Artificial Intelligence Review (2025) 58:131
https://doi.org/10.1007/s10462-025-11126-9

A review on EEG-based multimodal learning for emotion recognition

Rajasekhar Pillalamarri[1] · Udhayakumar Shanmugam[1]

Accepted: 30 January 2025 / Published online: 14 February 2025

Abstract

Emotion recognition from electroencephalography (EEG) signals is crucial for human–computer interaction yet poses significant challenges. While various techniques exist for detecting emotions through EEG signals, contemporary studies have explored the combination of EEG signals with other modalities. However, the field is still rapidly evolving, and new advancements are constantly being made. Comprehensive research is essential by distilling all factors in one manuscript to stay up-to-date with the latest research findings. This review offers an overview of multimodal leaning in EEG-based emotion recognition and discusses current literature in this domain from 2017 to 2024. Three primary challenges addressed are the fusion algorithm, representation of different modalities, and classification scheme. The review thoroughly explores the challenges of fusion algorithms, representation of different modalities, and classification schemes through empirical studies, offering a detailed analysis of their effectiveness. The approach of fusion algorithms is compared and evaluated based on convention and deep learning fusion methods. The research results show that poor performance is attributed to a lack of rigor and inadequate methods to identify correlated patterns across modalities to create a unified representation for experiments. This indicates a need for more thorough analysis and integration of data in future studies. When more than two modalities are involved, it becomes increasingly important to consider different aspects of classification schemes, such as the number of features and model selection. However, designing a classification scheme without considering the number of parameters and emotional categories may compromise the accuracy of classification. To aid readers in understanding the findings better, the results of different classification schemes and their corresponding accuracies are summarized. The tables in this draft display the fusion algorithms researchers utilize and evaluate the effectiveness of selected modalities, providing valuable insights for decision-making. Key contributions include a systematic survey of EEG features, an exploration of EEG integration with behavioral modalities, an investigation of fusion methods, and an overview of key challenges and future research directions in implementing multimodal emotion recognition systems.

Keywords Emotion recognition · Electroencephalography · EEG · Multimodal · Fusion

Extended author information available on the last page of the article

1 Introduction

Human beings undergo emotional reactions in response to various events or situations. The specific emotion a person feels is influenced by the particular circumstance that serves as the trigger. For example, an individual might feel excitement upon achieving a personal goal, while they may experience sadness in the face of a significant loss. Emotions, therefore, reflect our subjective responses to life's diverse experiences. Understanding human emotions is of paramount importance when it comes to ensuring effective communication and interaction among individuals as well as between individuals and machines. (Dwijayanti et al. 2022). Therefore, the recognition of emotions has been a focal point of extensive research in numerous studies including psychology, neuroscience, and computer science. (Zeng et al. 2021). Understanding the role of socio-affective inferential mechanisms in identifying social emotions is essential, given the crucial influence of social context on emotional exchanges. (Mumenthaler et al. 2020). Computers have the potential to respond to people in a manner that is emotionally suitable by employing the technique of emotion recognition (Loveys et al. 2022). Furthermore, understanding how socio-affective cues shape emotional recognition is essential for optimizing computer-mediated interactions. Moreover, through emotion recognition, users can receive feedback regarding their emotional state. This can be highly advantageous for users as it enables them to be more aware of their emotions and the impact they have on their behavior, ultimately aiding in their personal development and self-improvement (Bahreini et al. 2016).

Emotion recognition (ER) is important in various fields, including healthcare, retail and affective computing. It allows for identifying and evaluating emotional states, which can be helpful in medical diagnosis, brain-computer interfaces, and assisting individuals with disabilities. Typically, emotion recognition often relies on data gathered from facial expressions and auditory cues, and other non-verbal signals such as speech, behavior, and biosignals detected by respiration and blood volume pulse (Giannakakis et al. 2019). Traditional sensor data collection and labeling methods often involve exaggerated or intense emotions, which do not reflect real-world scenarios. Therefore, there is a need for more realistic and subtle emotional data to model emotionally intelligent machines. Realistic scenarios, such as natural conversations observed in the real world, are crucial because they capture the complexity and nuances of true emotional expressions. In contrast, lab-generated data often lacks this richness, leading to poor generalizability when applied outside controlled settings. For example, emotions expressed during a scripted lab conversation may not fully represent the variability and spontaneity of emotions seen in everyday interactions. Studies have shown that multimodal approaches, combining facial video data, speech, and physiological signals can improve the performance of ER systems (Sebe et al. 2005). In contemporary research, multimodal learning methods, known for their ability to integrate information from various sources, have emerged to capitalize on several benefits: (i) they maximize the correlation between two or more different modalities, (ii) they assist machine learning algorithms in understanding and training with multiple forms of input data (ii) they prevent redundant or repetitive information fusion when merging different modalities (iv) they are essential for robust processing, particularly when one modality becomes unreliable. While several papers have reviewed unimodal approaches (Yadav et al. 2022), our research aims to bridge the gap by offering a comprehensive review of fusion methodologies and classification techniques, thereby offering a more holistic understanding of multimodal research and insights into open challenges and research guidelines in this domain. Our research scope, however, is limited to studies involving EEG data within these multimodal frameworks. The scope of this draft is shown in Fig. 1.

Affective computing strives to develop machines (or computers) capable of comprehending and responding to human emotions. In reality, computers may find it challenging to recognize emotions because the internal nature of emotions renders them inaccessible to external machines (van den Broek et al. 2010). A standard process pipeline (Broek 2011) helps overcome this limitation (i) by gathering measurable information about a particular emotional state (indicators) (ii) by adding additional information i.e. labels to the gathered data which provide subjective annotations of emotions (annotations), and (iii) by analyzing the relationship between indicators and annotations (modelling). The goal of the modelling step is to use the indicators and annotations to predict or determine the user's emotional state. Essentially, the model can learn patterns or correlations in the data to predict true emotional states.

Various methods are employed to gather indicators, such as emotional words, facial expressions, physiological signs and voice (Vairamani 2024). When it comes to annotations, researchers choose between a discrete emotion model and a continuous emotion model, depending on whether the annotation scale is discrete or continuous. In the past, emotions were often sorted into distinct categories (Plutchik 1982; Ekman 1992), making it challenging to capture varying levels of emotional intensity (Gatti et al. 2018) and the changing nature of emotional experiences over time.

Moreover, emotions are best understood as continuous phenomena rather than discrete entities, and labels alone cannot accurately convey the intensity or strength of emotion (Soleymani et al. 2011). However, the use of continuous annotation based on dimensional models, like the circumplex model (Russell 2003; Posner et al. 2005), is now favored. Russell's Circumplex model is called the Valence-Arousal Spatial (VAS) model. The VAS

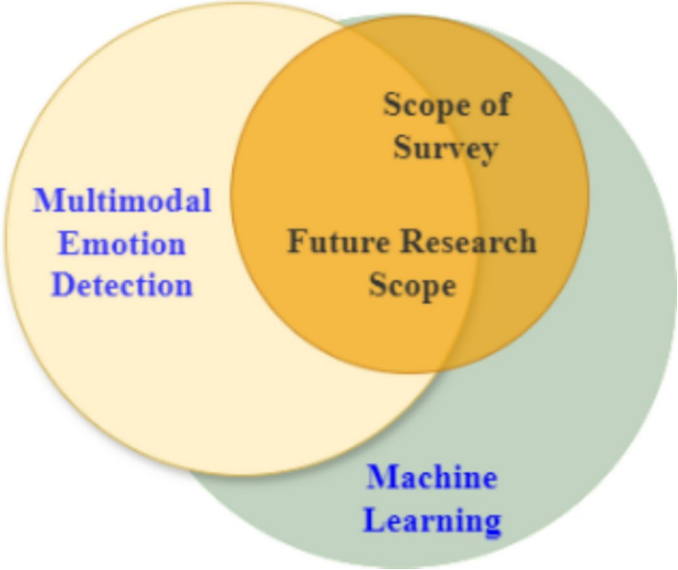


Fig. 1 Survey scope and future research direction

model maps affective concepts to a spatial point. Varieties of affective states occupy different points in the two-dimensional space and are distinguished by their spatial distance. For example, emotions like happiness (with high valence and arousal) and sadness (with low valence and arousal) can be distinguished based on their spatial distance within the multi-dimensional space. The valance scale ranges from negative or unpleasant to pleasant, whereas the arousal scale ranges from active or excited to inactive (Soleymani et al. 2011). The limitation of this approach is its inability to discriminate identical emotions. For instance, anger and disgust are both negative-aroused emotions, making them challenging to differentiate using only a two-dimensional approach. Mehrabian (Bakker et al. 2014) expanded the emotion model to 3-dimensions to address this issue. The additional dimension, represented by the dominance axis, reflects an individual's level of control over an emotion. Current research in emotion identification emphasizes dimensional representation (2D), which is more congruent with the results obtained from developmental studies of affect. This approach allows for a more precise definition of intensity and considers the dynamic and evolving nature of emotional experiences.

Lastly, various frameworks are used for detecting emotions by considering multiple features, which have been suggested and studied in past investigations. However, the accurateness of emotion prediction can be influenced by various factors, such as choosing effective methods for measuring, signal analysis, extracting features, designing experiments, how participants perceive stimuli, and, majorly, the algorithms used for modelling. Many authors frequently report these factors, which have significant implications in numerous studies. Measuring emotions and preparing input data is challenging and still an ongoing area of research in many studies. Experiments can happen in different settings, and two usual ways are in labs and real-life situations. It is challenging to thoroughly study the impacts of emotions by purposely making people feel specific ways. People often feel and react based on the emotions of those nearby, showing that our responses are affected by others' feelings. Recently, as far as the automatic tracking of emotions is concerned, various modeling algorithms have been developed, transitioning from simple shallow methods to more advanced deep neural networks (Liu et al. 2016).

The cutting-edge research on automatic emotion recognition (AER) systems has mainly focused on utilizing combined audio (speech) and visual (faces) signs from video recordings. Besides investigating the distinguishing cues (features), the studies also use several fusion strategies for combining multimodal information to improve emotion detector performance. Beyond audio-visual signs, combined physiological-physical responses improve multimodal emotion detector performance (Yang et al. 2023). We describe a multimodal emotion classification (MEC) system pipeline, shown in Fig. 2. The pipeline consists of four main components: multi-sensor-generated data, feature extraction, modality fusion network, and decision-making system, which classify emotions into various categories. The fundamental procedure is composed of the below four steps:

Step 1:*Identifying relevant modalities*

The complementary data sources that can enrich a comprehensive understanding of human emotions need to be identified and systematically integrated to maximize the effectiveness of multimodal emotion recognition approaches.

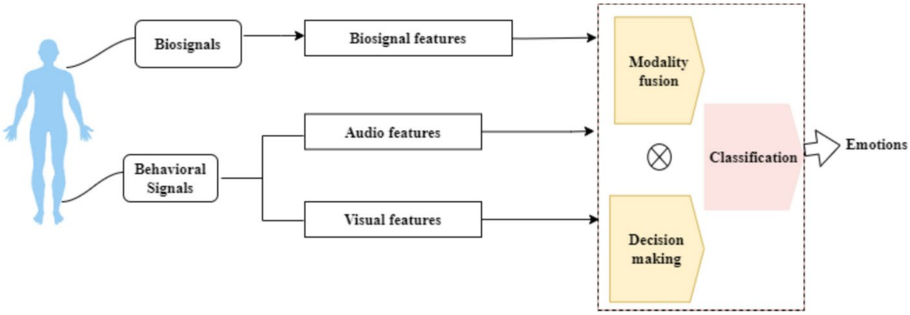


Fig. 2 A general pipeline for a multimodal emotion classification (MEC) system

Step 2:*Feature extraction*

EEG signals provide brainwave patterns, audio yields pitch and tone, video captures facial expressions and facial thermal images, and physiological data reflects bodily responses. This step showcases the extraction of distinct features from each modality.

Step 3:*Modality fusion*

This step involves visualizing and designing strategies for integrating multiple modalities to develop a model that effectively accommodates diverse input data and achieves seamless fusion for enhanced performance.

Step 4:*Decision making*

A refined approach to emotion recognition is fostered by utilizing multimodal inputs, including internal and external behavior paths, coupled with modelling algorithms such as machine learning techniques.

1.1 Biosignal feedback

Multiple data sources possess an incredible ability to convey a broad category of signs related to an individual's feelings, e.g., workload stress, sleepiness, and emotions, which exist within one's mind and are mental states (Eysenck et al. 1994). However, such feelings are not just at the psychical level; they also impact our organisms at physiological level (Yakovyna et al. 2021). For instance, strong emotions trigger various bodily reactions; researchers carefully tap into the bio signals to uncover these signs and better understand or decipher emotions. Researchers today primarily acknowledge two bio-signal measures: one involves the using peripheral physiological signals (PPS). At the same time the other employs brain electrical activity measures captured from the central nervous system (van den Broek et al. 2010). For emotion recognition, both are equally valuable and complementary methods.

Biosignal feedback (physiological measure) offers a precise means to monitor an individual's emotional state (Kim et al. 2005). Further, there is a growing interest in using biosignals, as opposed to facial or vocal expressions, to detect affect (Picard 1997). The similarity of physiological responses to emotions across different societies, as well as cultures, supports the efficacy of this approach (Park et al. 2013). Moreover, biosignal feedback provides a well-defined emotional state that is essentially undistorted. The key advantage lies in their ability to sidestep conscious influences, such as social masking. Modifications in physiological signals linked to emotional states are reflexive and involuntary, and individuals are generally unaware of these changes (Domínguez-Jiménez et al. 2020). Additionally, in cases where analyzing emotional individuals with verbal disabilities, relying on physical signals like voice may not always be feasible (Bakhshi and Chalup 2021). For instance, it can be challenging in speech emotion recognition to differentiate similar emotions like frustration and anger due to overlapping vocal characteristics or the intentional masking of true emotions in social contexts. Unlike these cues, EEG wave patterns reflect underlying neural activity and provide a more direct and involuntary representation of emotional states. This minimizes ambiguity and allows for more precise labeling of emotional states, ultimately improving the data quality used for model training. Currently, there is a trend favoring the adoption of physiological measures, particularly electroencephalography (EEG) based brain wave signals, to quantify inner emotional states. As a flexible physiological measure, EEG holds advantages in emotion recognition by clearly indicating brain activity. Also, emotion detectors based on EEG have demonstrated exemplary performance (Yang et al. 2018).

Combining brain wave data with details about emotional characteristics in physical (or expressive) responses is a compelling avenue of exploration. The integration of physiological and expressive signals strengthens the combined information, improves upon the limitations of each modality, and creates a more robust system. However, in the utilization of multimodal signals, a pivotal concern revolves around the methodology of integrating these signals, representing a field of growing interest within the research community. Nevertheless, integrating EEG data with various behavioral modalities has not been extensively explored across many multimodal studies.

1.2 Related work

Several studies, including those by Yu and Wang (Yu and Wang 2022) and Craik et al. (Craik et al. 2019), have explored methods to achieve high accuracy in emotion recognition. However, these reviews did not adopt multimodal approaches, leaving room for further exploration and analysis. Samal et al. (Samal and Hashmi 2024) conducted a review on EEG-based emotion recognition, encompassing multimodal analysis. However, the review did not delve into EEG fusion with various modalities. Ahmed et al. (Ahmed et al. 2023) presented a detailed review of multimodal emotion detection using learning algorithms, which also presented available datasets and data acquisition tools, deep learning, and conventional classifiers and their application areas.

Baltrušaitis et al. (Baltrušaitis et al. 2018) presented a survey on various fusion techniques and challenges of multimodal learning, such as co-learning; mainly discussed modalities are text, image, and audio. However, the review does not provide a detailed comparison of the various multimodal fusion patterns and does not evaluate their performance on benchmark datasets. Nonetheless, the review (Baltrušaitis et al. 2018) prompts the researcher's core challenges in multimodal representations using machine learning and can be utilized as a valuable resource by researchers in this field.

Bota et al. (Bota et al. 2019) reviewed the existing studies on AER systems using machine learning techniques to analyze physiological signals. The authors' main focus is to discuss several biosignals, such as ECG, GSR, PPG, EEG, etc., used for emotion recognition. Even though the survey highlights the challenges and limitations of AER using physiological modalities, it does not provide a comprehensive analysis of the approaches and techniques used to address these challenges. While the report provides some recommendations for future research, a more detailed discussion of specific methods and techniques for addressing the limitations of physiological signal-based AER would have been beneficial.

He et al. (He et al. 2020) surveyed the use of multiple modalities, such as EEG and other brain imaging signals, in combination with behavioral measures to enhance the accuracy of affect-based brain-computer interface (aBCI). Additionally, they presented multimodal fusion methods connected to behavior, brain signals, and hybrid neurophysiology modalities. On the other hand, it is also limited because it does not focus on state-of-the-art models across various modalities. For instance, speech modality is not covered.

Zhang et al. (Zhang et al. 2020) discuss the convergence of machine learning and information fusion for endowing machines with emotion recognition abilities. Various EEG-based methods are reviewed, covering feature extraction, reduction, and classifiers. Correlations between EEG rhythms and emotions, as well as brain region-emotion associations, are also explored. The review (Zhang et al. 2020) concludes with a comparison of machine learning (ML) and deep learning (DL) algorithms and highlights open issues in this rapidly evolving field. The authors reviewed various fusion techniques in general, whereas we reviewed different fusion techniques and multimodal databases comprehensively. In summary, the aim of our survey shares similarities with the related works (Zhang et al. 2020; He et al. 2020), which explores recent progress in fusion techniques and multimodal contribution with a specific emphasis on EEG features. Table 1 compares these related surveys on emotion recognition.

Recognizing the gaps in existing surveys, this paper undertakes a meticulous examination of the challenges associated with current EEG-based multimodal methods. The survey, which encompasses all relevant articles published from 2017 to 2024 and sourced from reputable databases including PubMed, MDPI, Science Direct, Springer Link, IEEE Xplore, and Google Scholar, primarily focuses on the fusion of EEG signals, behavioral, and other biosignals. The main contributions of this survey can be summarized below:

(i) Presentation of a systematic survey on EEG feature extraction methods,
(ii) Exploration of the fusion of EEG signals with diverse modalities (speech, facial, other biosignals),
(iii) Exploration of conventional and deep learning fusion techniques, and
(iv) Providing insights into this domain's open challenges and research guidelines..

