Stressed Emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020*; European Language Resources Association: Reykjavik, Iceland, 2020.
84. Hosseini, M.; Sohrab, F.; Gottumukkala, R.; Bhupatiraju, R.T.; Katragadda, S.; Raitoharju, J.; Iosifidis, A.; Gabbouj, M. Empathic-School: A multimodal dataset for real-time facial expressions and physiological data analysis under different stress conditions. *arXiv* **2022**, arXiv:2209.13542.
85. Tabbaa, L.; Searle, R.; Bafti, S.M.; Hossain, M.M.; Intarasisrisawat, J.; Glancy, M.; Ang, C.S. VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures. *Proc. Acm Interact. Mob. Wearable Ubiquitous Technol.* **2022**, 1–20. [CrossRef]
86. Oppelt, M.P.; Foltyn, A.; Deuschel, J.; Lang, N.R.; Holzer, N.; Eskofier, B.M.; Yang, S.H. ADABase: A Multimodal Dataset for Cognitive Load Estimation. *Sensors* **2023**, *23*, 340. [CrossRef]
87. Soon, P.S.; Lim, W.M.; Gaur, S.S. The role of emotions in augmented reality. *Psychol. Mark.* **2023**, *40*, 2387–2412. [CrossRef]
88. Dahiya, V. Interactive Emotional Resonance: Bidirectional Communication Between Heart Rate-Derived Player States, Game Music, and Gameplay Events. Master's Thesis, Drexel University, Philadelphia, PA, USA, 2025.
89. Paniagua-Gómez, M.; Fernandez-Carmona, M. Trends and Challenges in Real-Time Stress Detection and Modulation: The Role of the IoT and Artificial Intelligence. *Electronics* **2025**, *14*, 2581. [CrossRef]
90. Yan, J.; Yue, Y.; Yu, K.; Zhou, X.; Liu, Y.; Wei, J.; Yang, Y. Multi-Representation Joint Dynamic Domain Adaptation Network for Cross-Database Facial Expression Recognition. *Electronics* **2024**, *13*, 1470. [CrossRef]
91. Yan, L.; Gašević, D.; Echeverria, V.; Zhao, L.; Jin, Y.; Li, X.; Martinez-Maldonado, R. In Sync or Out of Sync? Understanding Stress and Learning Performance in Collaborative Healthcare Simulations through Physiological Synchrony and Arousal. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 2421–2452. [CrossRef]
92. Yadav, G.; Bokhari, M.U. Hybrid classifier for optimizing mental health prediction: Feature engineering and fusion technique. *Int. J. Ment. Health Addict.* **2024**, *22*, 1–41. [CrossRef]
93. Silva, R.; Salvador, G.; Bota, P.; Fred, A.; Plácido da Silva, H. Impact of sampling rate and interpolation on photoplethysmography and electrodermal activity signals' waveform morphology and feature extraction. *Neural Comput. Appl.* **2023**, *35*, 5661–5677. [CrossRef]
94. Li, Z.; Tian, Y.; Jin, Y.; Wei, X.; Wang, M.; Liu, J.; Liu, C. EDDM: A Novel ECG Denoising Method Using Dual-Path Diffusion Model. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2509815. [CrossRef]

95. Huang, W.; Chen, Y.; Jiang, X.; Zhang, T.; Chen, Q. GJFusion: A channel-level correlation construction method for multimodal physiological signal fusion. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 1–23. [CrossRef]
96. Iammarino, E.; Marcantoni, I.; Sbrollini, A.; Morettini, M.; Burattini, L. Normalization of Electrocardiogram-Derived Cardiac Risk Indices: A Scoping Review of the Open-Access Literature. *Appl. Sci.* **2024**, *14*, 9457. [CrossRef]
97. Kantharaju, P.; Vakacherla, S.S.; Jacobson, M.; Jeong, H.; Mevada, M.N.; Zhou, X.; Major, M.J.; Kim, M. Framework for personalizing wearable devices using real-time physiological measures. *IEEE Access* **2023**, *11*, 81389–81400. [CrossRef]
98. Orlhac, F.; Eertink, J.J.; Cottereau, A.S.; Zijlstra, J.M.; Thieblemont, C.; Meignan, M.; Boellaard, R.; Buvat, I. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J. Nucl. Med.* **2022**, *63*, 172–179. [CrossRef]
99. Rahman, S.; Karmakar, C.; Natgunanathan, I.; Yearwood, J.; Palaniswami, M. Robustness of electrocardiogram signal quality indices. *J. R. Soc. Interface* **2022**, *19*, 20220012. [CrossRef]
100. Bahador, N. Assessment of Neurological Function with Multimodal and Multichannel Physiological Signal Analysis Using Machine and Deep Learning Techniques. Ph.D. Thesis, University of Oulu, Oulu, Finland, 2024.
101. Liu, C.L.; Xiao, B.; Hsieh, C.H. Multimodal fusion of spatial-temporal and frequency representations for enhanced ECG classification. *Inf. Fusion* **2025**, *118*, 102999. [CrossRef]
102. Dalmeida, K.M.; Masala, G.L. HRV features as viable physiological markers for stress detection using wearable devices. *Sensors* **2021**, *21*, 2873. [CrossRef]
103. Schneider, M.; Kraemmer, M.M.; Weber, B.; Schwerdtfeger, A.R. Life events are associated with elevated heart rate and reduced heart complexity to acute psychological stress. *Biol. Psychol.* **2021**, *163*, 108116. [CrossRef]
104. Chen, T.; Ma, Y.; Pan, Z.; Wang, W.; Yu, J. Fusion of multi-scale feature extraction and adaptive multi-channel graph neural network for 12-lead ECG classification. *Comput. Methods Programs Biomed.* **2025**, *265*, 108725. [CrossRef] [PubMed]
105. Telangore, H.; Sharma, N.; Sharma, M.; Acharya, U.R. A novel ECG-based approach for classifying psychiatric disorders: Leveraging wavelet scattering networks. *Med. Eng. Phys.* **2025**, *135*, 104275. [CrossRef]
106. Rauf, U.; Saeed, S.M.U. Towards Improved Classification of Perceived Stress using Time Domain Features. *IEEE Access* **2024**, *12*, 51650–51664. [CrossRef]
107. Xiang, Y.; Zhang, X.; Zhang, W.; Dou, Z.; Wang, T. Wrist Motion Regression Using EMG Attention Feature Fusion Algorithm. *IEEE Sens. J.* **2025**, *early access*. [CrossRef]
108. Kartowisastro, I.H.; Trisetyarso, A.; Budiharto, W.; Sudimanto. Pain Classification Using Discrete Wavelet Transform Feature Extraction and Machine Learning Techniques. *IEEE Access* **2025**, *13*, 45912–45922. [CrossRef]
109. Ghosh, S.; Tripathi, K.; Garg, A.; Singh, D.; Prasad, A.; Bhavsar, A.; Dutt, V. Predicting Stress among Students via Psychometric Assessments and Machine Learning. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, Crete, Greece, 26–28 June 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 662–669.
110. Dogan, G.; Akbulut, F.P. Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress. *Neural Comput. Appl.* **2023**, *35*, 24435–24454. [CrossRef]
111. Rashid, N.; Mortlock, T.; Al Faruque, M.A. Stress detection using context-aware sensor fusion from wearable devices. *IEEE Internet Things J.* **2023**, *10*, 14114–14127. [CrossRef]
112. Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv* **2024**, arXiv:2404.18947. [CrossRef]
113. Hussain, M.; O'Nils, M.; Lundgren, J.; Mousavirad, S.J. A Comprehensive Review on Deep Learning-Based Data Fusion. *IEEE Access* **2024**, *12*, 180093–180124. [CrossRef]
114. Bodaghi, M.; Hosseini, M.; Gottumukkala, R. A multimodal intermediate fusion network with manifold learning for stress detection. In *Proceedings of the 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 13–14 April 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.
115. Duan, J.; Xiong, J.; Li, Y.; Ding, W. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Inf. Fusion* **2024**, *110*, 102536. [CrossRef]
116. Singh, R.; Ranjan, V.; Ganguly, A.; Halder, S. Physiological Patterns Classification of HRV Dynamics through Feature-Level Fusion and Machine Learning during Chi Meditation. *Eng. Lett.* **2025**, *33*, 1759.