2 Methods

To categorize studies and streamline the selection of pertinent papers for this research, a systematic review process called PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) was utilized. The PRISMA process consists of four key sequential steps: paper identification, the elimination of duplicates through scanning, eligibility filter-

Table 1 Comparison of previous and related surveys and reviews

References	Multi-modal analysis	Discussion on EEG fusion with		Fusion methods/operations	EEG feature extraction	Dataset discussion	Multi-modal use cases	Ad-dressing through RQs	Approach to survey/review
		Behavior signals	Bio-signals						
Yu and Wang (2022)	✗	✗	✗	✗	✓	✗	✗	✗	EEG-based emotion detection
Craik et al. (2019)	✗	✗	✗	✗	✗	✗	✗	✓	EEG-based emotion detection
Ahmed et al. (2023)	✓	✗	✗	✓	✗	✓	✗	✓	Emotion acquisition tools; fusion methods
Samal and Hashmi (2024)	✓	✗	✗	✗	✓	✓	✗	✗	Focus on BCI systems
Baltrušaitis et al. (2018)	✓	✗	✗	✓	✗	✗	✗	✗	Taxonomy of multi-modal machine learning
Bota et al. (2019)	✓	✗	✗	✓	✓	✓	✗	✗	Study of emotions; focus on biosignals
He et al. (2020)	✓	✓	✓	✓	✗	✓	✗	✗	Focus on aBCI system
Zhang et al. (2020)	✓	✗	✗	✓	✓	✓	✗	✗	Study of emotions; focus on EEG and other biosignals
Ours (-)	✓	✓	✓	✓	✓	✓	✓	✓	Study of emotions; focus on EEG fusion with behavior and other biosignals

ing, and the inclusion of papers to construct a final paper selection list. Figure 3 depicts the steps undertaken.

A total of 147 manuscripts were retrieved from various databases, including 72 from IEEE, 26 from Science Direct, and 20 from Springer, with additional articles obtained from MDPI, Frontiers, and PubMed via Google Scholar. After excluding 23 duplicated papers and 18 conference papers, further refinement involved eliminating non-related studies, those unrelated to EEG and articles published before 2017 from the survey. The present research followed the steps as recommended by (Moher et al. 2010).

2.1 Study selection

After identifying the research questions, papers from the selected studies were collected. This section covers the search strategy and the criteria used for study selection. To tailor our literature search and focus this review on specific aspects, we formulated the search query using the following terms: (a) Emotion Recognition, (b) Emotion Detection, (c) Electroencephalography/EEG, (d) physiological signal fusion, and (e) multimodal data fusion.

We conducted a systematic examination of related publications available in major online research repositories, such as IEEE Xplore, Science Direct, SpringerLink, PubMed, and MDPI, to identify the necessary articles. For instance, we searched the IEEE database with an example query: ((“emotion detection”) OR (“emotion recognition”)) AND (“EEG”) AND (“multimodal”).

In evaluating the systematic literature review (SLR), we applied inclusion and exclusion criteria to pinpoint relevant primary research works. Our primary focus was journal papers, with only minimal consideration of a few papers from reputed conferences. The initial query yielded 147 records. After removing duplicates during the identification phase, 124 records were selected for further screening. During the screening phase, we excluded datasets and reviewed articles based on their titles and publication types, reducing the count to 95. A comprehensive full-text review was then conducted, excluding 39 records that did not focus on EEG analysis methods or lacked specific insights on fusion techniques. Ultimately, 45 records were included in the final study selection. We categorized the final selected articles for data analysis based on several criteria. First, we grouped the studies according to the types of modalities used in fusion with EEG data.

Next, we further subdivided these studies based on the fusion methods employed. Finally, we classified the studies based on whether they utilized machine learning or deep learning techniques. This systematic segregation is designed to address the research questions posed in the study effectively. We utilized Zotero as a citation management tool to compile the bibliography. Figure 4 further categorizes the papers according to their respective publication years.

The inclusion criteria were strategically incorporated to streamline our search effectively. These criteria consisted of the following: (1) Inclusion of research articles written in English, (2) Inclusion of articles mentioning fusion or feature fusion either in the title or abstract, (3) Requirement for articles to address aspects of data or sensor fusion, and (4) Inclusion of articles within the scope of our domain. For example, publication titles that referenced multimodal and EEG but did not utilize multiple modalities were excluded from the study (Wu et al. 2022).

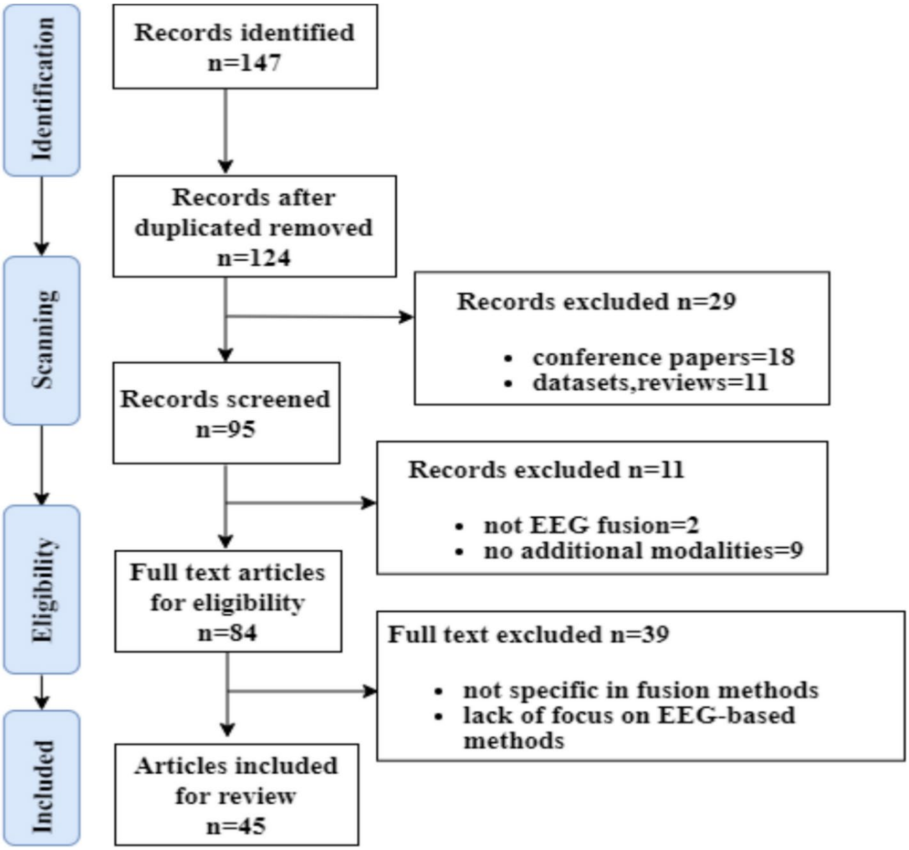


Fig. 3 Flow diagram illustrating the paper selection process based on PRISMA (n- represents the number of papers)

3 Research questions (RQs)

The primary survey is structured to address specific research inquiries (RQs), forming the main content of the survey. We have delineated three relevant research inquiries on integrating EEG data and behavioral modalities. These research questions are:

RQ1 What is the use of joint EEG and behavioral input to multimodal emotion recognizer? For example, joint Speech and EEG data measures can effectively detect behavioral dysfunction in medical-related studies.

RQ2 What fusion networks are currently available for EEG-based emotion recognition?

RQ3 What parameters or features are interesting for accurate emotion recognition?

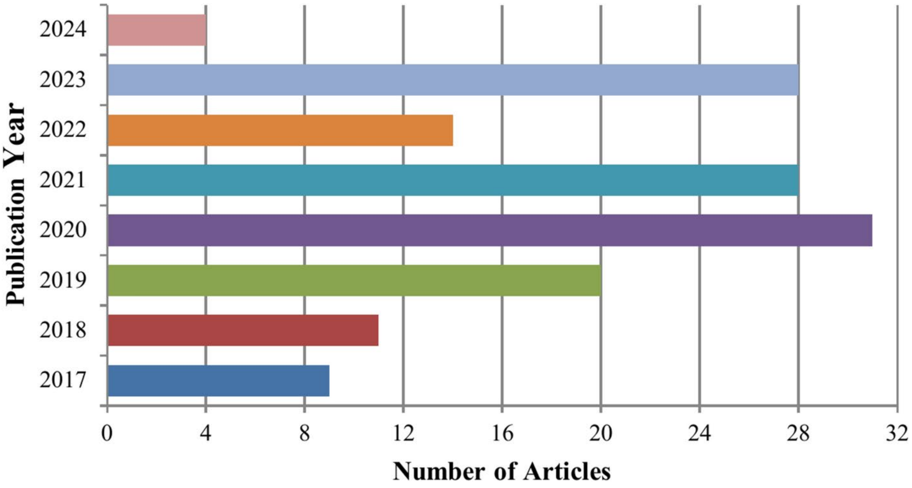


Fig. 4 Number of articles per year

3.1 RQ1: What is the use of joint EEG and behavioral input to multimodal emotion recognizer?

Using different modalities, coupled with EEG signals, significantly advances our understanding of complex cognitive processes and improves efficacy in multimodal emotion detection systems. This research enhances our current understanding and paves the way for future advancements in ER systems. By studying multiple sources of information, we can gain a more comprehensive understanding of how the brain responds to stimuli, ultimately creating a more accurate and effective emotion recognizer. The review structure of RQ1 comprises three primary components: data modalities, use cases, and datasets. Data modalities are further divided into two sub-components, namely biosignals and combinatory modalities. The investigation into use case components relies on combinatory modalities. The subsequent subsections provide detailed elaboration on the taxonomies and their corresponding components.

3.1.1 Affective modalities

The study of data modalities thoroughly examines the relationship or interplay, aiming to discern how diverse forms of data, ranging from physiological signals to individual behavioral expressions, contribute to the broad landscape of human emotions. A thorough investigation of various behavioral modalities, including visual modalities (facial and body gestures), audio modality, and text modality, was presented by Poria et al. (Poria et al. 2017). This sub-section highlighted various biophysical signals.

3.1.1.1 Peripheral physiological signals (PPS) PPS signals refer to the physiological responses of an individual's body that are linked to psychological processes and emotions.

Common psychophysiological measures include galvanic skin response (GSR), photoplethysmogram (PPG), electrocardiogram (ECG), and facial electromyogram (facial-EMG).

The physiological responses arise from two main mechanisms of the nervous system: the Central Nervous System (CNS), encompassing the brain and spinal cord, and the Autonomic Nervous System (ANS), reflecting changes in cardiovascular, electrodermal, and muscular activities. ECG is employed to measure cardiovascular activity, providing parameters such as heart rate (HR) and heart rate variability (HRV). Electrodermal activity (EDA) or GSR gauges skin's electrical conductivity, offering a reliable response to external stimuli. The blood volume pulse is monitored through PPG (Allen 2007), a device in contact with the skin that records changes in cardiac electrical potential over time. Additional physiological indicators of emotional changes include electromyography (EMG), which records muscle electrical activity, as well as respiration rate and body temperature. (Bontchev 2016). In a study (Chanel et al. 2007), GSR and EMG demonstrated a high correlation with arousal and valence across all subjects. Irregular respiration and rapid breathing are associated with heightened emotions such as fear.

ECG: Presently, emotion identification systems can be built using a variety of approaches and algorithms. The utilization of ECGs is widespread in many studies and applications, as reviewed in (Hasnul et al. 2021). Emotional experiences can cause changes in the heart rhythm, which can be sensed or measured using ECG readings, referring to specific features from the ECG waveform. Thus making them a helpful tool for designing emotion classifiers (Brás et al. 2018). A recent study (Sepúlveda et al. 2021) achieved good accuracy in classifying emotions from wearable device signals, outperforming previous studies with 95.3% accuracy from the AMIGOS dataset in a two-dimensional classification.

PPG: Today, PPG signals are gaining popularity in various applications due to their ease of use and suitability for wearable devices. PPG signals are derived from the changes in blood volume in peripheral blood vessels, typically measured using sensors placed on the skin surface. The user-friendly aspect implies that obtaining PPG signals does not require invasive procedures, making it effortless and convenient for individuals. Additionally, PPG signals are well-suited for wearable devices, allowing continuous monitoring of physiological parameters, including heart rate and blood oxygen levels, in real-time. Most reported emotion identification methods use PPG signals in a multi-modality approach but can also be used to detect multiple emotional states using PPG alone (Paul et al. 2023). The proposed algorithm shows superior performance, achieving 97.78% detection accuracy when evaluated on PPG signals from the DEAP dataset. The existing literature discusses various approaches to creating emotional recognition systems, but there is insufficient exploration of the performance of PPG signals within these systems (Lee et al. 2019). Examining the performance of PPG signals in emotional recognition systems is crucial, as it provides investigators with the opportunity to enhance the precision and efficiency of methods used for characterizing and classifying emotions based on PPG measures (Goshvarpour and Goshvarpour 2018).

GSR: Another commonly used biofeedback signal for developing automatic emotion classifiers is GSR. The signal known as GSR is produced by the electrical activity of the sweat glands and contains valuable information about the sympathetic nervous system. Unlike sensors that generate signals internally, modulating sensors rely on an external source. An example of such a sensor is the GSR sensor, which is a modulating sensor controlled involuntarily and measures skin conductance, as mentioned in (Ahmed et al. 2023). GSR measure is a remarkably sensitive indicator of emotional arousal (Balconi and Lucchiari 2008). Various studies have employed different GSR-based procedures to identify levels of stress for working professionals (Sriramprakash et al. 2017). In Ref. (Goshvarpour and Goshvarpour 2020), the potential of GSR for emotion classification is examined, and the outcomes demonstrated that the phase space geometry, and consequently, the dynamics of the signal, are influenced by the emotional content of the music video.

EMG: facial-EMG is a biosensor measure utilized to gauge affective empathy by monitoring the activity of facial muscles. This method is based on the premise that individuals tend to mimic the facial expressions of others, providing insights into their emotional experiences (Ferguson and Wimmer 2023). Research using electromyography (EMG) has uncovered that observers tend to mimic expressions of happiness or anger. According to Dimberg et al. 2011, the more empathic individuals exhibit more significant muscle contractions in response to these expressions. This finding suggests that highly empathic individuals may possess increased sensitivity to the emotions of others (Dimberg et al. 2011). Identifying emotions using EMG signals offers the potential for developing a classification system that relies solely on EMG information. The approach presented by (Pereira et al. 2022) provides benefits such as detecting micro-expressions and consistent signal collection and eliminating the necessity of acquiring facial images. This research marks an initial stride toward the automatic classification of emotions based exclusively on facial EMG. Additionally, the investigation of emotion sensing using wearable facial-EMG devices to evaluate emotional valence is discussed in (Gjoreski et al. 2022).

3.1.1.2 Neuroimaging modalities Brain signals provide a rich source of emotional information that complements audio-visual data and PPS signals. Compared to PPS signals, integrating CNS signals enhances the efficacy of user-centric affect recognition by capturing neural activity directly associated with emotional processing (Abadi et al. 2015). Three commonly used central nervous system (CNS) signals capture neural activity in the brain, and their measurement techniques include electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI). This section focuses on two key neuroimaging modalities—fMRI and MEG—each offering unique insights into brain function during emotional experiences (Babiloni et al. 2009). A detailed discussion of EEG will follow in the subsequent section. Biophysical signals associated with brain electrophysiology or hemodynamic and metabolic activity form the foundation of all non-invasive neuroimaging techniques. Among these, EEG and MEG are rooted in capturing neural processes through electrophysiological principles. In contrast, fMRI operates on principles related to hemodynamic and metabolic activity within the brain (Ogawa et al. 1992; Burle et al. 2015). The strengths and limitations of these modalities are primarily influenced by the spatiotemporal properties of the recorded signals relative to neuronal activity, as well as the variety of sensing and imaging techniques employed for each modality analysis (He and Liu 2008).

fMRI: In the study (Reske et al. 2009), fMRI was used as a non-invasive tool to investigate the neural correlates of emotion by capturing real-time brain activity in response to emotional stimuli (Erwin et al. 1992). The main findings revealed comparable behavioral performance between schizophrenia patients and healthy participants in emotion discrimination tasks. These results provide insights into the specific brain regions involved in processing different emotions, helping to understand how emotional processing may be altered in schizophrenia.

MEG: MEG offers a non-invasive method to capture real-time neural activity by measuring magnetic fields generated by neuronal currents (Cohen 1968). Its high temporal resolution makes it ideal for tracking rapid brain dynamics linked to emotional responses. The DECAF dataset leverages MEG to decode human emotional responses by capturing fine-grained neural activity. This approach allows for a detailed relationship analysis between participants' self-assessments and physiological responses. MEG's high temporal resolution provides insights into the timing and sequence of neural processes underlying emotional experiences. It is a valuable tool for exploring emotion-related neural dynamics in the DECAF (Abadi et al. 2015). In Ref. (Kheirkhah et al. 2021), the authors utilized MEG data to classify human emotions, building on insights from the DECAF framework. They demonstrated the feasibility of using MEG to analyze brain responses to emotional stimuli with high accuracy. The study successfully distinguished between pleasant, unpleasant, and neutral emotions by examining neural activity in response to picture stimuli, showcasing MEG's capability in emotion detection tasks.

3.1.2 Brain wave (EEG) signal

EEG remains one of the main utility in clinical neurophysiology and cognitive neuroscience. As per the International Federation of Clinical Neurophysiology (IFCN) Committee, the EEG is characterized as "(1) the science of relating to the electrical activity of the brain", and "(2) the techniques of recording electroencephalograms" (Steriade et al. 1990). The best aspect is that EEG is a real-time display and ongoing indicator of brain activity that offers a superior temporal resolution. (Thakor and Sherman 2012).

3.1.2.1 EEG signal and electrode distribution In 1929, Austrian psychiatrist Hans Berger began the practice of recording human EEG and subsequently published the first research paper on this topic (Berger 1929). Following Berger's work, electrophysiologists and neurophysiologists validated his findings, leading to significant advancements in EEG research within brain science and clinical medicine. Investigating EEG patterns allows for a better assessment of emotional shifts, with neuronal potentials serving as indicators of the functional and physiological changes within the central nervous system. The EEG captures the electrical activity of neuron groups in the specific brain region where the electrode is placed, providing a wealth of psychophysiological information. In medicine, EEG signals serve as an objective tool for diagnosing various diseases through advanced processing, classification, and analysis techniques. Meanwhile, in neuro-engineering, EEG-based technologies enable individuals with disabilities to control assistive devices like wheelchairs or robotic arms using thought or motion imagery, forming the core of the rapidly growing Brain-Computer Interface (BCI) field (Zhang et al. 2020).

Usually, EEG waveforms are recorded using scalp-mounted metal electrodes. During data acquisition, electrodes are attached to the subject's scalp, and the potential differences (pds) are recorded. These waveforms are used to interpret the several aspects of the electrical activity arising from human-brain. There are various methods to obtain EEG; one widely used standard method is called 10–20 system (Klem et al. 1999). The standard 10–20 system includes a total of 21 electrodes placed at specific distances of the skull, which are determined by measuring the distance between anatomical landmarks (nasion, inion, left preauricular and right preauricular points) on the head. The electrodes are labeled with letters indicating the brain regions they are placed over (Fp:frontoplor, F:frontal, T:temporal, O:occipital, P: parietal, and C:central), and numbers indicating the hemisphere they are located in ("odd numbers for the left hemisphere, even numbers for the right hemisphere"). The parietal and frontal electrode pairs are found to be significant in finding appropriate emotional states (Lin et al. 2010). Ref. (Rojas et al. 2017) outlines the 10–20 system (see Fig. 5). The selection of EEG electrode positions is significant for emotion classification. In Ref. (Zhang et al. 2016b) four electrode (Fp1, Fp2, F3, C4) positions were selected, and the best performances were achieved.