117. Wang, L.; Zhang, Y.; Zhou, B.; Cao, S.; Hu, K.; Tan, Y. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. *Comput. Electr. Eng.* **2024**, *118*, 109413. [CrossRef]
118. Mengara Mengara, A.G.; Moon, Y.K. CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts. *Mathematics* **2025**, *13*, 1907. [CrossRef]
119. Roy, S.; Ogidi, F.; Etemad, A.; Dolatabadi, E.; Afkanpour, A. A Shared Encoder Approach to Multimodal Representation Learning. *arXiv* **2025**, arXiv:2503.01654. [CrossRef]
120. Zhu, J.; Li, Y.; Yang, C.; Cai, H.; Li, X.; Hu, B. Transformer-based fusion model for mild depression recognition with EEG and pupil area signals. *Med. Biol. Eng. Comput.* **2025**, *63*, 2011–2027. [CrossRef]

121. Ghose, D.; Gitelson, O.; Scassellati, B. Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24), San Jose, Costa Rica, 4–8 November 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 22–32. [CrossRef]
122. Kim, H.; Hong, T. Enhancing emotion recognition using multimodal fusion of physiological, environmental, personal data. *Expert Syst. Appl.* **2024**, *249*, 123723. [CrossRef]
123. Wang, M.; Fan, S.; Li, Y.; Xie, Z.; Chen, H. Missing-modality enabled multi-modal fusion architecture for medical data. *J. Biomed. Inform.* **2025**, *164*, 104796. [CrossRef] [PubMed]
124. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomed. Signal Process. Control* **2023**, *82*, 104561. [CrossRef]
125. Pereira, L.M.; Salazar, A.; Vergara, L. A Comparative Analysis of Early and Late Fusion for the Multimodal Two-Class Problem. *IEEE Access* **2023**, *11*, 84283–84300. [CrossRef]
126. Ramteke, R.B.; Gajbhiye, G.O.; Thool, V.R. Discriminating psychological stress levels: Multi-level attentive LSTM approach. *Neural Comput. Appl.* **2025**, *37*, 25579–25599. [CrossRef]
127. Kim, S.H. Mifu-ER: Modality Quality Index-based Incremental Fusion for Emotion Recognition. *IEEE Access* **2025**, *13*, 112703–112719. [CrossRef]
128. Frey, S.; Spacone, G.; Cossettini, A.; Guermandi, M.; Schilk, P.; Benini, L.; Kartsch, V. BioGAP-Ultra: A Modular Edge-AI Platform for Wearable Multimodal Biosignal Acquisition and Processing. *arXiv* **2025**, arXiv:2508.13728. [CrossRef]
129. Zhao, S.; Hu, Y.; Chen, J.; Wang, W.; Hu, X. Multi-source Signal Fusion with Contrastive AutoEncoder for Emotion Classification. *IEEE J. Biomed. Health Inform.* **2025**, 1–14. [CrossRef]
130. Farmani, J.; Bargshady, G.; Gkikas, S.; Tsiknakis, M.; Rojas, R.F. A CrossMod-Transformer deep learning framework for multi-modal pain detection through EDA and ECG fusion. *Sci. Rep.* **2025**, *15*, 29467. [CrossRef] [PubMed]
131. Li, A.; Wu, M.; Ouyang, R.; Wang, Y.; Li, F.; Lv, Z. A Multimodal-Driven Fusion Data Augmentation Framework for Emotion Recognition. *IEEE Trans. Artif. Intell.* **2025**, *6*, 2083–2097. [CrossRef]
132. Mansourian, N.; Mohammadi, A.; Ahmad, M.O.; Swamy, M. ECG-EmotionNet: Nested mixture of expert (NMoE) adaptation of ECG-foundation model for driver emotion recognition. *arXiv* **2025**, arXiv:2503.01750.
133. Gkikas, S.; Kyprakis, I.; Tsiknakis, M. Tiny-biomoe: A lightweight embedding model for biosignal analysis. In *Proceedings of the Companion Proceedings of the 27th International Conference on Multimodal Interaction, Canberra, Australia, 13–17 October 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 117–126.
134. Li, Y.; Li, Y.; He, X.; Fang, J.; Zhou, C.; Liu, C. Learner's cognitive state recognition based on multimodal physiological signal fusion. *Appl. Intell.* **2025**, *55*, 127. [CrossRef]
135. Choi, H.S. Emotion Recognition Using a Siamese Model and a Late Fusion-Based Multimodal Method in the WESAD Dataset with Hardware Accelerators. *Electronics* **2025**, *14*, 723. [CrossRef]
136. Muke, P.Z.; Kozierkiewicz, A. Machine learning techniques to improve the cognitive workload classification using multimodal sensors' data. *IEEE Access* **2025**, *13*, 173415–173443. [CrossRef]
137. Guo, Y.; Yang, K.; Wu, Y. A multi-modality attention network for driver fatigue detection based on frontal EEG, EDA and PPG signals. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 4009–4022. [CrossRef]
138. Fang, C.; Sandino, C.; Mahasseni, B.; Minxha, J.; Pouransari, H.; Azemi, E.; Moin, A.; Zippi, E. Promoting cross-modal representations to improve multimodal foundation models for physiological signals. *arXiv* **2024**, arXiv:2410.16424. [CrossRef]
139. Hou, K.; Zhang, X.; Yang, Y.; Zhao, Q.; Yuan, W.; Zhou, Z.; Zhang, S.; Li, C.; Shen, J.; Hu, B. Emotion recognition from multimodal physiological signals via discriminative correlation fusion with a temporal alignment mechanism. *IEEE Trans. Cybern.* **2023**, *54*, 3079–3092. [CrossRef]
140. Tang, Z.; Qi, J.; Zheng, Y.; Huang, J. A Comprehensive Benchmark for Electrocardiogram Time-Series. *arXiv* **2025**, arXiv:2507.14206.
141. Roy, K.; Rao, A.C.S. Self-Supervised Learning of Cardiac Dynamics Using Masked Volume Modeling (MVM). *TechRxiv* **2025**. [CrossRef]
142. Nourbakhsh, A.; Mohammadzade, H. Deep Time Warping for Multiple Time Series Alignment. *arXiv* **2025**, arXiv:2502.16324. [CrossRef]
143. Kurtek, S.; Wu, W.; Christensen, G.E.; Srivastava, A. Segmentation, alignment and statistical analysis of biosignals with application to disease classification. *J. Appl. Stat.* **2013**, *40*, 1270–1288. [CrossRef]
144. Yang, H.C.; Lee, C.C. An attribute-invariant variational learning for emotion recognition using physiology. In *Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1184–1188.
145. Zhang, Y.; Cai, H.; Wu, J.; Xie, L.; Xu, M.; Ming, D.; Yan, Y.; Yin, E. EMG-based cross-subject silent speech recognition using conditional domain adversarial network. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 2282–2290. [CrossRef]

146. Li, W.; Hou, B.; Shao, S.; Huan, W.; Tian, Y. Spatial-temporal constraint learning for cross-subject EEG-based emotion recognition. In *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
147. Viana-Matesanz, M.; Sánchez-Ávila, C. Adaptive normalization and feature extraction for electrodermal activity analysis. *Mathematics* **2024**, *12*, 202. [CrossRef]
148. Hou, M.; Zhang, Z.; Liu, C.; Lu, G. Semantic alignment network for multi-modal emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5318–5329. [CrossRef]
149. Pang, L. Contrastive Learning Neural Network with Multimodal Physiological Signal Fusion for Early Detection of Cognitive Impairment. In *Proceedings of the 2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Beijing, China, 20–22 June 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1691–1695.
150. Ramaswamy, M.P.A.; Palaniswamy, S. EOG and PPG fusion for subject independent multimodal emotion recognition: A prototypical networks approach. In *Proceedings of the 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 23–25 October 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1635–1640.