3.1.2.2 EEG-signal characteristics and rhythms The EEG signal directly mirrors brain activity and is essential for studying the physiological processes of the human brain. Its key characteristics include:

(a) Noise: EEG recordings are highly sensitive to environmental disturbances, making them noise-prone. Their low amplitude ranges from 50 to 100 µV. Major artifacts, distortions, and other physiological signals such as EOG, ECG, and EMG often contaminate EEG signals.

(b) Non-linear: EEG signals are influenced by other peripheral physiological signals during acquisition, which can impact the recorded potentials. This interaction, along with physiological adaptations in human tissues, results in a highly non-linear nature of the EEG signals.

(c) Nonstationary: EEG signals are sensitive to surrounding environmental factors and can vary unpredictably, signifying a nonstationary nature.

(d) Frequency-domain characteristics: EEG signals are spread across frequency ranges from 0.5 to 100 Hz; however, a lower EEG bandwidth (0.5–30 Hz) is found to be

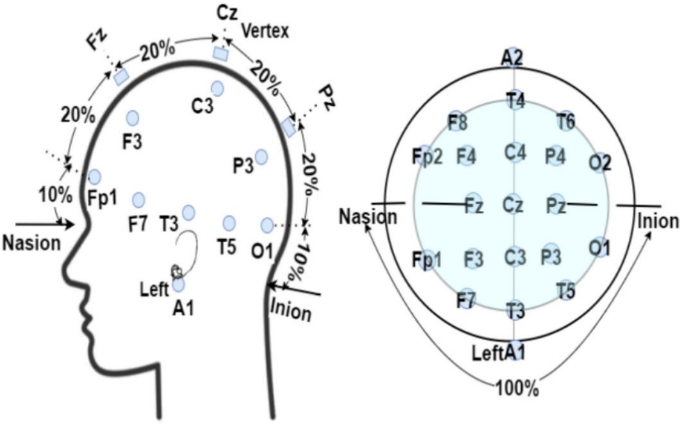


Fig. 5 Electrode placement in the EEG system follows the 10–20 international system (Bromfield et al. 2006)

relevant for recognition Consequently, it is a common approach to classify EEG signals into five frequency bands, each corresponding to a specific kind of cognitive task.

(e) Non-invasive EEG recording: Non-invasive EEG recordings offer a valuable tool for studying brain activity in humans, making them widely usable for both clinical and research purposes (Ball et al. 2009). The key advantage of non-invasive EEG is its ability to deliver fast and cost-effective measurements, which play an important role in emotion computing applications.

An important issue is the selection of appropriate frequency bands: EEG signal is a blend of waveforms and can be classified based on its magnitude, frequency, spatial distribution, reactivity, and wave shape. In particular, the frequency band is a standard classification method in many studies. The frequency band of human EEG can be classified into five categories depending on the frequency range and amplitude characteristics: delta (<4 Hz), theta (3–7 Hz), alpha (8–14 Hz), and gamma (>30 Hz). Each frequency band is related to specific activities or brain states. For example, delta (wave frequency is very low) waves are observed when a subject is unconscious and in deep sleep. In contrast, gamma (very high-frequency wave) waves are associated with intense brain activity (Dadebayev et al. 2022). Table 2 presents a short description of EEG frequency bands. The EEG frequency bands can provide information about brain states and are used in several applications, including clinical diagnosis to find neurological disorders (Newson and Thiagarajan 2019).

3.1.2.3 EEG and brain-network Despite advancements in EEG-based emotion detection, most existing approaches have been limited to single-channel feature extraction analysis and classification. This narrow focus undermines the capability of EEG to capture the complex dynamics of emotional processing. There is a growing interest in moving towards multi-class, multi-channel automatic emotion identification, where EEG data is analyzed from a network perspective to reveal the interconnections between different brain regions (Li et al. 2019c). Moreover, the authors investigated the potential of emotion-related brain networks by utilizing spectral power analysis to capture activation differences across multiple brain regions and constructing networks using Phase Locking Value (PLV). By integrating both activation and connectivity information from EEG data, they introduced a multi-feature fusion approach that significantly enhanced the performance of emotion recognition. The results, validated on three public emotion databases, revealed that the combined features

Table 2 Short description of widely used EEG rhythms and correlation with mental activity (Kawala-Sterniuk et al. 2021)

EEG rhythms (Hz)	Mental activity	Location on brain
Delta (0–4)	Sleep, dreaming-no focus, unconsciousness	Frontal
Theta (4–8)	Imaginary, deep relaxation	Midline, temporal
Low alpha (8–10)	Calm, wakeful relaxation	Frontal, occipital
High alpha (10–12)	Self-awareness	
Low beta (12–18)	Active thinking, problem-solving	Frontal, distributed on both sides
High beta (18–30)	Start to alert	
Low gamma (30–50)	Intelligence, self-control	Frontal, central, somatosensory cortex
High gamma (50–70)	Cognitive tasks: reading and speaking	

outperformed single features based on power distribution or network characteristics, underscoring the effectiveness of combining activation and connectivity patterns for improved emotion classification. This approach emphasizes the value of leveraging brain network connectivity to develop human–computer interaction systems capable of adapting to human emotions in real-world applications. By leveraging multi-channel information, these methods can better account for the brain's functional connectivity patterns, enabling more accurate and comprehensive emotion identification across multiple emotional states. This shift towards multi-channel approaches is essential to advancing the field, as it offers a deeper understanding of the neural correlates of emotions and enhances the overall performance of emotion recognition systems.

In a study (Li et al. 2019a), the authors have placed a strong emphasis on feature extraction from brain networks constructed using EEG signals to improve emotion recognition. Their proposed approach demonstrated superior classification results, highlighting the potential of brain network-based features for distinguishing between different emotions. Unlike standard methods, which focus on single-channel analysis, they introduced a novel feature extraction technique that leverages brain network spatial topologies to enhance emotion differentiation. Although various methods exist for constructing brain networks, this work also utilized PLV to establish undirected brain network structures. The spatial topologies derived from these networks effectively captured distinct patterns in different emotional states, thereby boosting classification performance. Furthermore, the study considered spatial network features across four frequency bands—theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–48 Hz)—to ensure a comprehensive analysis of the brain's functional connectivity about emotion recognition.

In Ref. (Wu et al. 2022), the authors presented a novel algorithm for selecting emotion-relevant critical subnetworks and investigated multiple EEG functional connectivity network features to improve emotion recognition. Additionally, they constructed a multimodal emotion recognition model by integrating functional connectivity features from EEG with features derived from eye movements and physiological signals. This multimodal approach complements the earlier work by providing a more holistic representation of emotional states, combining central and peripheral physiological signals for enhanced classification performance. Together, these studies underscore the potential of using advanced network-based feature extraction methods and multimodal integration to more effectively capture the complex patterns of brain activity associated with different emotions, paving the way for more robust and comprehensive emotion recognition models. In Ref. (Li et al. 2024a), the authors focused on EEG brain networks for emotion recognition. Their approach targeted multicategory classification by utilizing multiple spatial network topology patterns to better capture the unique characteristics of different emotional states. This method demonstrated the effectiveness of employing complex network structures for distinguishing between ranges of emotions, further advancing this field through EEG-based brain network analysis. In Ref. (Li et al. 2024b), the authors proposed a novel method using a spectral graph filtering method called BF-GCN to investigate the applicability of EEG-derived brain graphs further. Additionally, the attention mechanism was incorporated into the framework to capture brain activation patterns, achieving an impressive accuracy of 97.44%.

3.1.3 Use cases

EEG-ECG: Physiological measures like ECG and EEG are widely studied for medical diagnosis and biometric recognition. A multimodal biometric recognition system is proposed (Barra et al. 2017), combining ECG and EEG signals, using fiducial (peak) features from ECG and spectrum features from EEG, with good recognition performances. In a recent investigation, researchers have explored the depth of anesthesia in a clinical setting through the joint measurement of EEG and ECG. This joint measurement approach aims to provide a enhance understanding of the patient's physiological signs, mainly focusing on the level of anesthesia (Bahador et al. 2021).

EEG-SPEECH: Researchers are identifying the close relationship between stress and emotions. Emphasizing that stress is often not considered an emotion in traditional emotion models. Especially long-term stress can lead to cognitive, emotional, and behavioral dysfunctions. The neglect of empathy and stress management in medicine is highlighted in Ref. (Duwenbeck and Kirchner 2024), and the authors aimed to develop AI-Assistant to detect stress for the benefit of various health applications by combining EEG-Speech to enhance both remote and direct healthcare. Nonetheless, a challenge associated with this approach lies in the availability of a comprehensive multi-modal dataset.

EEG-FACIAL: Multimodal emotion recognition, which combines EEG-FACIAL signals, outperforms single-modal emotion recognition (Yang et al. 2022). This enhanced approach is particularly noteworthy in the context of deaf individuals, where using both EEG signals and facial expressions in tandem demonstrates superior efficacy compared to relying on a single modality for emotion recognition. Integrating these modalities contributes to a more comprehensive and accurate emotional states, emphasizing the potential benefits of multimodal approaches in diverse populations (Yang et al. 2022).

EEG-TEXT: Studying customer loyalty involves looking at brain activity, especially in the emerging field of neuromarketing (Kumar et al. 2019). Researchers use tools like EEG and MEG to understand how our brains respond to products (Pei and Li 2021). Recently, there is a growing trend to combine brain tracking (EEG) with sentiment analysis to detect emotions. An innovative system is proposed (Panda et al. 2020) to combine brain signals and customer reviews (text) to improve how we understand and recognize emotions during product experiences. The EEG signals and review text have different structures. To make them compatible, both inputs are transformed into a shared feature space. The encoder of EEG converts features into a 2D vector, while the encoder of text uses a bag-of-words model to extract relevant features, resulting in a 2D vector. These features from both modalities are then combined into a common feature space to classify emotions. This approach unifies diverse data types to enhance emotion classification.

3.1.4 Datasets

Numerous datasets have been made publicly available to facilitate research in the realm of multimodal emotion learning. It is worth noting that most of these datasets are typically either multi-physical or multi-physiological. In contemporary times, the scope of multimodal studies has expanded to encompass the combination of various modality elements (EEG and visual components, for example). Table 3 provides details of these datasets, outlining their modality elements and the number of subjects. Few of them are recent and accessible for public use. While most datasets incorporate at least four modalities, some, like DEAP, include as many as nine distinct modalities.

DEAP (Koelstra et al. 2012) and SEED-IV (Zheng et al. 2019) are popular benchmark datasets for studying emotions using physiological signals. SEED-IV collected EEG and Eye movement data from 44 subjects watching video clips inducing sad, fear, happiness, and neutrality. DEAP consists of peripheral physiological and EEG recordings from 32 subjects viewing one-minute 40-music videos. These datasets provide valuable resources for emotion research in multimodal environments. Both datasets include recordings of physiological signals, particularly EEG data, which captures brain activity and peripheral physiological signals. In addition to EEG data, SEED-IV also includes eye movement recordings, while DEAP incorporates multiple peripheral physiological signals alongside the EEG data.

The DREAMER dataset (Katsigiannis and Ramzan 2018) encompasses multimodal data, incorporating ECG and EEG signals captured during affect elicitation via audio-visual stimuli. This database sets a benchmark for affect recognition by leveraging EEG and ECG-based features, yielding results comparable to those obtained from datasets employing costly medical-grade devices. The AMIGOS (Miranda-Correa et al. 2021) database includes neurophysiological signals, video recordings, and annotations of participants' emotions. It will be publicly available for researchers to evaluate the suitability of low-cost devices for affect recognition applications. It allows for studying affective responses concerning mood, personality, and social context.

The MAHNOB-HCI (Soleymani et al. 2012) is a widely used multimodal dataset for studying affect interfaces in human–computer interaction. It focuses on capturing multimodal data, including physiological signals, audio and visual signals during various interactive scenarios.

ASCERTAIN (Subramanian et al. 2018) is another multimodal database that connects personality traits and emotional states using physiological responses, containing data from 58 users recorded while viewing affective movie clips. This database contributes to the study of human–computer interaction by addressing the need for agents to understand and adapt to users' affective states, considering the influence of personality. With a similar motivation, a recent study (Pant et al. 2023) presented a Physiological database for Multimodal Emotion Recognition (PhyMER) to study emotion through physiological response with personality as a context. The PhyMER dataset includes EEG along with other physiological data (EDA, BVP, and skin temperature) collected from 30 subjects. The researchers conducted experiments to examine emotion elicitation in response to video-based stimuli, assessing inter-rater agreement and the correlation between experienced emotions and personality traits (Pant et al. 2023).

In order to study emotions in social interactions effectively, a novel dataset named K-EmoCon (Park et al. 2020) has been introduced. This multimodal dataset offers detailed annotations of continuous emotions during naturalistic conversations. K-EmoCon comprises diverse measurements, including audiovisual recordings, EEG, and peripheral physiological signals. The data was collected through off-the-shelf devices during 16 sessions, each lasting approximately 10 min, featuring paired debates on a social issue. What distinguishes K-EmoCon from existing datasets is incorporation of self, debate partner, and external observers emotion annotations.

We identified new datasets MED4 and PME4. The MED4 (Wang et al. 2022) database is a collection of synchronized signals recorded from participants, including EEG, PPG,

Table 3 Summary of behavioral-physiological databases

Database	Year	Signal type										Emotion labels	Subjects
		EEG	ECG	EMG	EOG	GSR	RSP	PPG	Others	Audio	Visual		
DEAP	2012	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	arousal, valance, liking	32
MAHNOB - HCI	2012	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	valance, arousal, dominance	27
SEED- IV	2018	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	positive, negative and neutral	44
PME4	2022	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	anger, disgust, fear, happy, sad, surprise, neutral	11
MED4	2022	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	happy, sad, angry, neutral	16
DREAMER	2018	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	valance, arousal, dominance	26
ASCERTAIN	2018	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	valance, arousal, dominance, and predictability	58
Phy-MER	2023	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	valance, arousal, 7 discrete emotions	30
K-EmoCon	2020	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	valance, arousal	32
AMIGOS	2018	✓	✓	✗	✗	✓	✗	✗	✓	✓	✓	valance, arousal, liking	40

speech, and facial images. The subjects were exposed to video stimuli specifically created to elicit happiness, sadness, anger, and neutrality. The experiment involved 32 participants and was conducted in two different environments: anechoic chamber and natural noises. The MED4 dataset is a resource for researching emotions and their physiological and behavioral correlates in different environmental conditions. Multimodal emotional datasets with physiological data from posed emotions are lacking. While spontaneous expressions offer naturalness, posed ones are less ambiguous and more intense (Chen et al. 2022). The researchers have created a comprehensive dataset named PME4 (Chen et al. 2022). This dataset comprises synchronized physiological and non-physiological signals, facilitating the comparison of emotions expressed by subjects. It encompasses four modalities: EEG, EMG, audio, and video, thus serving as a novel resource for investigating posed multimodal emotions.

3.1.5 Challenges

Multimodal emotion recognition using EEG faces several challenges: (a) integrating different types of information, (b) finding generalizable features, (c) handling non-stationary signals, and (d) addressing computational requirements.

Firstly, it is difficult to effectively integrate the spatial, spectral, and temporal information of EEG signals to realize better emotion classification performance (Gong et al. 2023). Analyzing non-stationary random signals solely in the time domain is not sufficient. To address this limitation, fusion feature analysis is introduced, allowing simultaneous consideration of dynamic changes in both time and frequency domains (Zheng et al. 2021). However, this approach has added processing steps, making it computationally more demanding than analyzing signals in just one domain (Gao et al. 2020b). This increased computational load could be challenging, particularly in real-time applications and for users without specialized expertise in signal processing techniques.

Secondly, finding discriminative features that generalize well to different EEG datasets is still challenging (Taha et al. 2023). Additionally, EEG signals are non-stationary continuous sequential signals, making feature extraction a challenging task (Liu et al. 2023). Furthermore, the increasing amount of computing power required for deep learning poses challenges in providing real-time detection and improving the robustness of deep neural networks (Pan et al. 2023). These challenges necessitate the development of models that can learn spatial and temporal representations of EEG signals and incorporate effective feature fusion mechanisms to improve discrimination power. Researchers have explored various combinations of features and physiological signals, but there is no consensus on which ones are the most informative or compelling. This lack of clear evidence suggests that further investigation is needed to determine the optimal combination of features and physiological signals for emotion detection (Qiu et al. 2018). Moreover, a notable gap exists between model sensitivity and fusion input cues, remaining underexplored in multimodal studies.

Modalities of varying importance may accommodate differing levels of reliability. This gap prompts a comprehensive literature survey to focus on EEGs' role in multimodal emotion recognition.

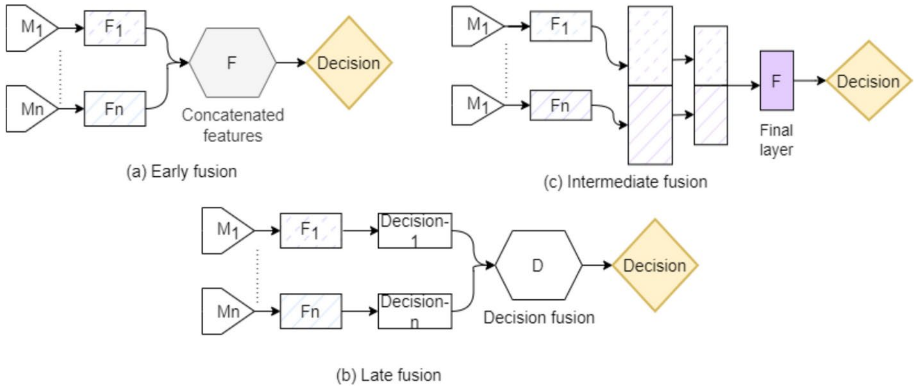
3.2 RQ2: What fusion networks are currently accessible, particularly those integrating EEG data?

In the context of the multimodal approach to modality fusion, two key issues stand out as critical: the choice of fusion (fusion level) and the fusion method (how to fuse). The fusion level determines when, in the process, data from different sources are merged—whether it occurs early on, at the feature level, or later, at the decision level. Different taxonomies are established for fusion architectures; however, they are broadly classified as early, late, and intermediate fusion, as illustrated in below figure (Fig. 6).