151. Han, E.G.; Kang, T.K.; Lim, M.T. Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features. *Electronics* **2023**, *12*, 2933. [CrossRef]
152. Vieluf, S.; Hasija, T.; Kuschel, M.; Reinsberger, C.; Loddenkemper, T. Developing a deep canonical correlation-based technique for seizure prediction. *Expert Syst. Appl.* **2023**, *234*, 120986. [CrossRef]
153. Zhang, T.; El Ali, A.; Wang, C.; Hanjalic, A.; Cesar, P. Weakly-supervised learning for fine-grained emotion recognition using physiological signals. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2304–2322. [CrossRef]
154. Demirel, B.U.; Holz, C. Shifting the Paradigm: A Diffeomorphism Between Time Series Data Manifolds for Achieving Shift-Invariancy in Deep Learning. *arXiv* **2025**, arXiv:2502.19921. [CrossRef]
155. Ji, J.; Cao, Y.; Ma, Y.; Yan, J. TITD: Enhancing optimized temporal position encoding with time intervals and temporal decay in irregular time series forecasting. *Appl. Intell.* **2025**, *55*, 415. [CrossRef]
156. Zhao, S.; Ye, Z.; Adhin, B.; Vuori, M.; Laukkanen, J.; FinnGen; Fisch, S. Cardiorenal Interorgan Assessment via a Novel Clustering Method Using Dynamic Time Warping on Electrocardiogram: Model Development and Validation Study. *JMIR Med. Inform.* **2025**, *13*, e73353. [CrossRef]
157. Wang, M.; You, C.; Zhang, W.; Xu, Z.; Liang, Q.; Li, Q. Causal ECGNet: Leveraging causal inference for robust ECG classification in cardiac disorders. *Front. Physiol.* **2025**, *16*, 1543417. [CrossRef] [PubMed]
158. Manchanda, R.; Panchal, S.; Sandiri, R.; Sudhamsu, G.; Mehta, A.; Gupta, R.; Bhowmik, A.; Bukate, B.B. Energy-efficient clustering and routing for IoT-enabled healthcare using adaptive fuzzy logic and hybrid optimization. *Sci. Rep.* **2025**, *15*, 34619. [CrossRef]
159. Jiménez-Guarneros, M.; Fuentes-Pineda, G.; Grande-Barreto, J. Multimodal semi-supervised domain adaptation using cross-modal learning and joint distribution alignment for cross-subject emotion recognition. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2518612. [CrossRef]
160. Ghasemigarjan, R.; Mikaeili, M.; Setarehdan, S.K.; Saboori, A. Enhancing EEG-based sleep staging efficiency with minimal channels through adversarial domain adaptation and active deep learning. *J. Neural Eng.* **2025**, *22*, 046043. [CrossRef]
161. Li, G.; Wu, C.; Liang, Z. Unsupervised Pairwise Learning Optimization Framework for Cross-Corpus EEG-Based Emotion Recognition Based on Prototype Representation. *arXiv* **2025**, arXiv:2508.11663.
162. Edder, A.; Ben-Bouazza, F.E.; Tafala, I.; Manchadi, O.; Jioudi, B. Self Attention-Driven ECG Denoising: A Transformer-Based Approach for Robust Cardiac Signal Enhancement. *Signals* **2025**, *6*, 26.
163. Yu, J.; Ru, Y.; Lei, B.; Chen, H. GBV-Net: Hierarchical Fusion of Facial Expressions and Physiological Signals for Multimodal Emotion Recognition. *Sensors* **2025**, *25*, 6397. [CrossRef] [PubMed]
164. Fang, X.; Jin, J.; Wang, H.; Liu, C.; Cai, J.; Nie, G.; Li, J.; Li, H.; Hong, S. PPGFlowECG: Latent Rectified Flow with Cross-Modal Encoding for PPG-Guided ECG Generation and Cardiovascular Disease Detection. *arXiv* **2025**, arXiv:2509.19774.
165. Sethi, S.; Chen, D.; Statchen, T.; Burkhart, M.C.; Bhandari, N.; Ramadan, B.; Beaulieu-Jones, B. ProtoECGNet: Case-Based Interpretable Deep Learning for Multi-Label ECG Classification with Contrastive Learning. *arXiv* **2025**, arXiv:2504.08713.
166. Wang, Y.; Hu, S.; Liu, J.; Wang, A.; Zhou, G.; Yang, C. PULSE: A personalized physiological signal analysis framework via unsupervised domain adaptation and self-adaptive learning. *Expert Syst. Appl.* **2025**, *278*, 127317. [CrossRef]
167. Srivastava, S.; Kumar, D.; Jiwari, R.; Seth, S.; Sharma, D. rECGnition_v2. 0: Self-Attentive Canonical Fusion of ECG and Patient Data using deep learning for effective Cardiac Diagnostics. *arXiv* **2025**, arXiv:2502.16255.