In the fusion process, the choice of fusion operation plays a crucial role in determining how data merging occurs, whether through concatenation, weighted fusion, logical operations (max/min/OR), or more advanced techniques like deep learning architectures. These operations exhibit effectiveness in specific contexts and can be categorized as rule-based or classification-based approaches. Rule-based fusion methods entail basic principles, including weighted fusion, max/min/AND/OR operations, and majority vote rule, while classification-based methods are employed to categorize multimodal observations into predefined classes. Decision fusion, for instance, combines classifier observations linearly with equal weights assigned to each observation class. However, with the advent of deep learning, the concept of feature-level fusion has evolved to include intermediate fusion, expanding the possibilities for combining information from multiple modalities (Ramachandram and Taylor 2017). First, we provide basic fusion operations discussed in the literature, and next, we answer the second research question.

In early fusion, first low-level descriptors are directly extracted from each modality and are merged using a simple concatenation operation before being sent to a classification scheme. Usually classification scheme involves a single classifier trained to learn correlations and interactions among low level features of each modality. Another commonly employed merging operation in the early-fusion approach is the weighted sum method. This method is frequently utilized in early-fusion scenarios, involving the multiplication of each modality's feature vector by a corresponding weight and then summing them together. We denote the single classifier/model as ‘g’ and each input signal/modality feature vector as ‘X’, Then the final prediction can expressed as:

a. Concatenation



(a) Early fusion

(c) Intermediate fusion

(b) Late fusion

Fig. 6 Illustration of different levels of fusion of multi-modal signals

$$X = [X^1, X^2, \ldots\ldots\ldots, X^N] \tag{1}$$

$$FinalPrediction = g(X)$$

b. Weighted feature fusion

$$X = W^1X^1 + W^2X^2 + \cdots + W^NX^N \tag{2}$$

$$= \sum_{i=1}^{n} W^iX^i$$

$$FinalPrediction = g(X)$$

where, X^i represents individual feature vectors extracted.W^i is the weight of the ith feature vector and X represents the merged (or final) vector.

The choice of weighting scheme in multimodal fusion significantly affects the fusion process. Weighted sum fusion is a common strategy involving assigning weights to each modality's feature vectors either manually based on prior knowledge or automatically during training to adaptively assign weights based on the data.

Late-fusion often creates multiple independent (called sub-classifiers in this case) models for joint decision-making through a range of fusion mechanisms (F) aimed at merging and ensemble the sub-classifier observations into a unified decision. If model g_n is used on modality n (n = 1,……,N) the resulting prediction is (Liu et al. 2018):

$$FinalPrediction = F[(g_1(f_1), g_2(f_2), \ldots\ldots\ldots, g_n(f_n)] \tag{3}$$

where $g_n(f_n)$ denotes the outcome of each sub-classifier. $g_n(f_n)$ can be either a confidence score, classification probabilities, classification distance or any other relevant metrics about various information classes. The decision fusion methods can opt for any of the following mechanisms:

a. Max or min rule fusion

In the max rule fusion, the final decision is determined by choosing the class with the maximum confidence score across all classifiers or modalities. Mathematically it can be represented as:

$$FinalPrediction = \mathrm{arg}max_i\{Conf_i\} \tag{4}$$

In the min rule fusion, the final decision is determined by selecting the category with the lowest confidence score across all classifiers or modalities. Mathematically it can be represented as:

$$FinalPrediction = \mathrm{arg}min_i\{Conf_i\} \tag{5}$$

where $\{Conf_i\}$ represents the confidence scores (probability or score) associated with class$_i$ from each classifier or modality, and $\mathrm{arg}max_i$ (or $\mathrm{arg}min_i$) denotes the index of the class with the maximum (or minimum) confidence score.

b. Majority voting fusion

In the majority voting rule, the final decision is determined by selecting the class that receives the most votes across all classifiers or modalities. Mathematically, it can be represented as:

$$FinalPrediction = \mathrm{arg}max_i \{Votes_i\} \tag{6}$$

where $\{Votes_i\}$ represents the number of votes received by class$_i$ from each classifier or modality, and $\mathrm{arg}max$ denotes the index of the class with a maximum number of votes.

c. Decision-weighted sum

Another decision-level fusion method, “decision-weighted sum,” entails processing and training each modality with its classifiers. Afterwards, the average prediction scores for each category generate the outcome. In this method, the fusion process involves weighing the prediction scores for each category from each modality and combining them. The final prediction is determined by the max-rule. To ascertain the best fusion weights, techniques like cross-validation on a validation dataset, reinforcement learning, or beam search methods are usually employed. We denote P_{im} as prediction score of ith modality for the m-th category.

$$Decisionweightedsum = \sum_{i=1}^{n} W_i P_i$$

$$FinalPrediction = max_m \langle \sum_{m=1}^{m} W_{im} P_{im} \rangle \tag{7}$$

Intermediate fusion involves combining individual modalities at a mid-level processing state before being sent to a decision-making system. Intermediate fusion is an alternative approach, expanding feature-level fusion by integrating deep learning frameworks. This method offers greater flexibility, particularly in facilitating the fusion of representations across various levels of abstraction. Multimodal shared representations can be produced either directly through a dedicated fusion layer or progressively through multiple fusion layers operating at different hierarchical levels (Ramachandram and Taylor 2017). Domain-specific neural networks are commonly employed on various modalities to generate their respective representations, which are merged or aggregated. A straightforward technique for merging hidden representations involves concatenating weights from diverse modalities. Subsequently, a prediction is typically made based on the aggregated representation, often using another neural network to learn complex mappings between input and output. This characteristic is evident from Eq. (2).

The major trends in EEG signal fusion with various affective modalities such as facial, speech, and other PPS signals are the focus of this review section. Various types of fusion approaches have been considered for emotion detection with shallow and deep learning methods. However, the fusion networks reviewed in this article are divided into conventional (32% studies), deep learning (47% studies), attention-based (9% studies), and autoencoders-based (12% studies) methods. This categorization is based on their underlying fusion and classification strategies. The conventional approaches implement early and late fusion strategies and use machine-learning methods. Figure 7 displays the aggregated information.

3.2.1 Conventional fusion methods

Conventional fusion methods have been used to integrate information from diverse modalities before the emergence of deep learning. They often include feature-level (early) fusion, decision-level (late) fusion, and hybrid-fusion. In this section, we provide a comprehensive review of these conventional multimodal fusion strategies for emotion detection reviewed in this paper. Table 4 illustrates the modalities combined and fusion rules employed in the studies reviewed within the conventional fusion category.

3.2.1.1 Early fusion Regarding feature-level fusion, the data acquired from diverse sources are merged or stacked. Feature-level fusion approaches have been increasingly used to exploit the complementary representation properties among various signals. Several studies have shown that different modalities complement each other in emotion detection for classifying various emotional states.

In Ref. (Zhao et al. 2019), authors adopted a simple feature concatenation strategy between eye movement (representing external subconscious behavior) and EEG (representing internal neural patterns) to ensure the proper representation of information from each modality. The results indicate that five emotions display distinguishable neural patterns, with pupil diameter demonstrating relatively high discriminatory capability compared to other eye movement features.

In Ref. (Fu et al. 2023), the researchers adopted a novel early fusion method to effectively combine eye movement and EEG signals by extracting complementary information and performing feature fusion using a dual-branch feature extraction module and a multiscale feature fusion module with cross-channel soft attention.

Researchers have developed several methods for extracting and fusing EEG features, but they often face challenges like high time complexity and insufficient accuracy. In the realm of multimodal emotion recognition, achieving high predictive accuracy requires effective methods to extract and fuse features. Deep learning, which emphasizes end-to-end learning,

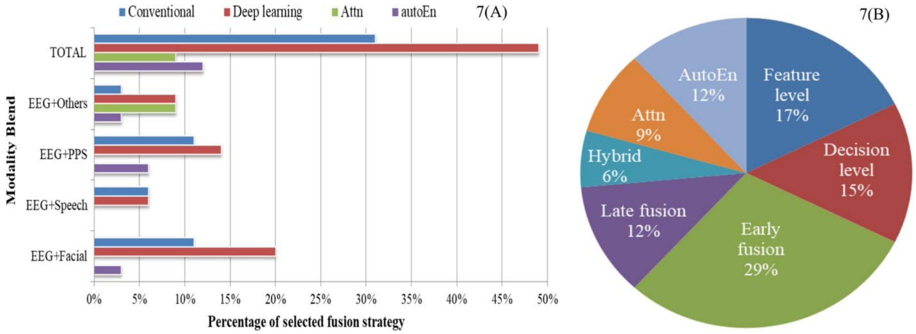


Fig. 7 **A** The percentage of selected strategy, **B** The aggregated information based on the method of fusion strategies

Table 4 Compilation of representative works under the conventional fusion category

	Fusion level	Modality combination	Fusion rule	Studies
Conventional Fusion Techniques	Early fusion	EEG + Facial	Concatenation	Chaparro et al. (2018); Mutawa and Hassouneh (2024)
			Element-wise summation	Muhammad et al. (2023); Wang et al. (2023)
			Tensor fusion	Li et al. (2022)
		EEG + Speech	Concatenation	Wang et al. (2022); Li et al. (2021)
		EEG + EyM	Concatenation	Zhao et al. (2019)
			Element-wise summation	Fu et al. (2023)
		EEG + Phy	Concatenation	Zhang et al. (2021b)
		EEG + BVP	Concatenation	Nakisa et al. (2020)
	Late fusion	EEG + Facial	Weighted sum rule	Huang et al. (2017); Lu et al. (2021); Wang et al. (2021)
		EEG + Speech	Weighted sum rule	Ghoniem et al. (2019); Wang et al. (2022)
		EEG + EI	Fuzzy fusion rule	Su et al. (2019)
		EEG + ECG + Speech	Majority voting rule	Guo et al. (2020)
		EEG + Facial + GSR + PPG	Majority voting rule	Saffaryazdi et al. (2022)

EyM eye movement

has emerged as a promising approach. Consequently, deep learning models are increasingly employed in multimodal emotion recognition and classification based on EEG data. Due to the dynamic nature and low signal-to-noise ratio of signal data, manually extracted features often fall short in accuracy. To address this, researchers incorporate conventional features alongside manual ones, optimizing their weights based on output accuracy to enhance ER models. The authors (Zhang et al. 2021b) proposed an early fusion method using weighted feature fusion to combine EEG and four different PPS.

The majority of emotion analysis methods, as indicated by references (Chaparro et al. 2018; Li et al. 2022; Wang et al. 2023; Moin et al. 2023; Mutawa and Hassouneh 2024), focus on fusing facial and EEG data to increase the overall recognition accuracy. By combining these two modalities at the feature level, the researchers aim to leverage the complementary features of brain activity captured through EEG and facial expressions, potentially leading to more robust and accurate ER systems. The work (Li et al. 2022) aims to classify the emotional expressions in subjects with hearing impairment based on EEG and facial expressions. In this work, the authors aim to explore the emotional changes in individuals with hearing impairment, utilizing a feature fusion strategy that combines EEG topographic maps (ETM) and facial expressions to extract effective emotional representation features across different modalities. However, the main challenge tackled by existing emotion detection methods in fusing facial expressions is the necessity to train a new model for facial feature extraction, which consumes substantial computing resources. To address this issue a CNN-based feature extraction model combined with the weighted fusion method is adopted to enhance multimodal emotion recognition by enriching the combined EEG-Facial feature set (Wang et al. 2023).

Poor performance in EEG-based ER models is attributed to the absence of emotion-related information in some EEG data channels. Additionally, it is not recommended to include all EEG channels in a deep learning model due to concerns about time complexity. Canonical correlation analysis (CCA) is crucial for fusing EEG and facial expressions as it identifies shared patterns across modalities, facilitating the creation of a unified representation for better emotion recognition accuracy. CCA enables the extraction of correlated features from EEG and facial data, enhancing the understanding of emotional states by capturing complementary information from multiple sources. The authors (Muhammad et al. 2023) proposed a lightweight deep CCA-based MEC model. By applying DCCA fusion, the method aimed to extract and combine relevant information from the EEG and facial video clips. The field of video emotion recognition continues to hold significant promise for future exploration (Noroozi et al. 2019). In Ref. (Xing et al. 2019), the researchers aimed to design a video emotion classifier by fusing EEG data, and video (both audio and visual) data. Notably, the authors emphasize the importance of video brightness in influencing the audience's mood and the overall atmosphere within video scenes. To capture this aspect, the luminance coefficient of the video is specifically extracted as a low-level visual feature. The study then proceeds by merging EEG and video features as early fusion, suggesting a comprehensive exploration of both physiological and visual elements in understanding and classifying emotional responses to videos.

The integration of audio signals into EEG responses is an area with limited research. In the study (Wang et al. 2022), the researchers investigated the complementary information between audio and EEG responses using a feature concatenation strategy. This approach allowed them to explore the combined representation of audio and EEG to enhance the efficacy of emotion recognition tasks potentially. Although models based on EEG data can capture the most authentic emotional states, the EEG signal's signal-to-noise ratio tends to be low, which can impact emotion recognition accuracy. Since audio signals also convey emotional information, fusing these two signal modalities may improve emotion recognition effectiveness. The authors (Li et al. 2021) subscribed to the aforementioned perspective and employed a feature concatenation fusion strategy to fuse EEG and audio-signal representations. The opinion of the researchers (Jaswal and Dhingra 2023) is that despite feeling and discourse recognition being distinct research domains, integrating voice and EEG signals can lead to a robust psychological model for emotion recognition. They propose exploring novel methods that combine emotion and speech to enhance the accuracy of individual and group emotion detection. They explore the integration of input signals derived from both EEG and audio sources in their analysis as early fusion.

3.2.1.2 Late fusion Regarding decision-level fusion, the outputs obtained from two classifiers in a bi-modal system (combining EEG with facial or speech expressions) are concatenated or fused. EEG is utilized as an inner channel to complement facial and speech expressions, and improve the overall accuracy and reliability of the ER system. Effective methods for decision-level fusion include averaging /weighted sum of class probabilities (Ghoniem et al. 2019), production rules (Huang et al. 2017), and fuzzy-fusion strategies (Su et al. 2019). When more than two modalities are involved, a multiple decision-making approach based on voting (Guo et al. 2020) is performed to determine the correct emotional category. The category receiving the highest number of votes prevails, and the final decision outcome is determined by the emotion category with the most significant number of votes.

This voting-based method helps to integrate the information from multiple modalities and arrive at a more robust emotion classification.

The fusion problem of the EEG-Expression combination is addressed using the max-weight fusion strategy, as proposed in the reference (Wang et al. 2021). According to the maximum rule, the outcome with the highest score is considered as the final result. In Ref. (Lu et al. 2021), a deep-learning-based fusion is proposed to combine expression and EEG features. The main contribution of the research in Ref. (Saffaryazdi et al. 2022) is centred around the hypothesis that identifying and analyzing the most emotional moments in stimuli allows for a better understanding of the body's reactions to emotions. The primary objective of the fusion in their work is to advance emotion recognition by integrating strategies for analyzing facial micro-expressions with EEG and other physiological signals. A majority voting fusion strategy is adopted for this work.

3.2.1.3 Hybrid fusion A hybrid temporal fusion model was used to combine the EEG and BVP signals (Nakisa et al. 2020). The model enabled joint learning and exploration of highly correlated emotional representations across both modalities. The approach involved training each modality with separate deep networks and then integrating their learned representations to capture the complementary information between BVP and EEG for more effective emotion classification.

In a different study (Cimtay et al. 2020) the researchers conducted experiments using three modalities: EEG, GSR, and facial expressions. For early fusion, EEG and GSR data were combined and input into a deep network. Additionally, a separate classifier was trained for facial expressions. The final emotion class was determined by feeding the outputs of both classifiers into a decision tree.

3.2.2 Deep learning fusion methods

Within the broad category of multimodal fusion, this section also explores the significant deep fusion techniques including attention and autoencoders networks. Table 5 illustrates the modalities combined and fusion rules employed in the studies reviewed within the deep fusion category.

Table 5 Compilation of representative works under the deep fusion category

	Fusion level	Modality combination	Fusion rule	Studies
Deep learning fusion techniques	IF	EEG + Facial	IF/LSTM	Li et al. (2019b); Wu et al. (2020)
		EEG + Phy	IF/LSTM	Ma et al. (2019)
	AN	EEG + Video	Attention weighted sum	Choi et al. (2020)
		EEG + EYE + EIG	Attention weighted sum	Lan et al. (2020); Zhao et al. (2021)
	AE	EEG + Facial	Shared representations	Zhang (2020)
		EEG + Eye	Shared representations	Zhao et al. (2019); Zheng et al. (2019)

IF intermediate fusion, *AN* attention network, *AE* autoencoders

3.2.2.1 Intermediate fusion Intermediate fusion is a new paradigm of multi-modal data fusion strategy using deep neural networks, often known as joint fusion models. The intermediate fusion model primarily involves combining intrinsic feature representations from the hidden layers of deep models with representations from heterogeneous modalities, which are then used as input to a decision function. During the fusion process, a shared or joint representation layer (called a fusion layer) is formed where the multi-modal features are merged. The simplest way to merge multi-modal features is to use an additive approach (Ma et al. 2019), in which modality-specific semantics are combined in a distinct hidden layer to form a fused vector.

Different from the late-fusion method, which can model heterogeneous modalities, the joint fusion model can adequately capture the cross-modality and the inter-modality complementary correlations of multi-modal data. Ref (Ma et al. 2019) presented a joint-fusion method for EEG and PPS signals using a multi-modal "Long Short Term Memory", LSTM network. The model has a four-layer LSTM network for each modality, which enables the capture of intra-correlations among EEG and PPG signals, where concatenation of weights or representations occurs at each layer during the training process. The work of (Li et al. 2019b) used a special framework called LSTM to combine information from EEG and facial expressions to classify emotions better. The model allowed an understanding of how emotions change over time, making emotion recognition more accurate. Recent advancements in the intermediate fusion approach have integrated typical models such as encoder-decoder architectures to capture shared semantics across EEG and eye movements (Zheng et al. 2019).

Additionally, attention mechanisms have been employed to selectively highlight and combine the most complementary features from EEG and facial video (Wu et al. 2020). This strategy enables the fusion layer to effectively integrate information from multiple sources at an intermediate stage, enhancing the model's ability to learn a cohesive representation of the input data.

3.2.2.2 Attention-based fusion Attention mechanisms have shown promising results across various tasks, such as question answering and machine translation. The attention mechanism enables a neural network to learn adaptable fusion weights for several modalities, leading to enhanced multimodal fusion and identification of emotions (Ahmed et al. 2023). In recent studies, multimodal attention networks have been investigated to enhance the complementary relationship between EEG and other modalities in superior MEC systems.

Various modalities may have varying levels of impact on the overall task or prediction of emotions. In an emotion recognition task utilizing facial video modality, different types of features may contribute differently to identifying an individual's emotional state accurately. Some individuals may experience negative or positive emotions without displaying facial expressions. To accurately identify genuine emotions, incorporating EEG signals alongside facial videos is preferred (Choi et al. 2020). Attention mechanisms allow the model to dynamically weigh the importance of features from each modality, enabling it to selectively integrate the most relevant information from EEG and facial-video modalities to understand a person's emotional state. The proposed method features a multimodal attention network, serving as a pivotal module. In this network, the intermediate features extracted from two modality networks are taken, and the weights of each modality are evaluated by comparing their respective features. Subsequently, fusion occurs through the multimodal attention network during the output stage. Both (Lan et al. 2020) and (Zhao et al. 2021) leverage attention mechanisms to combine information from EEG, raw eye images, and eye movements for multimodal fusion. Ref (Lan et al. 2020) introduced an attention mechanism among these three modalities and examined the contributions of each modality to emotion recognition. However, ignoring the interaction or mutual influence between modalities may lead to suboptimal fusion results (Zhao et al. 2021). Conversely, in Ref. (Zhao et al. 2021), the attention mechanism is used to improve the accuracy by enhancing the weights of critical feature channels and establishing correlations among different modalities. This is achieved through the novel co-attention layer, which focuses on the most relevant information from each modality and captures inter-modality interactions.

3.2.2.3 Autoencoders Presently, the dual-model fusion strategy employing deep autoencoders (BADE) has shown promising results in effectively discriminating interactions across EEG and behavioral modalities. In their study, the authors demonstrated the interactions between EEG and facial videos (Zhang 2020) by utilizing automatic encoder structures to extract high-level representations. These representations enabled a deeper focus on the relationships between different modalities, leading to improved performance in the recognition tasks. However, the extent to which EEG and eye movements offer complementary representations for discriminating between the five emotions (happy, sad, fear, disgust, and neutral) remains unclear and explored in (Zhao et al. 2019). Additionally, eye movement signals not only provide physiological cues but also convey crucial subconscious activities, thereby offering contextual hints for emotion recognition. The authors also employed BADE for modal fusion. During the encoding stage, two Restricted Boltzmann Machines (RBMs) are utilized to process EEG and eye movement features separately. Subsequently, the outputs of these RBMs, represented by two hidden layers (hEEG and hEye), are concatenated directly as the input for an upper autoencoder. The decoding network mirrors the structure of the encoding network, intending to reconstruct the original EEG and eye movement inputs. Lastly, the newly derived shared representations are inputted into a linear Support Vector Machine (SVM) to acquire the recognition results.

A recent work (Zhou et al. 2018) suggested the use of Convolutional Auto-Encoder (CAE) to fuse EEG and multi-type physiological (EP) signals (Zhou et al. 2018) that exhibited a commendable performance CAEs can achieve multimodal fusion by integrating inputs from various sources, leveraging convolutional neural networks (CNNs).

CAEs differ from conventional autoencoders (AEs) in that their weights are shared across all locations in the input data, thereby preserving spatial locality (Masci et al. 2011). The trained encoder (seven-layer CNN) of the CAE is utilized to acquire features from multichannel EEG and EP signals for the final result.

A review of research comparing major fusion techniques for emotion analysis from selected studies is shown in Fig. 8.

To better compare the performance of different fusion methods, we analyzed the number of emotional categories recognized in each study alongside the respective fusion approaches. Figure 8 illustrates that nearly all selected studies achieved average accuracies above 80%, with the exception of one study (Nakisa et al. 2020). Notably, the best performance of

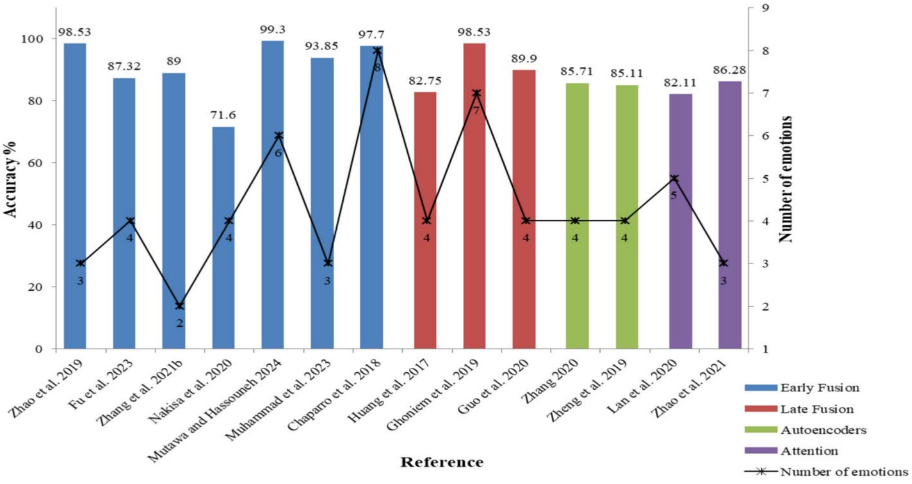


Fig. 8 Comparison of the accuracy of various studies in terms of fusion strategies and the number of emotions recognized

97.7% was observed in Ref. (Chaparro et al. 2018), which achieved this accuracy while recognizing the maximum of eight emotion categories.

3.3 RQ3: What parameters and methods are interesting for accurate emotion recognition?

Features can be extracted or obtained from time-, frequency-, or time–frequency domains from selected modalities in various ways. Successful approaches for the task of automatic emotion identification include wavelet transforms, differential entropy, non-linear dynamics and many more. Fused modalities, fusion level, and the number of parameters dictate the decision strategy for multimodal performance. The previous section presents the fusion level and fused modalities (Sect. 3.2). This section aims to understand the number of parameters and methods used in predicting emotions from the selected modalities. In this regard, we delve into the analysis of EEG features, exploring various feature extraction methods and assessing different techniques for feature selection.

3.3.1 EEG features

The accuracy of AER systems relies on the EEG signals used to describe emotions. Moreover, the better the quality of the descriptor, the more accurate the classification ability. The field of AER uses various EEG features to identify and comprehend different emotional states experienced by individuals. These features can be classified into three categories: time-series features, spectral features, and time–frequency features.

Recently, several non-linear estimators have been utilized to quantify the complexity of EEG signals. The non-linear dynamic features unlock rich emotion-specific patterns. Below are some of the most frequently used procedures for analyzing EEG signals.

3.3.1.1 Time-domain (TD) features Many researchers have attempted TD features to analyze the performance of EEG data to extract emotion-specific patterns. Time-domain features for EEG signals are statistical and descriptive characteristics that are extracted from the raw EEG data in the time domain. These features provide valuable information regarding the temporal characteristics of brain activity recorded by EEG electrodes.

The popular time-series feature is Hjorth parameters (Hjorth 1970). Hjorth parameters are characteristics derived from the variations in the EEG signal's derivatives (mobility, activity, and complexity). Additionally, the mean absolute value of mobility, activity, and complexity can also serve as features (Wang et al. 2011).

Among the several categories of temporal features, one frequently employed category is instantaneous statistics (Motamedi-Fakhr et al. 2014). The instantaneous statistical information is directly derived from moments of the EEG signal. The metrics like mean, absolute value, kurtosis, skewness and similar (Wang et al. 2011; Ghoniem et al. 2019) are widely used in many studies. They also encompass measures related to the probability density function of the waveform, such as the mode, median, and entropy. Instantaneous statistics treat each data point within the time series as an independent univariate process without taking into account any potential correlations or relationships between these individual data points.

Apart from the relatively understandable statistical attributes, we also have the option to derive more sophisticated characteristics from the data. For instance, advanced features like the fractal dimension, FD (such as the Hurst Exponent) are linked to the long-term memory of a time series (Ruiz-Padial and Ibáñez-Molina 2018). Additionally, one can consider the concept of entropy, which measures the level of both regularity and unpredictability in fluctuations across our time series (Wang et al. 2018). Even though statistical metrics are less utilized individually, emotion identification using physiological data shows superior performance. For instance, the time-domain features of PPG signals exhibit a strong correlation with emotional arousal (Fu et al. 2022). Although time-domain features retain the richness of EEG data due to the complex EEG waveform, a universally applicable method for analyzing these features is lacking. Hence, time-domain analysis requires significant expertise and understanding.

3.3.1.2 Frequency-domain (FD) features In the frequency domain, the corresponding affective descriptors are analyzed in terms of their frequency components. For any given time-based signal, we can transform a time-series signal into its corresponding spectral representation. This transformation enables us to observe the distinct frequencies that constitute emotion-related features.

In the context of spectral analysis, spectral estimation is about determining spectral density (Li et al. 2019b). Power spectral density (PSD) or simply the spectral density of the signal is the common FD attribute utilized in EEG analysis (Zhang 2019). In the frequency domain, statistical metrics such as mean, variance, median, standard deviation, skewness, kurtosis, and similar metrics are also frequently employed.

Another widely used frequency-domain feature is the differential entropy feature. A study (Lan et al. 2020) utilized differential entropy (DE) features in an EEG-based emotion recognition system. Two other common features that can be computed include ratios and differences between DEs -RASM and DASM respectively. These features are used to quantify the asymmetry or differences in the statistical properties of EEG signals between the left and right hemispheres of the brain (Duan et al. 2013).

3.3.1.3 Time–frequency (TF) features When dealing with a non-stationary signal, relying solely on its spectrum becomes less informative. Therefore, it becomes imperative to introduce a time dimension and estimate the signal's frequencies at each time interval (Morales and Bowers 2022). To analyze spectral properties of the non-stationary EEG signals, various methods have been adopted: power distribution methods and signal decomposition methods. Power distribution methods focus on calculating the distribution of signal power across different frequency bands as the signal evolves. Signal decomposition methods involve breaking down a complex signal into its constituent components, enabling the analysis of individual components separately (Zhang 2019).

Short-Time Fourier Transform (STFT) and Wavelet Transform (WT) techniques are used to convert waveform to EEG frequency spectrum. The STFT does not achieve a balanced frequency and time resolution; whereas the WT proves to be better suited for analyzing non-stationary EEG signals (Murugappan et al. 2010).

Wavelet transform is a time–frequency decomposition method. Therefore, wavelet transform emerges as a more appropriate method for breaking down EEG signals into separate frequency bands. Several emotion-related features from wavelets (energy and entropy) are generally utilized. However, it might not be as efficient in handling noise components within the signal (Yong and Menon 2015).

Recently, empirical mode decomposition (EMD) methods have been used in the EEG emotion model (Salankar et al. 2021). EMD is a "time–frequency analysis method to deal with non-linear and non-stationary signal" like EEG waves (Zhang et al. 2016b). Using EMD and Second Order Difference Plot (SODP), several features can be extracted, such as the area of the SODP, mean distance, and central tendency measure. Hilbert Huang Transform (HHT), Wavelet Packet Decomposition (WPD) (Ting et al. 2008; Zhang et al. 2017b), Matching Pursuit (MP) (Franaszczuk et al. 1998) and Morlet Wavelet Transform (MWT) (Cohen 2019) are also crucial approaches for time–frequency feature analysis. In multimodal emotion identification based on EEG data, commonly utilized feature extraction methods are detailed in Sect. 3.3.2.

3.3.1.4 Features of nonlinear dynamical systems Recent studies indicate that the human brain behaves as a nonlinear dynamic system, attributing EEG signals to its output (Xingyuan et al. 2009). As a result, researchers are focusing on nonlinear dynamics to scrutinize EEG data (Sharma et al. 2020). Nonlinear dynamics methods can be broadly classified into chaos theory and information theory. Chaos theory-based approaches involve metrics like correlation dimension (Khalili and Moradi 2009), Kolmogorov entropy (Aftanas et al. 1997), Lyapunov exponent (Übeyli 2010), and Hurst exponent. Information theory methods encompass Approximate Entropy (ApEn) (Pincus 1991), Sample Entropy (SampEn)

(Richman and Moorman 2000), Permutation Entropy (PeEn) (Bandt and Pompe 2002), and measures of complexity.

3.3.2 EEG feature extraction methods

The stage after domain analysis is feature extraction. After eliminating noise from the selected modality data, it is crucial to extract important and diverse features before passing them to the classifier. Feature extraction methods play a critical role in emotion analysis, capturing information that accurately reflects an individual's emotional state. These extracted features serve as the foundation for algorithms employed in emotion classification tasks. The significance of these methods lies in their ability to determine the accuracy and effectiveness of emotion identification, highlighting the importance of selecting and implementing feature extraction techniques that faithfully represent the nuances of emotional states. The study covers various feature extraction techniques.

3.3.2.1 Feature extraction using STFT For stationary signals, the Fourier Transform (FT) is often regarded as the most commonly used operation for obtaining the spectrum of the signal. As for non-stationary signals, the same spectrum representation is not purposeful. A spectrogram, on the other hand, is helpful for tracking how different frequencies evolve, which is crucial for understanding a signal's time-varying phenomena (Brynolfsson 2012). In general, a spectrogram is generated by calculating the periodogram for short-time segments of a signal, which is realized through the STFT method. Compared to FT, the STFT is a more localized method that breaks the signal into smaller, overlapping segments using a time window and then applies the Fourier Transform to each segment separately (Bruns 2004). A widely used window function in STFT analysis is the Hamming window, as described by Harris in 1978 (Harris 1978).

The conventional means of EEG signal spectrum computation involves a classic approach, realized through Fourier Transform. On the other hand, PSD computation varies among different studies, majorly categorized as parametric and nonparametric methods.

Parametric methods assume that the signal can be modeled using a certain equation or model with a set of parameters. For example, parametric methods often involve modeling the signal as an autoregressive (AR) process. Unlike parametric methods, which assume a specific model beforehand, nonparametric methods often rely on the Discrete Fourier Transform (DFT).

Welch's method stands out among the non-parametric approaches, while the AR method is a prevalent choice among parametric approaches (Zhang 2019). The main drawback of nonparametric methods is that the computation involves data windowing, which can introduce distortion in the resulting power spectral densities (PSDs).

Welch's method can also be applied to estimate the spectrogram. Welch's method uses a modified segmentation scheme and average periodogram. Averaging the periodogram values reduces the statistical oscillations, which generates a stable Welch PSD estimator. Computationally, the periodogram $S_x^l(f)$ is obtained from the signal's windowed segment, $X^l(f)$ and window power, U. For *l–th* segment the computations are as follows:

$$X^{l}(f)=\sum_{n=0}^{N-1}x_{l}\left(n\right)w(n)e^{-i2\pi fn} \tag{8}$$

$$U=\frac{1}{N}\sum_{n=0}^{N-1}\left|w(n)\right|^{2} \tag{9}$$

$$S_{x}^{l}(f)=\frac{1}{NU}\left|X^{l}(f)\right|^{2} \tag{10}$$

To gain PSD with Welch, it is needed the average of the modified periodogram, can be found as:

$$\widetilde{S_{x}}=\frac{1}{Q}\sum_{l=1}^{Q}S_{x}^{l}(f) \tag{11}$$

3.3.2.2 Feature extraction using WT A wavelet is a transient wave function that bears some properties such as limited period and finite energy and is widely used in digital signal processing. Wavelet transform is an optimal spectral estimation method that replaces any time-series sinusoidal wave as an infinite series of wavelets (known as a window function) by translations and dilations. Two distinct WTs are identifiable discrete WT (DWT) and continuous WT (CWT). By transform, a series of elementary functions $\psi_{a,\mathrm{b}}$ (t) are generated by shifting and dilating a fixed basis function ψ(t) which is correlated with the original signal. The wavelet transform at dilation ‘a’ and translation ‘b’ can be described by:

$$\psi_{\mathrm{a,b}}\ (\mathrm{t})=\frac{1}{\sqrt{\mathrm{a}}}\psi\left(\frac{\mathrm{t}-\mathrm{a}}{\mathrm{a}}\right) \tag{12}$$

where wavelet family and wavelet coefficient are denoted by $\psi(\mathrm{t})$ and $\psi_{a,b}(\ \mathrm{t}\ )$.

In spectrum estimation using DWT, the signal’s decomposition into a set of sub-bands is accomplished through successive high-pass and low-pass filtering of the time-domain signal. The outputs of these filters undergo down-sampling by a factor of 2. This process leads to the generation of the detail component (D) from the high-pass filter, and the other is approximation (A) from the low-pass filter. The DWT decomposition for a given signal $x\left(t\right)$ is obtained as (Subasi 2007):

$$x\left(t\right)=\int_{k=-\infty}^{k=+\infty}C_{n,k}\varphi\left(2^{-n}t-k\right)+\int_{j=1}^{n}\int_{k=-\infty}^{k=+\infty}D_{j,k}2^{-j/2}\psi\left(2^{-j}t-k\right) \tag{13}$$

where $D_{j,k}$ denotes j-th level k-th detail component, $C_{n,k}$ represents k-th approximate component, while $\psi\left(\mathrm{t}\right)$ and $\varphi\left(\mathrm{t}\right)$ correspond to wavelet and scaling functions, respectively.

It should be noted that the incorporation of all DWT coefficients from multiple frequency bands into a single feature vector is regarded as impractical and unnecessary (Gandhi et al.

2011). In the reviewed papers, a preference was observed for the utilization of either statistical or non-statistical parameters rather than the direct utilization of wavelet coefficients. For example, the coefficients can be summed to provide estimates of the mean and standard deviation of detailed coefficients. (Subasi 2007). In the reviewed papers, the following are the main DWT features that received significant attention (Table 6).

3.3.2.3 Differential entropy feature extraction method In the field of thermodynamics, entropy is a measure of the amount of disorder or randomness in a system. Whereas in information theory, which was developed by Claude Shannon, entropy represents the average amount of information or surprise associated with a random variable or a probability distribution (Shannon 1948).

Differential entropy (DE) is the concept from information theory and probability theory, which quantifies the uncertainty and captures the degree of randomness and intricacy present in EEG waveforms. Higher DE values may indicate more intricate or less predictable EEG patterns, which could be linked to certain neurological or cognitive states. DE is used to evaluate the complexity of continuous and random variables. It quantifies relative or changes in uncertainty, rather than computing an absolute degree of uncertainty (Michalowicz et al. 2013).

Several studies have demonstrated that entropy metrics possess significant potential for studying patterns and regularities within EEG datasets (Acharya et al. 2013). Therefore, the effectiveness of distinguishing various emotions based on their level of irregularity in the EEG signal is demonstrated by entropy (Patel et al. 2021). Distinct kinds of emotions and entropy measures are discussed in the review by Patel et al. (2021). Duan et al. indicated that for a fixed-length EEG sequence, estimating DE is equivalent to the logarithm of the energy spectrum in a specific frequency band (Duan et al. 2013). The fundamental expression of DE can be defined as

$$h\left(X\right) = -\int_{-\infty}^{\infty} p(x)log[p(x)]dx \tag{14}$$

where $p(x)$ represents the probability density function (PDF) of a continuous random sample X.