168. Zheng, Y. Fusing Cross-Domain Knowledge from Multimodal Data to Solve Problems in the Physical World. *Acm Trans. Intell. Syst. Technol.* **2025**, *17*, 1–13. [CrossRef]
169. Berwal, D.; Vandana, C.; Dewan, S.; Jiji, C.; Baghini, M.S. Motion artifact removal in ambulatory ECG signal for heart rate variability analysis. *IEEE Sens. J.* **2019**, *19*, 12432–12442. [CrossRef]
170. Bari, D.; Aldosky, H.; Tronstad, C.; Kalvøy, H.; Martinsen, Ø. Electrodermal responses to discrete stimuli measured by skin conductance, skin potential, and skin susceptance. *Ski. Res. Technol.* **2018**, *24*, 108–116. [CrossRef]

171. Fan, Y.; Liang, J.; Cao, X.; Pang, L.; Zhang, J. Effects of noise exposure and mental workload on physiological responses during task execution. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12434. [CrossRef]
172. Scarciglia, A.; Catrambone, V.; Bonanno, C.; Valenza, G. Physiological noise: Definition, estimation, and characterization in complex biomedical signals. *IEEE Trans. Biomed. Eng.* **2023**, *71*, 45–55. [CrossRef] [PubMed]
173. Venton, J.; Harris, P.M.; Sundar, A.; Smith, N.A.; Aston, P.J. Robustness of convolutional neural networks to physiological electrocardiogram noise. *Philos. Trans. R. Soc.* **2021**, *379*, 20200262. [CrossRef]
174. de Jong, I.P.; Sburlea, A.I.; Valdenegro-Toro, M. Uncertainty Quantification in Machine Learning for Biosignal Applications–A Review. *arXiv* **2023**, arXiv:2312.09454.
175. Frachi, Y.; Takahashi, T.; Wang, F.; Barthet, M. Design of emotion-driven game interaction using biosignals. In *Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 160–179.
176. Parreira, J.D.; Chalumuri, Y.R.; Mousavi, A.S.; Modak, M.; Zhou, Y.; Sanchez-Perez, J.A.; Gazi, A.H.; Harrison, A.B.; Inan, O.T.; Hahn, J.O. A proof-of-concept investigation of multi-modal physiological signal responses to acute mental stress. *Biomed. Signal Process. Control* **2023**, *85*, 105001. [CrossRef]
177. Mühl, C.; Jeunet, C.; Lotte, F. EEG-based workload estimation across affective contexts. *Front. Neurosci.* **2014**, *8*, 114. [CrossRef]
178. Niu, L.; Chen, C.; Liu, H.; Zhou, S.; Shu, M. A deep-learning approach to ECG classification based on adversarial domain adaptation. *Healthcare* **2020**, *8*, 437. [CrossRef]
179. O'Shea, R.; Katti, P.; Rajendran, B. Baseline drift tolerant signal encoding for ECG classification with deep learning. In *Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 15–19 July 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5.
180. Gholamiangonabadi, D.; Kiselov, N.; Grolinger, K. Deep Neural Networks for Human Activity Recognition With Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection. *IEEE Access* **2020**, *8*, 133982–133994. [CrossRef]
181. Han, J.; Wei, X.; Faisal, A.A. EEG decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks. *J. Neural Eng.* **2023**, *20*, 066027. [CrossRef] [PubMed]
182. Dissanayake, T.; Fernando, T.; Denman, S.; Ghaemmaghami, H.; Sridharan, S.; Fookes, C. Domain Generalization in Biosignal Classification. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 1978–1989. [CrossRef]
183. Pup, F.D.; Atzori, M. Applications of Self-Supervised Learning to Biomedical Signals: A Survey. *IEEE Access* **2023**, *11*, 144180–144203. [CrossRef]
184. Liu, Y.; Du, S.; Han, H.; Chen, X.; Zeng, W.; Tian, Z. Adaptive Distraction Recognition via Soft Prototype Learning and Probabilistic Label Alignment. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 18701–18713. [CrossRef]
185. Zhang, T.; Ali, A.E.; Hanjalic, A.; Cesar, P. Few-Shot Learning for Fine-Grained Emotion Recognition Using Physiological Signals. *IEEE Trans. Multimed.* **2023**, *25*, 3773–3787. [CrossRef]
186. Jeong, H.; Son, J.; Kim, H.; Kang, K. Defensive Adversarial Training for Enhancing Robustness of ECG based User Identification. In *Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 3362–3369. [CrossRef]
187. Luganga, A. Emofusion: Toward Emotion-Driven Adaptive Computational Design Workflows. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences, Niterói, RJ, Brazil, 3–6 June 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 473–478. [CrossRef]
188. Rim, B.; Sung, N.J.; Min, S.; Hong, M. Deep learning in physiological signal data: A survey. *Sensors* **2020**, *20*, 969. [CrossRef] [PubMed]
189. Nayak, S.K.; Pradhan, B.; Mohanty, B.; Sivaraman, J.; Ray, S.S.; Wawrzyniak, J.; Jarzębski, M.; Pal, K. A review of methods and applications for a heart rate variability analysis. *Algorithms* **2023**, *16*, 433. [CrossRef]
190. Orguc, S.; Khurana, H.S.; Stankovic, K.M.; Leel, H.; Chandrakasan, A. EMG-based Real Time Facial Gesture Recognition for Stress Monitoring. In *Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 8–21 July 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 2651–2654. [CrossRef]
191. Torres-Valencia, C.; Álvarez López, M.; Orozco-Gutiérrez, Á. SVM-based feature selection methods for emotion recognition from multimodal data. *J. Multimodal User Interfaces* **2017**, *11*, 9–23. [CrossRef]
192. Hao, T.; Xu, K.; Zheng, X.; Li, J.; Chen, S.; Nie, W. Towards mental load assessment for high-risk works driven by psychophysiological data: Combining a 1D-CNN model with random forest feature selection. *Biomed. Signal Process. Control* **2024**, *96*, 106615. [CrossRef]
193. Gunawan, M.D.; Setiawan, R.; Hikmah, N.F. Estimation of Sleep Quality Based on HRV, EMG, and EEG Parameters with K-Nearest Neighbor Method. In *Proceedings of the 2024 International Seminar on Intelligent Technology and Its Applications (ISITIA), Kuala Lumpur, Malaysia, 20–22 July 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 651–656. [CrossRef]
194. Patil, M.S.; Patil, H.D. Logistic regression based model for pain intensity level detection from biomedical signal. *Int. Res. J. Multidiscip. Scope* **2024**, *5*, 652–662. [CrossRef]

195. Dutsinma, L.I.F.; Temdee, P. VARK learning style classification using decision tree with physiological signals. *Wirel. Pers. Commun.* **2020**, *115*, 2875–2896. [CrossRef]
196. Rivas, J.J.; Orihuela-Espina, F.; Sucar, L.E. Recognition of Affective States in Virtual Rehabilitation using Late Fusion with Semi-Naive Bayesian Classifier. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, Trento, Italy, 20–23 May 2019*; PervasiveHealth'19; Association for Computing Machinery: New York, NY, USA, 2019; pp. 308–313. [CrossRef]
197. Kyamakya, K.; Al-Machot, F.; Haj Mosa, A.; Bouchachia, H.; Chedjou, J.C.; Bagula, A. Emotion and stress recognition related sensors and machine learning technologies. *Sensors* **2021**, *21*, 2273. [CrossRef] [PubMed]
198. Naidu, G.; Zuva, T.; Sibanda, E.M. A review of evaluation metrics in machine learning algorithms. In *Proceedings of the Computer Science On-Line Conference, Virtual, 3–5 April 2023*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 15–25.
199. Lin, Z.; Wang, Y.; Zhou, Y.; Du, F.; Yang, Y. Ste-mamba: Automated multimodal depression detection through emotional analysis and spatio-temporal information ensemble. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
200. Tian, C.; Ma, Y.; Cammon, J.; Fang, F.; Zhang, Y.; Meng, M. Dual-encoder VAE-GAN with spatiotemporal features for emotional EEG data augmentation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 2018–2027. [CrossRef] [PubMed]
201. Thant, A.M.; Panitanarak, T. Emotion Recognition Through Advanced Signal Fusion and Kolmogorov-Arnold Networks. *IEEE Access* **2025**, *13*, 93259–93270. [CrossRef]
202. Jain, P.; Kar, P. Non-convex optimization for machine learning. *Found. Trends® Mach. Learn.* **2017**, *10*, 142–363. [CrossRef]