Table 6 Summary of time–frequency (wavelet-based) measures

Feature	Formula	Definition
ED(i)	$ED_i = \sum_{k=1}^{N_i} D_{i,j}^2$	Wavelet Energy
R(i)	$R(i) = \frac{ED_i}{\sum_{j=1}^{N} ED_j}$	Wavelet energy ratio
ENT	$ENT = \sum_{i=1}^{N} R\left(i\right)\left[\ln R\left(i\right)\right]$	Wavelet Entropy
μ_i	$\mu_j = \frac{1}{N_j}\sum_{k=1}^{N_j} D_{j,k}$	Mean of detailed coefficients (N_j:j-th level detailed coefficients)
σ_j	$\sigma_j = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}\left(D_{j,k}-\mu_j\right)^2}$	Standard deviation of detailed coefficients

Experimental studies have shown that after band-pass filtering, several sub-frequency bands of EEG signals closely adhere to the Gaussian distribution N(μ,σ^2), and their associated DE can be determined as

$$h\left(X\right)=-\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma_i^2}\right)\log\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma_i^2}\right)\right]dx$$

$$h_l\left(\mathrm{X}\right)=\frac{1}{2}\log\left(2\pi e\sigma_i^2\right) \tag{15}$$

where μ is the mean of the sample X and σ_i^2 is signal variance. According to Eq. (15)**,**h_l is the DE of the EEG signal in the i^{th} frequency-band. Here 'e' is the Euler's constant.

However, estimating the DE feature is quite challenging due for two reasons (i) it requires the estimation of the density of X, and (ii) not suitable for CNN-type classifiers (Hwang et al. 2020). Finally, Eq. (16) represents methods to extract DASM and RASM features.

$$DASM=h\left(x_j^{left}\right)-h\left(x_j^{right}\right)andRASM=h(x_j^{left})/h(x_j^{right}) \tag{16}$$

3.3.2.4 Spatial feature extraction method The common spatial pattern (CSP) is a prominent spatial feature extraction technique for classifying EEG signals, leveraging spatial filters to enhance the discriminability between two classes (Ai et al. 2018). The CSP algorithm has been extensively utilized for feature extraction in EEG-based brain-computer interface (BCI) systems, particularly for motor imagery (MI) applications (Kirar and Agrawal 2016). The technique employed for pattern extraction from the data is derived from the CSP method, which was first introduced in EEG analysis by Koles et al. (Koles et al. 1990). The core idea of CSP is to enhance the differentiation between two classes of EEG data by optimizing "spatial filters" that maximize variance for one class while minimizing it for the other. The authors of (Saha et al. 2021) address the limitations of traditional Common Spatial Pattern (CSP) by introducing an enhanced method called Frequency-domain Common Spatial Pattern (FCSP). In Ref. (Pan et al. 2022), CSP is used to extract spatial features to recognize fear emotion. While conventional CSP struggles to preserve discriminative features between classes in the time domain, FCSP overcomes this by transforming time-domain EEG signals into their power spectral density (PSD), allowing for the identification of event-related variations in the frequency domain. CSP is then applied to the PSD data to capture spatial patterns, resulting in more effective feature extraction and improved class reparability.

3.3.2.5 Feature extraction using EMD and SODP The authors (Salankar et al. 2021) propose an EMD-based approach for feature extraction, where the signal is decomposed into multiple oscillatory Intrinsic Mode Functions (IMFs). Distinguishing features are then extracted from the SODP, with a focus on the area covered by the elliptical shape of the

SODP, which serves as the basis for feature extraction and contributes to achieving good classification accuracy. The method can be expressed as:

Let $x(t)$ is a time-series EEG signal. The number of IMFs and residue is denoted as $Mandr_M(t)$ respectively. Then $x(t)$ can be expressed as mentioned in (Salankar et al. 2021):

$$x\left(t\right)=\sum_{m=1}^{M}D_m\left(t\right)+r_M(t) \tag{17}$$

The mean of two envelopes is $m\left(t\right)$, and is computed as.

$m(t)=\left(e_u\left(t\right)+e_l\left(t\right)\right)/2$ Where, $e_u\left(t\right)\,and e_l\left(t\right)$ are the upper and lower envelopes.

Finally, IMF, $d_i\left(t\right)$ is calculated by subtracting the mean from $x(t)$ and the final residual r $\left(t\right)$ is obtained as

$$x\left(t\right)=\sum_{i}^{M}d_m\left(t\right)+r(t)$$

After calculating and selecting IMFs, the SODP is constructed by plotting these second-order differences, revealing non-linear patterns and trends within the IMF. The necessary mathematical representation is given Eq. (17).

$$SODP_i\left(t\right)=d_i\left(t+2\right)-2d_i\left(t+1\right)+d_i\left(t\right) \tag{18}$$

The features extracted from the SODP include the area of the elliptical region, central tendency measure, and mean distance.

3.3.2.6 Time-domain feature extraction methods Most EEG acquisition systems capture signals in the time domain, making time-domain analysis methods highly suitable. These methods primarily focus on the geometric characteristics of EEG signals, resulting in minimal information loss. The time-domain characteristics offer valuable insights into the statistical patterns present in the EEG data. Two commonly used approaches for time-domain feature extraction are statistical function-based methods and entropy function-based methods. Statistical function-based methods include features such as maximum, mean, standard deviation, variance, skewness, kurtosis, and the mean of absolute values of normalized first and second differences. The Eq. (19–27) are the respective mathematical expressions of the aforementioned features (Álvarez-Jiménez et al. 2024), where $XandN$ are the EEG signal and total number of samples (experiments).

(1) Maximum:

$$\max=\left(X_i\in C\right) \tag{19}$$

X_i is considered the maximum element of the set C if every other element in the set is less than or equal to X_i

(2) Mean:

$$mean = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{20}$$

(3) Standard deviation

$$stddev = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (X_i - mean)^2} \tag{21}$$

(4) Variance

$$var = \frac{1}{N}\sum_{i=1}^{N} (X_i - mean)^2 \tag{22}$$

(5) Skewness

$$skewness = \frac{\sum_{i=1}^{N} (X_i - mean)^3/N}{(stddev)^3} \tag{23}$$

(6) Kurtosis

$$kurtosis = \frac{\sum_{i=1}^{N} (X_i - mean)^4/N}{(stddev)^4} \tag{24}$$

(7) Shannon entropy

$$ShEn = \sum_{i=1}^{N} P_i(X) log P_i(X) \tag{25}$$

(8) Mean absolute of 1st difference

$$AFD_N = \frac{1}{N-1}\sum_{i=1}^{N-1} |X(i+1) - X(i)| \tag{26}$$

(9) Mean absolute of 2nd difference

$$ASD_N = \frac{1}{N-2} \sum_{i=1}^{N-2} |X(i+2) - X(i)| \tag{27}$$

The idea behind using combinatorial features (a.k.a. mixed domain feature set) is to improve the discriminative power of the feature set and enhancing the model's ability to differentiate between different emotional states. Finding the right set of features to achieve the highest accuracy is a challenging task (Wen and Zhang 2017). Combinatorial features can be created by combining features from within the same domain or from different domains. For instance, one can combine frequency-domain and time-domain features or features within the time–frequency domain. The choice of which combinatorial features to use often depends on the specific research task, dataset, and the algorithms applied in the analysis.

Indeed, to create combinatorial feature sets, a strategy that combines various feature extraction methods is required (Zhang et al. 2017a). Finally, many literature studies have suggested that by incorporating EEG data into emotional studies, it is possible to identify the inner sense of a person's emotions. However, the key problem in using EEG data is extracting the stable and correlated features to form the feature vectors to analyze and find relevant emotions. Feature vectors are "mathematical representations of signals" This implies that during EEG signal analysis, appropriate analytical methods must be employed to ensure meaningful features. The standard quantitative assessment of EEG data is based on linear analysis, which primarily analyzes the signal's characteristics in the time domain and frequency domain. The current evolution in the field: Some researchers have suggested that "non-linear analysis methods" can provide new insights into how EEG signals can be used for emotion analysis (Sharma et al. 2020).

3.3.3 EEG feature reduction/selection

Feature selection (FS) and Dimensionality reduction (DR) are both methodologies utilized in machine learning and data analysis, yet they serve distinct purposes and employ diverse approaches. FS is a mechanism that selects the best subset of features from the candidate feature set to effectively minimize the feature space based on some criteria, thus enabling machine learning classifiers to operate within the constraints of available computational resources and amplify their predictive accuracy. Unlike other DR methods, such as those utilizing projection (like principal component analysis) or compression (such as information theory-based approaches), the FS process does not change the "original representation" of the variables. Instead, they simply choose a subset of them while discarding irrelevant and redundant ones (Saeys et al. 2007).

The FS methods can be grouped into four distinct categories: wrapper methods, filter methods, embedded methods, and hybrid methods (Somol et al. 2010). These categories vary in their methodologies and degrees of interaction with the classification model.

In filter methods, feature selection is dependent on the intrinsic properties and the relevancies that exist among features. Stated differently, filter methods operate independently of any specific learning algorithm (Yu and Liu 2003). In contrast, wrapper methods involve model-specific evaluations and accepting feedback from specific classifiers (Kohavi and John 1997). Similar to wrapper methods, embedded techniques are tailored to specific needs. However, embedded models treat FS as an optimization problem, tightly integrating it with a specific classifier to enhance its performance in classification tasks. The features that are selected are chosen deliberately to have a favourable result on the current classification task. The Hybrid method combines the strengths of more than one approach. Table 7 presents a general taxonomy of feature selection methods, outlining each technique's key merits and demerits as specified in Ref. (Saeys et al. 2007).

Most of the reviewed papers used linear DR methods like principle component analysis (PCA). Researchers find nonlinear DR methods particularly useful when they face complex data relationships that linear techniques cannot effectively capture. Li et al. (2019b) used nonlinear DR techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) to enhance regression performance by removing features that were either irrelevant or redundant.

Besides DR techniques, FS algorithms ReliefF (Zhang et al. 2016a), mRMR (Atkinson and Campos 2016) have been used for EEG feature subset selection. Decision tree (DT)-based FS algorithm method used in Ref. (Zhang 2020).

In multi-modal fusion, the redundant features or high dimensionality from processing each modality can hinder the effectiveness of commonly used learning algorithms like neural nets. This necessitates optimization. One way to address this issue is by reducing the input space dimensions to a more manageable size. Clustering emerges as a crucial solution employed to mitigate this issue by organizing groups of objects or patterns into clusters, thereby reducing the dimensionality of training data as implemented in Ref. (Ghoniem et al. 2019).

Furthermore, we also looked for the FS methods on a combined feature set to generate the relevant feature subset. In Ref. (Xefteris et al. 2022), the evolutionary approach like the genetic algorithm (GA) was used as the combined feature selection method. To compare the number of parameters and extraction methods, we selectively selected the articles based on their works, which are listed in Table 8.

3.3.4 Classifiers and classification schemes for emotion analysis

3.3.4.1 Classifiers Advanced pattern recognition algorithms are crucial in developing modern classifiers, enabling the system to identify and categorize complex patterns within vast datasets. These algorithms leverage sophisticated techniques such as neural networks and deep learning to distinguish between different categories accurately, thus enhancing the performance of various tasks in fields of affective computing applications. In this subsection,

Table 7 Taxonomy of widely used FS methods and representative works

FS models	Merits	Demerits	Example
Filter	Scaled to high-dimension data, Classifier independence, Computationally simple, Avoids over fitting	Overlooks interaction with classifier	ReliefF Kononenko (1994) mRMR Peng et al. (2005)
Wrapper	Simple, Inference with classifier, Relies on model-specific features, Risk of model over fitting	Risk of over fitting, Classifier dependent selection	Recursive feature elimination-RFE AbdelAal et al. (2018)
Embedded	Inference with classifier, Relies on model-specific features	Classifier dependent selection	Decision tree Zhang (2020)
Hybrid	Combined strength of Filter -Wrapper schemes	Ignores feature-feature interaction	MIC and QPSO Chen et al. (2023) RFI-NCA Tuncer et al. (2021)

Table 8 Summary of parameters and methods used to detect emotions from fused modalities

References	EEG features	EEG dimensions (ch x feat)	Fused modality	Feature extraction	Combined Vs individual FS method
(Ghoniem et al. 2019)	δ,θ,α,β rhythms	14 × 23	Speech	For EEG mixed set of features extracted Speech features are extracted using MFCC, and fundamental frequency	Fuzzy-clustering and GA
(Wang et al. 2022)	δ,θ,α,β, γ rhythms	32 × 5	Speech	160-dimensional differential entropy features obtained for EEG 36-dimensional frame-level MFCC features for speech	ICA
(Su et al. 2019)	δ,θ,α,β, γ rhythms	30 × 6 + 30 × 6 + 12 × 5	Eye movement	420-dimensions of EEG features; average energy, and power spectral features 20-dimensional PSD and differential entropy features from eye movement data	Individual/PCA
(Huang et al. 2017)	δ,α1,α2,β1,β2,γ1,γ2 rhythms	Not specific	Facial expression	169-dimensional facial image features obtained For EEG, power spectrum density features	Individual/PCA
(Li et al. 2019b)	θ,α,β, γ rhythms	14 × 4	Facial expression	Facial geometric features obtained by considering 29 landmarks located at the eyes and mouth 56 PSD features from EEG	Individual/t-SNE
(Zhang 2020)	θ,α,β, γrhythms	64 × 48	Facial expression	14-dimensional WPD features of EEG Sparse representation of facial features	Individual/decision tree
(Zhao et al. 2019)	δ,θ,α,β,γ rhythms	62 × 5	Eye movement	310-dimensional DE features of EEG 33 eye movement features, pupil diameter, and event statistics	Combined/Autoencoder
(Guo et al. 2020)	δ,θ,α,β rhythms	–	ECG	40-dimensional, 5040 power and central frequency features for EEG are obtained 20-dimensional, 1064 parameters from ECG data	Individual/PCA
(Xing et al. 2019)	θ,α,β,γ rhythms	64 × 5	Auditory and Visual	320-dimensions of PSD features for EEG 88 different acoustic features extracted from video 2-dimensions of colour energy, and luminance coefficient as visual features from video	Combined/PCA

ch channels, *feat* features

we delve into various learning techniques to comprehend different emotion classification schemes. Emotional classifiers can be classified into two main types: shallow and deep learning models.

Shallow-models

Support vector machines (SVM): As a standard machine learning method, SVM classifiers play a crucial role in multimodal emotion classification, offering a powerful tool for distinguishing complex emotional states by leveraging diverse data sources. SVM classifiers are commonly used in multimodal emotion classification because of their ability to handle various types of data and find the optimal discriminant hyperplane that separates different emotion classes (Vapnik 1999). By integrating multiple modalities such as facial expressions, voice intonations, and physiological signals, SVMs can enhance the accuracy of ER systems (Zhao et al. 2019), (Huang et al. 2017). The classifier transforms the input data into a higher-dimensional space where a clear separation between classes can be achieved. This makes SVMs particularly useful in scenarios where emotions are subtle and overlap significantly, ensuring a more precise and reliable classification.

Random forest: Random Forest is an ensemble learning method highly effective for emotion classification tasks. It constructs multiple decision trees during training and outputs the mode of the classes of the individual trees (Breiman 2001). Random Forest can handle various types of input data, making it suitable for multimodal emotion classification, where it can simultaneously process facial expressions, voice features, physiological signals, and textual data.

One of the strengths of Random Forest is its ability to evaluate the importance of different features. This can be particularly useful in emotion classification to determine which features (e.g., face key points) are most indicative of particular emotions (Chaparro et al. 2018). Random Forest can maintain accuracy even when some of the data is missing, which is common in real-world emotion classification scenarios where all sensors or modalities might not always be available.

Fuzzy-c means clustering (FCM): FCM, also known as soft clustering, was developed by Bezdek (1981) using fuzzy set theory (Bezdek 1981); this method assigns observations to one or more clusters based on similarity measures. The most commonly used similarity measures are distance, intensity, or connectivity. The basic FCM algorithm, being centroid-based, requires a predetermined number of clusters and is highly sensitive to initial centroid placement. However, many emotion recognition problems lack prior knowledge of the optimal number of clusters. To address this, evolutionary approaches such as genetic algorithms (GA) offer alternative optimization methods by employing stochastic principles to evolve clustering solutions (Ghoniem et al. 2019).

Ensemble learning: Ensemble learning is a powerful machine learning technique that combines the predictions from multiple models to improve overall performance and robustness. By aggregating the outputs of various classifiers, ensemble methods can reduce the risk of overfitting, enhance generalization, and increase accuracy. This approach is particularly important in classification tasks, where diverse models can capture different patterns and nuances in the data, leading to more reliable and precise results. Techniques such as bagging, boosting, and stacking are commonly used to create strong ensemble models that outperform individual classifiers. Today, ensemble learning methods have increasingly garnered the attention of researchers and achieved significant success in emotion identification from combined EEG data. This success is due to their ability to combine the strengths of multiple models, enhancing accuracy and robustness in detecting complex emotional states from brainwave data. Furthermore, ensemble approaches help to mitigate the variability and noise inherent in EEG signals, providing more reliable and consistent classification outcomes.

AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that aims to improve the accuracy of weak classifiers by combining them into a strong classifier. Its core principle is based on iterative refinement. By iteratively focusing on the hard-to-classify examples, AdaBoost effectively enhances the performance of the overall classifier, making it robust against overfitting and improving its generalization ability. A step-by-step AdaBoost working description is provided below (Freund and Schapire 1997).

- Training samples are selected, and each observation is assigned an initial weight.
- Weak classifiers are trained on the weighted samples.
- Misclassified instances for each weak classifier are identified.
- Update the weights of the correctly and misclassified samples.
- Finally, a strong classifier is obtained by combining the weighted predictions of the weak classifiers.

Artificial neural networks (ANN): ANN consist of interconnected layers of nodes (neurons) that process data in a manner similar to biological neural networks. ANNs are particularly powerful for pattern recognition tasks due to their ability to learn from large datasets and capture complex, non-linear relationships within the data. The structure of an ANN consists of layers of interconnected neurons: an input layer that receives the data, one or more hidden layers that process the data through weighted connections, and an output layer that delivers the final prediction or classification, as illustrated in Fig. 9. In the context of emotion detection, ANNs are utilized to identify and classify emotional states from numerous data sources, such as facial expressions, voice intonations, textual content, and physiological signals. By training on labeled datasets, ANNs can learn to recognize subtle emotional cues and generate accurate predictions, making them a valuable tool in affective computing.