203. Ramachandram, D.; Taylor, G.W. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
204. Fan, Y.; Xu, W.; Wang, H.; Wang, J.; Guo, S. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 20029–20038.
205. Hatipoglu Yilmaz, B.; Kose, C.; Yilmaz, C.M. A novel multimodal EEG-image fusion approach for emotion recognition: introducing a multimodal KMED dataset. *Neural Comput. Appl.* **2025**, *37*, 5187–5202. [CrossRef]
206. Li, C.; Xie, L.; Wang, X.; Pan, H.; Wang, Z. A disentanglement mamba network with a temporally slack reconstruction mechanism for multimodal continuous emotion recognition. *Multimed. Syst.* **2025**, *31*, 169. [CrossRef]
207. Wang, X.; Li, C.Z.; Sun, Z.; Xu, Y. Design and Analysis of a Closed-Loop Emotion Regulation System Based on Multimodal Affective Computing and Emotional Markov Chain. *IEEE Trans. Syst. Man, Cybern. Syst.* **2025**, *55*, 2426–2437. [CrossRef]
208. Fang, L.; Chai, B.; Xu, Y.; Wang, S.J. KANFeel: A Novel Kolmogorov-Arnold Network-Based Multimodal Emotion Recognition Framework. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Yokohama Japan, 26 April–1 May 2025*; CHI EA '25; Association for Computing Machinery: New York, NY, USA, 2025. [CrossRef]
209. Moorthy, S.; Moon, Y.K. Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion. *Mathematics* **2025**, *13*, 1100. [CrossRef]
210. Erdem Güler, S.; Patlar Akbulut, F. Multimodal Emotion Recognition: Emotion Classification Through the Integration of EEG and Facial Expressions. *IEEE Access* **2025**, *13*, 24587–24603. [CrossRef]
211. Pavan, K.; Singh, A.; Pawar, D.S.; Ganapathy, N. Multimodal Wearable-Based Automated Driver Inattention State Assessment Using Multidevices and Novel Cross-Modal Attention Framework. *IEEE Sens. Lett.* **2025**, *9*, 1–4. [CrossRef]
212. Can, Y.S.; Benouis, M.; Mahesh, B.; André, E. Application of Multimodal Self-Supervised Architectures for Daily Life Affect Recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 2454–2465. [CrossRef]
213. Thaduri, V.R.; R, R.; Rafi, M.; Fernandez, F.M.H.; Lakumarapu, S. Integrating Graph Neural Networks and Temporal Graph Convolutions for Enhanced Multimodal Stress Detection in Physiological and Behavioral Data Streams. In *Proceedings of the 2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare (64220), Vellore, India, 24–27 July 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [CrossRef]
214. Niu, Y.; Chen, X.; Fan, J.; Liu, C.; Fang, M.; Liu, Z.; Meng, X.; Liu, Y.; Lu, L.; Fan, H. Explainable machine learning model based on EEG, ECG, and clinical features for predicting neurological outcomes in cardiac arrest patient. *Sci. Rep.* **2025**, *15*, 11498. [CrossRef]
215. Khuntia, S.; Amjad, A.; Tarekegen, R.B.; Tai, L.C. Deep Learning-Based Emotion Recognition Using Fusion of Multimodal Affective Data From Consumer-Grade Wearable ECG and Speech Sensors. In *Proceedings of the 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–14 January 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
216. Feng, G.; Manimurugan, S.; Yi, B.; Feng, Y. Towards Precision Cardiac Healthcare: Deep Learning and IoT Integration for Real-Time Monitoring and Personalized Diagnosis. *IEEE Internet Things J.* **2025**, *early access*. [CrossRef]
217. Wen, Y.; Chen, W. A Multi-Modal Emotion Recognition Method Considering the Contribution and Redundancy of Channels and the Correlation and Heterogeneity of Modalities. *Measurement* **2025**, *258*, 119247. [CrossRef]
218. Tryon, J.; Guillermo Colli Alfaro, J.; Luisa Trejos, A. Effects of Image Normalization on CNN-Based EEG–EMG Fusion. *IEEE Sens. J.* **2025**, *25*, 20894–20906. [CrossRef]

219. Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.P.; Ebrahimi, T.; Lalanne, D.; Schuller, B. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognit. Lett.* **2015**, *66*, 22–30. [CrossRef]
220. Zang, Z.; Yu, X.; Fu, B.; Liu, Y.; Ge, S.S. Contrastive reinforced transfer learning for EEG-based emotion recognition with consideration of individual differences. *Biomed. Signal Process. Control* **2025**, *106*, 107622. [CrossRef]
221. Cañellas, M.L.; Casado, C.Á.; Nguyen, L.; López, M.B. A self-supervised multimodal framework for 1D physiological data fusion in remote health monitoring. *Inf. Fusion* **2025**, *124*, 103397. [CrossRef]
222. Houssein, E.H.; Mohsen, S.; Emam, M.M.; Abdel Samee, N.; Alkanhel, R.I.; Younis, E.M. Leveraging explainable artificial intelligence for emotional label prediction through health sensor monitoring. *Clust. Comput.* **2025**, *28*, 86. [CrossRef]
223. Gutiérrez-Martín, L.; López-Ongil, C.; Miranda-Calero, J.A. DeepBindi: An End-to-End Fear Detection System Optimized for Extreme-Edge Deployment. *IEEE J. Biomed. Health Inform.* **2025**, *30*, 688–699. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.