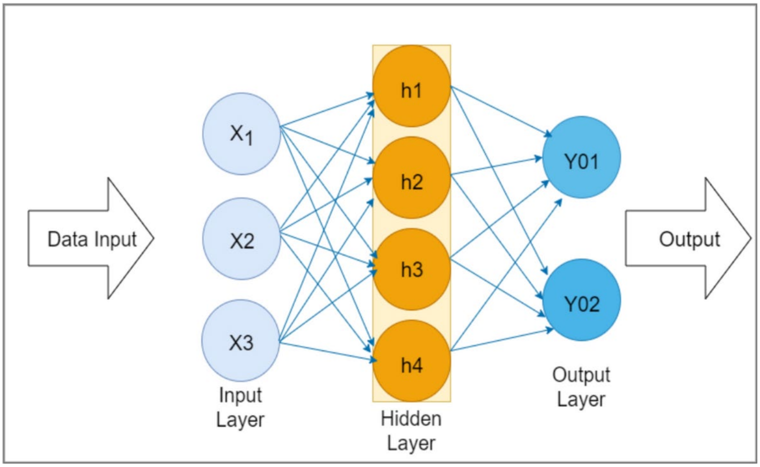


Fig. 9 Structural diagram of ANN

3.3.4.2 Deep learning models DL models are a subset of ML techniques that use multiple layers (deep architecture), enabling them to learn high-level representations directly from the input stream, for instance, sound or image. The practice of using DL models for emotion identification involves two essential steps: at the core, the model learns modality-specific features from emotional data, which is called "feature or representation learning", and then it applies a classification scheme via appropriate classifier. However, for effective multimodal classification, it is crucial to first extract modality-specific representations before performing fusion. A variety of similar architectures are introduced for this purpose, but their essential components for extracting features specific to each modality may vary (Guo et al. 2019). Here, we will present some of the typical DL models that are widely used.

Deep belief net (DBN): DBN is a form of generative graphical deep model. Structurally, the backbone network of DBN uses a restricted Boltzmann machine (RBM). The RBMs are a kind of "undirected graphical model", which consists of hidden and visible units. Notably, the RBM model lacks connections between the units within the same layer. A visible layer (or hidden layer) is constructed out of stochastic visible units (or stochastic hidden units). Typically, a DBN is formed by combining a series of given RBMs, referred to as layers. The hidden units of the initial (or bottom) RBM serve as the visible units for the subsequent RBM. As we progress through these layers, the bottom layer captures the basic structure of the actual data, while the subsequent layers extract high-level or more abstract features, as illustrated in Fig. 10. However, a multimodal DBN is advantageous for more effective modeling of complex patterns from diverse input modalities.

Autoencoder (AE): Another key architecture in the model is AEs (Rumelhart et al. 1986), typically a combination of encoder and decoder modules. Structurally, an encoder-decoder system has three primary components: input, hidden, and output processing layers. The input layer is primarily used to obtain raw input. Next, the hidden layer (intermediate units) is utilized to encode raw signal into compressed representations, whereas reconstruction of raw signals occurs at output layer. These DL models have recently been brought to the attention of the multimodal research community because of their significant potential to learn

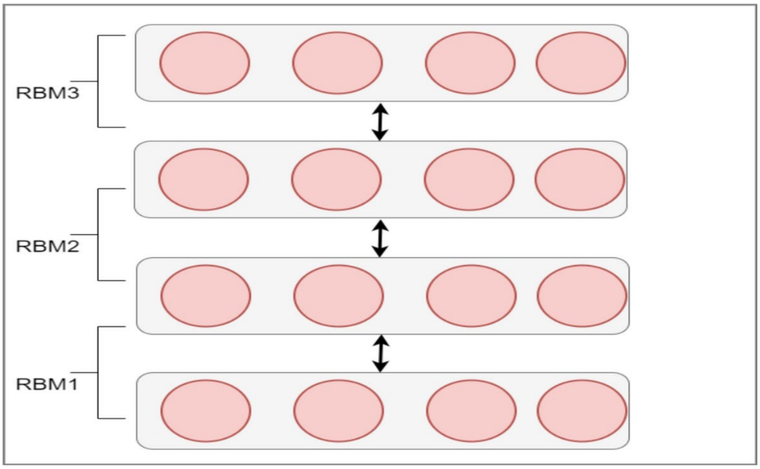


Fig. 10 Structural diagram of DBN with three RBMs (Hua et al. 2015)

"robust multimodal representations" from diverse modalities, thereby enhancing the performance of multimodal training. For instance, a multimodal encoder-decoder scheme with multiple inter-mediate layers (called stacked AE) achieves joint compression while preserving correlations among different modalities. Autoencoders (AEs) are generally implemented using various deep-learning networks. For instance, the encoder and decoder components can be realized by different neural networks. Based on the concept of encoder-decoder, Masci et al. (Masci et al. 2011) implemented the encoder with CNNs. They proposed the Convolutional Autoencoder (CAE), which adopts an unsupervised learning algorithm for encoding. In recent years, CAEs have achieved good results for unsupervised feature extraction. Figure 11 illustrates the structure of the CAE, as mentioned in (Zhou et al. 2018).

Convolution neural network (CNN): CNNs are also known as ConvNets. ConvNets are exceptionally effective at automatically and efficiently extracting spatial hierarchies of features from visual modalities, which makes them particularly powerful for tasks such as image dehazing and video recognition. To extract visual features from images, the essential CNN components are (a) the input layer, (b) linearly stacked convolution layers, and (c) the fully connected output layer. A robust CNN model can be employed to analyze EEG signals effectively. First, the EEG signals are converted into visual transformations such as spectrograms. The spectrograms serve as input to the CNN, allowing it to learn hidden representations within the data. By leveraging these hidden representations, CNN can accurately interpret and classify the underlying patterns in the EEG signals.

Recurrent neural networks (RNN): RNN is a kind of DL model that deals with variable-length serial data, which includes sound and time-series data. Mostly, RNNs map the input activations to the output and then transfer the hidden states to the output through a recurrent feedback connection. Unlike other DL models, RNN computing uses forward-pass computing and backpropagation-through-time computing. For backpropagation computing, it proceeds to process the current input and hidden states simultaneously. In addition to their effectiveness in monomodal tasks, RNNs have demonstrated their utility in various multimodal problems that necessitate modelling both long- and short-temporal dependencies within input sequences. For example, Wu et al.(Wu et al. 2020) used a stacked LSTM-based network for multimodal emotion classification.

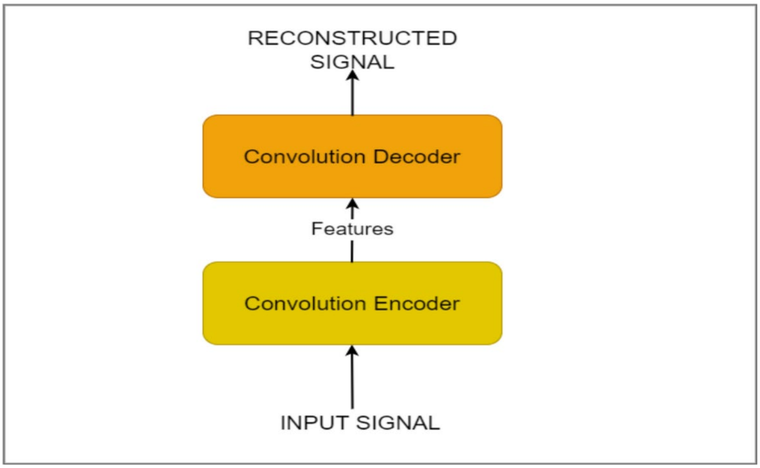


Fig. 11 Illustration of Convolution AE (CAE)

3.3.4.3 Multimodal classification schemes In Ref. (Lan et al. 2020), researchers introduce a novel deep-learning architecture termed "Deep Generalized Canonical Correlation Analysis with an Attention Mechanism" (DGCCA-AM). They emphasize the efficacy of multimodal learning in capitalizing on the complementary aspects of diverse signals, often surpassing the performance of approaches relying on single modalities. The DGCCA-AM model integrates Canonical Correlation Analysis (CCA) to manage multiple signals while incorporating an attention mechanism for adaptive fusion. This fusion mechanism dynamically adjusts weights to exploit generalized correlations across various modalities. Experimental findings reveal the effectiveness of their fusion strategy, achieving an overall classification accuracy and standard deviation of 82.11 and 2.76%, respectively, for the classification of emotions utilizing three modalities.

In (Li et al. 2019b), researchers propose a framework that amalgamates EEG and facial data for continuous emotion recognition. Employing late fusion, the framework aims to enhance the understanding of emotional states by leveraging both modalities. In a separate investigation documented by (Li et al. 2022), the author introduces a novel multimodal fusion strategy, merging EEG signals with facial expressions to classify emotions in individuals with hearing impairment. Experimental findings reveal that after fusion, the mean classification accuracy for emotion recognition reached 78.32%. Notably, this mean accuracy was surpassed by the accuracy achieved using solely facial expressions (67.90%) or EEG signals (69.43%). For the classification task, the study utilized a deepNeuralNet named CBAM_ResNet34.

In their work, Zhongjie Li et al. (Li et al. 2021) introduce a unique framework tailored for multimodal emotion classification. Their approach integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) models to effectively capture both spatial and temporal characteristics inherent in raw EEG data, thereby circumventing the need for manual feature engineering. Empirical results demonstrate their fusion technique's efficacy, yielding higher valence precisions (93.20%±2.55%) and arousal (93.18%±2.71%), respectively. On another front, Nakisa et al. (Nakisa et al. 2020) delve into the intricate task of emotion recognition utilizing EEG and BVP signals. They propose a temporal multimodal fusion method employing a deep learning algorithm to capture the intricate (non-linear) emotional correlations existing within and between modalities. The ConNet-LSTM model, devised within this framework, achieves a notable accuracy of 71.6% in classifying human emotions into dimensional quadrants. These findings underscore the potential of deep learning methodologies in unravelling the complexities of emotion recognition across diverse modalities.

The issue of poor accuracy in traditional emotion-detection methods that rely on facial expressions is addressed in (Wang et al. 2021). The paper proposes a "weighted fusion" for decision-level fusion with CNN. The experimental outcomes exhibit excellent precision, wherein the multi-scale feature extraction network attains an accuracy of 94.4% on the CK+database. In their research, Cimtay et al. (Cimtay et al. 2020) introduce a cutting-edge ER system that integrates EEG, facial expressions, and GSR data. The system is designed to address the complexity of human emotions through a hybrid-fusion approach, leveraging Convolutional Neural Networks (CNN). Impressively, this model achieves a notable accuracy of nearly 81.2% on the LUMED-2 database across emotion categories, including neutral, happy, and sad. The proposed model is especially valuable for accurately identifying emotional states, even in situations involving naturally deceptive facial expressions.

Combining features from several modalities can lead to high-dimensionality and complexity in learning processes for many machine learning algorithms. To address these challenges (Ghoniem et al. 2019), introduced a novel approach using hybrid fuzzy-evolutionary computation methodologies for feature learning and dimensionality reduction. The proposed system uses multi-hierarchical neural nets to fuse EEG and Speech modalities. Public databases MAHNOB-HCI and SAVEE were used to test the model and obtained 98.21 and 98.26% accuracies respectively. Tables 9 and 10 compare classifiers, databases used, and performance measures in terms of accuracy, along with fusion approaches.

Based on Tables 9 and 10, it is clear that not all (21 reports) research papers focused on specific types and number of emotion categories. Only two studies reported using a dimensional model (valence and arousal scale), while the remaining 19 papers used discrete emotions. The number of emotion categories reported varied, with the least being three and the highest being eight. The most frequently analyzed emotions were happiness, fear, and sadness (57%). Out of 80 emotional tags that appear in the studies, happy (20%) and sad (18.75%) are commonly analyzed. However, neutral emotion was included in almost all studies. The tables (Tables 9 and 10) also show the classifiers used and their corresponding performance. The classifiers are ranked by popularity as follows: LSTM, CNN, and SVM (each at 23.81%), RF, MLP, DT, and ELM (each at 4.76%), LSTM (4.2%). SVM, LSTM and CNN were the most commonly used methods for multimodal emotion identification, achieving high performance with maximum accuracies of 89.90% (SVM) LSTM (99.30%, and 94.44% (CNN). However, other classification schemes for multimodal emotion detection also performed exceptionally well, surpassing the 90% margin. For example, the FCM-GA-NN model achieved 98.53%. One study even achieved a remarkable 99.3% accuracy in predicting six emotions during real-time monitoring. This suggests that alternative classification methods can also yield strong results.

The reported accuracies, however, only highlight the best-performing models and do not reflect a consensus on emotion classification performance. The effectiveness of each model varies depending on factors such as the database used, feature selection, number of modalities, and specific emotion categories considered in the classification scheme—even the

Table 9 Comparison of Shallow-models used for multi-modal emotion learning and its performance

References	Dataset	Classifier	Emotion classes	Acc (%)
Wang et al. (2022)	MED4	ELM	Hap, Ang, Neu, Sad	89.65%
Chaparro et al. (2018)	MAHNOB	Random Forest	Hap, Dis, Neu, Ang, Surp, Anx, Fear, Sad	97.70%
Huang et al. (2017)	Private data	SVM	Hap, Neu, Fear, Sad	82.75%
Su et al. (2019)	Private data	SVM	Arousal, Valence	a:72.80 v:69.3%
Guo et al. (2020)	DEAP	SVC	Hap, Ang, Fear, Sad	89.90%
Zhang (2020)	Private data	LIBSVM	Hap, Neu, Fear, Sad	85.71%
Zhao et al. (2019)	SEED V	SVM	Hap, Dis, Neu, Fear, Sad	79.70%
Ghoniem et al. (2019)	Private data	FCM-GA-NN	Hap, Anx, Dis, Neu, Surp. Fear,Sad	98.53%
Cimtay et al. (2020)	Lumed2	Decision tree	Hap, Neu, Sad	81.2%
Zhao et al. (2021)	SEED	MLP	Postive, Negative, Neu	86.01%

ACC accuracy, *Hap* happy, *Dis* disgust, *Neu* neutral, *Anx* anxiety, *Surp* surprise, *Ang* anger

Table 10 Comparison of DL-models used for multi-modal emotion learning and its performance

References	Dataset	Classifier	Emotion classes	Acc (%)
Li et al. (2022)	Private data	CBAM-Resnet34	Hap, Fear, Sad, Calmness	78.32%
Zhang et al. (2021b)	MAHNOB	HFCNN	Arousal, Valence	89.00%
Mutawa and Hassouneh (2024)	Private data	LSTM	Hap, Dis, Ang, Sup, Fear, Sad	99.30%
Jaswal and Dhingra (2023)	Private data	CNN	Hap, Neu, Sad	94.44%
Wang et al. (2021)	Private data	CNN	Hap, Ang, Neu, Sup, Scared	92.60%
Lu et al. (2021)	MAHNOB	Stacked LSTM	Hap, Ang, Dis, Sup, Fear, Sad	89%
Nakisa et al. (2020)	Private data	LSTM	HA-negative, HA-positive, LA-positive, LA-negative	71.6%
Lan et al. (2020)	SEED V	LSTM	Hap, Dis, Neu, Fear, Sad	82.11%
Ma et al. (2019)	DEAP	LSTM	Arousal, Valence	a:92.87%, v:92.30%
Muhammad et al. (2023)	MAHNOB	CNN	Hap, Neu, Sad	93.86%
Fu et al. (2023)	SEED-IV	CNN	Hap, Neu, Fear, Sad	87.32%

ACC accuracy, *Hap* happy, *Dis* disgust, *Neu* neutral, *Anx* anxiety, *Surp* surprise, *Ang* anger, *HA* high arousal, *LA* low arousal, *LV* low valence, *HV* high valence, *a* arousal, *v* valence

selection of emotion classes in combination. For instance, the combination of happy and sad in one study may not yield the same performance as a combination of sad and angry. Consequently, it is challenging to compare the actual classification performance across various models directly. Furthermore, the number of emotion categories also significantly affects performance. For instance, in most reported studies, three-class predictions performed better than four-class predictions. The lack of clear evidence regarding optimal classification performance highlights the need for further research to determine the most effective combination of features and physiological signals for emotion detection. Such efforts will advance emotion recognition technology and ensure more consistent and reliable performance.

Evaluating emotion recognition models requires considering experimental settings, including subject-dependent and subject-independent experiments, and the number of emotion categories used. These factors can influence classification outcomes, leading to varied results. Out of the total reports, 10 (50%) conducted subject-independent experiments, while 5 (25%) employed subject-dependent experiments. The remaining 25% did not specify the type of experimental setting used. The number of participants in the studies ranged from a minimum of 12 to a maximum of 35 (Lu et al. 2021), with an average of 32 participants reported in many studies.

4 Open issues and discussion

4.1 Open issues and further research

EEG-based multimodal emotion recognition has made significant strides, but several open issues continue to challenge researchers in this field. Addressing these issues is crucial to enriching the accuracy and utility of ER systems in various domains. The following points outline a set of open issues and further research within the realm of multimodal emotion detection acknowledged during our comprehensive review.

4.1.1 Multiple feature collection

MEC systems commonly initiate by deriving low or high-level attributes after pretreatment procedures (such as preprocessing blocks). This field of research is still in its developmental stages, with no unanimous agreement on the optimal features to employ for particular tasks. In numerous investigations, practitioners simply aggregate a multitude of attributes to construct a matrix of numbers for classification purposes. If the chosen feature collection fails to accurately capture the inherent emotional cues of the unprocessed signal, the system's performance will suffer, irrespective of the employed algorithm (Lu et al. 2015). Consequently, selecting unique information is pivotal in enhancing overall system efficacy.

4.1.2 Data fusion criteria

In multimodal analysis, the task of selecting the optimal fusion scheme is indeed a complex research endeavor. Presently, decisions based on equal weighting or voting mechanisms fail to guarantee the precision of a system. Furthermore, a significant number of studies tend to overlook the distinct and interconnected emotional cues present during early fusion (Zhang et al. 2021a). Due to the nature of a late-level fusion approach, where modalities are integrated after independent processing, the impressive accuracy attained by a specific modality might be compromised by the relatively poor accuracy of another modality. Conversely, deep-leaning fusion schemes, which are in their initial stages, pose ongoing challenges (Gao et al. 2020a). Consequently, the design or selection of an optimal fusion strategy is a crucial concern for practical systems.

4.1.3 Constrained dataset with labels

A scarce set of classified data poses significant challenges in machine learning endeavors. The foremost concern is compromised model performance, resulting in suboptimal outcomes. To mitigate the issue, practitioners often resort to co-learning techniques. Co-learning presents a research challenge that involves the necessity of transferring domain knowledge between different modalities utilized in a production setting. These challenges become especially significant (Baltrušaitis et al. 2018) in applying various deep learning algorithms. Several extensions have been proposed to tackle co-leaning issues (Rahate et al. 2022).

4.1.4 Modality blend

Much of the research conducted thus far has assessed various hybrid fusion approaches that combine biosensor feedback with external effective measures. However, these studies are also constrained by the scarcity of combinations of selected modalities within real-life contexts. Considerable research efforts may be required to experiment with diverse modality combinations (e.g., using the EEG modality with speech and linguistic cues), aiming to achieve more human-like behavior, which stands as an important issue for the future. Furthermore, achieving a balanced blend requires robust methods to mitigate the impact of missing or class imbalance data from any modality (Fan et al. 2024). Another critical challenge in modality combination lies in effectively modeling the interactions between complementary modalities (Cohen et al. 2024). In addition, effective blending requires addressing challenges such as variability in signal quality and ensuring the explainability of fusion methods. Validating the incorporation of contextual factors could further enhance the adaptability of modality blends in global contexts (Izard 1994).

4.1.5 Multi-environment datasets

Having access to annotated multimodal datasets is essential for the advancement of emotion classification algorithms. Traditional emotion recognition systems are limited by their controlled settings, resulting in models that perform well in similar conditions but fail to generalize to real-life scenarios (Siddharth et al. 2019). Multi-environment datasets, which incorporate diverse environmental factors like lighting conditions and locations, enable the development of more robust and generalizable models. To represent a variety of conditions, data can be collected from public places, workplaces, homes, and outdoors (Fleury et al. 2010). While numerous multimodal datasets are available, there is still a lack of such datasets that incorporate physiological information gathered from significant human subjects. Another promising area for further research is the gathering of data under multi-environmental (both natural and anechoic) conditions (Wang et al. 2022).

4.1.6 Domain generalization issues

Generalizability issues are inherent when data is collected from non-stationary signals like EEG, causing data shift problems. Due to the transient/stochastic nature of EEG signals, achieving “cross-subject and cross-session” generalization remains a significant challenge. To address this issue, much of the research conducted has utilized various domain adaptation and domain generalization techniques. However, several issues still exist regarding inter and intra-subjective generalization methods tested in multi-modal studies focusing on EEG data. In Ref. (Apicella et al. 2024), the authors identified several factors contributing to low generalization performance related to “cross-subject and cross-session” problems. They highlighted that various sources of uncertainty can lead to generalization problems, including (i) measuring, which pertains to the specific features of the EEG signals being measured; (ii) instrumentation, referring to the different devices and setups used for data acquisition; and (iii) environmental factors that can influence the signals collected. Understanding these challenges is crucial for developing effective solutions to improve model generalization in emotion classification.

4.2 Discussion

Drawing inspiration from existing literature, it is evident that analyzing user emotions frequently involves utilizing audiovisual data within multimodal networks. These networks often employ fusion techniques to capitalize on information derived from speech signals and temporal cues from video signals. Research demonstrates that integrating audiovisual data significantly improves recognition accuracy. Nevertheless, much of multimodal research has yet to explore the amalgamation of EEG physiological data with behavioral modalities such as audio or visual cues. Combining brain wave data obtained from EEG signals with the behavioral manifestations of emotions offers an intriguing avenue for investigation.

With this motivation, an extensive literature review is undertaken, focusing specifically on examining the role of EEG data in multimodal emotion identification. Tables 4 and 5 present an overview of EEG-based multimodal fusion methodologies employed for emotion identification and classification.

However, implementing a multimodal approach for human emotion analysis involves effectively integrating physiological signals with external behavioral cues to identify human emotions accurately. Additional research and innovative approaches are required to overcome this obstacle.

When the inputs from different modalities are well-aligned in time, a feature-based fusion approach is deemed appropriate. Also, feature-level fusion is suitable when the study modalities provide complementary information and are relatively low-dimensional. Notably, in cases where the modalities possess varying levels of importance, feature weighting may become necessary as an additional pre-processing step. This technique ensures that every modality contributes to the conclusive decision-making process proportionally to its significance in the overall emotion recognition task. By assigning appropriate weights to features from different modalities, the fusion process can be optimized to reflect the significance of each source of information, thereby enhancing the accuracy and effectiveness of the multimodal ER system. Decision-level fusion is considered appropriate when the modalities provide independent or redundant information and when different levels of reliability or confidence exist among them, allowing for a weighted combination. In practice, the choice between decision and feature-level fusion is determined by aspects such as the specific utility, the nature of the data, the availability of modalities, and the desired performance goals. Moreover, in specific scenarios, utilizing a combination of both strategies can be beneficial by employing a hybrid approach. This approach leverages the advantages of feature-level fusion and decision-level fusion techniques. By integrating features from multiple modalities at an early stage and then combining the outputs of different classifiers or fusion algorithms at a later stage, the hybrid approach aims to capitalize on the strengths of each strategy. This allows for a more comprehensive and robust multimodal ER system that can effectively adapt to diverse data characteristics and task requirements.

Subsequently, the tasks primarily focusing on the selection of merged modalities were categorized into four main groups: facial (34%), speech (11%), peripheral physiological signals (31%), and other modalities (23%). The category labeled as "others" encompassed studies that chose more than two modalities or visual data. Interestingly, a significant proportion of studies (34%) opted to utilize facial data, as illustrated in Fig. 7. Among these, 11% (4 articles) employed conventional and 20% (7 articles) deep learning fusion methods for facial data.

Following facial data, PPS signals were the next prevalent modality (31%), with 11% (4 articles) employing conventional fusion methods and 14% (5 articles) using deep learning fusion. Lastly, another subset of studies integrated speech data. Within this subset, most (2 articles) applied deep learning techniques for feature-level fusion. Notably, no articles mentioned the combination of speech signals with attention networks or autoencoders. Finally, Table 11 summarizes significant studies on diverse multimodal databases.

Table 11 Comparative table: summary of significant studies on multimodal datasets

Datasets used	References	Emotional states	Accuracy (%)
MED4	Wang et al. (2022)	Happy, sad, angry, neutral	89.65%
CASIA	Guo et al. (2020)	Angry, happy, sad, fear	95.16%
SEED IV	Zhao et al. (2019)	Happy, fear, sad, disgust, and neutral	79.71%
DEAP	Zhang et al. (2021b)	Valence, arousal	valence 84.7% arousal 83.2%
MAHNOB-HCI	Xing et al. (2019)	Anger, fear, disgust	89%
PME4	Chen et al. (2022)	Happiness, sadness, surprise, neutral	–
PhyMER	Pant et al. (2023)	Disgust, happy, angry, neutral, sad, surprise, fear	NA
K-EmoCon	Park et al. (2020)	Valence, arousal	–

‘–’ refers no research found for multimodal studies, *NA* not applicable

5 Conclusions

Our review highlights that leveraging features from both EEG and behavioral modalities is a highly effective approach for uncovering human emotions. It is a valuable combination for applications in medical-IOT and education systems. Integrating EEG and behavioral modalities through sensor fusion enables greater context awareness, allowing the IOT-enabled healthcare system to detect and respond to unusual expressions or behavioral patterns more effectively. This makes sensor fusion a powerful tool for advancing medical IoT applications and providing a deeper understanding of patient states. Looking ahead, integrating these advanced methods and machine learning applications for emotion assessment holds excellent potential for diverse fields such as psychotherapy and e-learning systems. Engaging professionals in the process is essential to create systems that can effectively be used in domains such as psychotherapy and e-learning. These intelligent systems offer the potential for more personalized therapeutic interventions and adaptive educational experiences, thereby revolutionizing computational intelligence approaches and paving the way for context-aware solutions that cater to diverse healthcare and education needs.

We provided a comprehensive overview of multimodal systems, exploring various approaches in the field. We categorized existing emotion models into those focused on physical concepts and those based on physiological concepts. We also highlighted several challenges faced in multimodal emotion detection. In our review, we examined EEG signals and their various use cases, discussing modern methods for multimodal fusion that have emerged recently. Furthermore, affective computing necessitates multimodal databases for training models based on machine learning and deep learning. We considered established and new datasets, highlighting several primarily utilized for emotion detection.

Furthermore, we discussed the key components of the EEG-based multimodal emotion classification pipeline. This general approach encompasses several steps: modality selection, feature extraction, fusion, and classification. We investigated a range of parameters and features crucial for accurate emotion recognition. Additionally, we explained various classification strategies, comparing machine learning and deep learning methods from multiple perspectives concerning emotional feature fusion in EEG-based research. Based on our review, there does not appear to be a single fusion or classification strategy that universally excels for all applications in multimodal emotion recognition. The optimal choice must relate specifically to the fusion algorithms and tasks involved. Therefore, exploring and investigating multiple fusion strategies and synchronized data modalities is advisable to evaluate the proposed methods' effectiveness comprehensively. Before settling on a strategy that yields satisfactory performance for a given task, one should compare various features and techniques to ensure the best results.

Despite a decade of research, many studies have focused on static environments with controlled stimuli, which limit their applicability to real-world scenarios. Most reviewed studies rely heavily on features derived from external manifestations, such as facial expressions or physiological signals, which often lack specificity and generalizability. As a result, these methods are not well-suited for accurately investigating emotions in real-time or inferring annotated emotions reliably. To extend this research, the next step should involve adopting a multi-perspective analysis of human emotions. The approach can be further investigated and validated using multi-perspective datasets from naturalistic conversational sources, such as the E-EmoCon database, which is publicly available. Future research in multimodal emotion recognition should focus on enhancing the effectiveness of fusion networks. It is essential to address the issue of degraded performance that often arises from the current approaches.

Moreover, there exists considerable ambiguity concerning the reported efficacy of fusion networks, which can be profoundly affected by the chosen features and implemented classifiers. Consequently, further studies should establish standardized metrics and reporting methodologies to yield more precise insights into the effectiveness of diverse fusion strategies. Lastly, integrating explainable AI into these systems will be crucial, as it can enhance transparency and facilitate a better understanding of the decision-making processes within fusion networks. Overall, this review underscores the significance of multimodal methodologies in emotion identification and emphasizes the need for sustained exploration and innovation to unlock its full potential for practical utility. Given the promising potential of integrating EEG data with behavioral modalities like audio and visual cues, future research should delve into more comprehensive EEG-based multimodal emotion detection systems.

- Future research may explore fusion techniques for disparate modalities in multimodal-based emotion identification systems.
- The selection of modalities beyond facial and speech, such as text and electrophysiological data, should be more diverse.
- More comprehensive information about feature selection methods may be included.
- Explore benchmark datasets encompassing a unique combination of modalities, encouraging and fostering comparisons between different approaches.
- Methods for designing ML and DL classifiers may be further included in subsequent investigations.

Author contributions R.P and U.S. conceived and presented the idea; R.P performed the analysis, drafted the manuscript and designed the figures and table; U.S provided a research perspective and helped shape the full manuscript. All authors reviewed the manuscript.

Data availability Data availability does not apply to this research.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Abadi MK, Subramanian R, Kia SM et al (2015) DECAF: MEG-based multimodal database for decoding affective physiological responses. IEEE Trans Affect Comput 6:209–222. https://doi.org/10.1109/TAFFC.2015.2392932

AbdelAal MA, Alsawy AA, Hefny HA (2018) EEG-based emotion recognition using a wrapper-based feature selection method. In: Hassanien AE, Shaalan K, Gaber T, Tolba MF (eds) Proceedings of the international conference on advanced intelligent systems and informatics. Springer, Cham, pp 247–256

Acharya UR, Vinitha Sree S, Swapna G et al (2013) Automated EEG analysis of epilepsy: a review. Knowl-Based Syst 45:147–165. https://doi.org/10.1016/j.knosys.2013.02.014

Aftanas LI, Lotova NV, Koshkarov VI et al (1997) Non-linear analysis of emotion EEG: calculation of Kolmogorov entropy and the principal Lyapunov exponent. Neurosci Lett 226:13–16. https://doi.org/10.1016/S0304-3940(97)00232-2

Ahmed N, Al Aghbari Z, Girija S (2023) A systematic survey on multimodal emotion recognition using learning algorithms. Intell Syst Appl 17:200171

Ai Q, Liu Q, Meng W, Xie SQ (2018) EEG-based brain intention recognition. In: Ai Q, Liu Q, Meng W, Xie SQ (eds) Advanced rehabilitative technology. Academic Press, pp 135–166

Allen J (2007) Photoplethysmography and its application in clinical physiological measurement. Physiol Meas 28:R1. https://doi.org/10.1088/0967-3334/28/3/R01

Álvarez-Jiménez M, Calle-Jimenez T, Hernández-Álvarez M (2024) A comprehensive evaluation of features and simple machine learning algorithms for electroencephalographic-based emotion recognition. Appl Sci 14:2228. https://doi.org/10.3390/app14062228

Apicella A, Arpaia P, D'Errico G et al (2024) Toward cross-subject and cross-session generalization in EEG-based emotion recognition: systematic review, taxonomy, and methods. Neurocomputing 604:128354. https://doi.org/10.1016/j.neucom.2024.128354

Atkinson J, Campos D (2016) Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. Expert Syst Appl 47:35–41. https://doi.org/10.1016/j.eswa.2015.10.049

Babiloni C, Pizzella V, Gratta CD et al (2009) Fundamentals of electroencefalography, magnetoencefalography, and functional magnetic resonance imaging. International review of neurobiology. Academic Press, pp 67–80

Bahador N, Jokelainen J, Mustola S, Kortelainen J (2021) Multimodal spatio-temporal-spectral fusion for deep learning applications in physiological time series processing: a case study in monitoring the depth of anesthesia. Inform Fusion 73:125–143. https://doi.org/10.1016/j.inffus.2021.03.001

Bahreini K, Nadolski R, Westera W (2016) Towards multimodal emotion recognition in e-learning environments. Interact Learn Environ 24:590–605. https://doi.org/10.1080/10494820.2014.908927

Bakhshi A, Chalup S (2021) Multimodal emotion recognition based on speech and physiological signals using deep neural networks. In: Del Bimbo A, Cucchiara R, Sclaroff S et al (eds) Pattern recognition. ICPR international workshops and challenges. Springer, Cham, pp 289–300

Bakker I, van der Voordt T, Vink P, de Boon J (2014) Pleasure, arousal, dominance: mehrabian and Russell revisited. Curr Psychol 33:405–421. https://doi.org/10.1007/s12144-014-9219-4

Balconi M, Lucchiari C (2008) Consciousness and arousal effects on emotional face processing as revealed by brain oscillations. A gamma band analysis. Int J Psychophysiol 67:41–46. https://doi.org/10.1016/j.ijpsycho.2007.10.002

Ball T, Kern M, Mutschler I et al (2009) Signal quality of simultaneously recorded invasive and non-invasive EEG. Neuroimage 46:708–716. https://doi.org/10.1016/j.neuroimage.2009.02.028

Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41:423–443

Bandt C, Pompe B (2002) Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 88:174102. https://doi.org/10.1103/PhysRevLett.88.174102

Barra S, Casanova A, Fraschini M, Nappi M (2017) Fusion of physiological measures for multimodal biometric systems. Multimed Tools Appl 76:4835–4847. https://doi.org/10.1007/s11042-016-3796-1

Berger H (1929) Über das Elektrenkephalogramm des Menschen. Archiv f Psychiatrie 87:527–570. https://doi.org/10.1007/BF01797193

Bezdek JC (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Springer US, Boston, MA

Bontchev B (2016) Adaptation in affective video games: a literature review. Cybernet Inform Technol 16:3–34

Bota PJ, Wang C, Fred AL, Da Silva HP (2019) A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. IEEE Access 7:140990–141020

Brás S, Ferreira JHT, Soares SC, Pinho AJ (2018) Biometric and emotion identification: an ECG compression based method. Front Psychol 9:467

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Broek E van den (2011) Affective signal processing (ASP): unraveling the mystery of emotions. https://doi.org/10.3990/1.9789036532433

Bromfield EB, Cavazos JE, Sirven JI (2006) Slide 14, [10/20 System of EEG Electrode Placement]. https://www.ncbi.nlm.nih.gov/books/NBK2510/figure/A44/. Accessed 3 Jun 2024

Bruns A (2004) Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches? J Neurosci Methods 137:321–332. https://doi.org/10.1016/j.jneumeth.2004.03.002

Brynolfsson J (2012) Time frequency analysis of EEG measured when performing the flanker task

Burle B, Spieser L, Roger C et al (2015) Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. Int J Psychophysiol 97:210–220. https://doi.org/10.1016/j.ijpsycho.2015.05.004

Chanel G, Ansari-Asl K, Pun T (2007) Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In: 2007 IEEE International conference on systems, man and cybernetics. pp 2662–2667

Chaparro V, Gomez A, Salgado A, et al (2018) Emotion recognition from EEG and facial expressions: a multimodal approach. In: 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp 530–533

Chen J, Ro T, Zhu Z (2022) Emotion recognition with audio, video, EEG, and EMG: a dataset and baseline approaches. IEEE Access 10:13229–13242. https://doi.org/10.1109/ACCESS.2022.3146729

Chen W, Cai Y, Li A et al (2023) EEG feature selection method based on maximum information coefficient and quantum particle swarm. Sci Rep 13:14515. https://doi.org/10.1038/s41598-023-41682-5

Choi DY, Kim D-H, Song BC (2020) Multimodal attention network for continuous-time emotion recognition using video and EEG signals. IEEE Access 8:203814–203826

Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S (2020) Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access 8:168865–168878. https://doi.org/10.1109/ACCESS.2020.3023871

Cohen D (1968) Magnetoencephalography: evidence of magnetic fields produced by alpha-rhythm currents. Science 161:784–786. https://doi.org/10.1126/science.161.3843.784

Cohen MX (2019) A better way to define and describe Morlet wavelets for time-frequency analysis. Neuroimage 199:81–86. https://doi.org/10.1016/j.neuroimage.2019.05.048

Cohen O, Hazan G, Gannot S (2024) Multi-microphone and multi-modal emotion recognition in reverberant environment. https://doi.org/10.48550/arXiv.2409.09545

Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16:031001. https://doi.org/10.1088/1741-2552/ab0ab5

Dadebayev D, Goh WW, Tan EX (2022) EEG-based emotion recognition: review of commercial EEG devices and machine learning techniques. J King Saud Univ Comput Inform Sci 34:4385–4401. https://doi.org/10.1016/j.jksuci.2021.03.009

Dimberg U, Andréasson P, Thunberg M (2011) Emotional empathy and facial reactions to facial expressions. J Psychophysiol 25:26–31. https://doi.org/10.1027/0269-8803/a000029

Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos JC et al (2020) A machine learning model for emotion recognition from physiological signals. Biomed Signal Process Control 55:101646. https://doi.org/10.1016/j.bspc.2019.101646

Duan R-N, Zhu J-Y, Lu B-L (2013) Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS conference on neural engineering (NER). IEEE, pp 81–84

Duwenbeck R, Kirchner EA (2024) Auditive emotion recognition for empathic AI-Assistants. Künstl Intell. https://doi.org/10.1007/s13218-023-00828-3

Dwijayanti S, Iqbal M, Suprapto BY (2022) Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network. IEEE Access 10:89876–89886

Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200

Erwin RJ, Gur RC, Gur RE et al (1992) Facial emotion discrimination: I. Task construction and behavioral findings in normal subjects. Psychiatry Res 42:231–240. https://doi.org/10.1016/0165-1781(92)90115-J

Eysenck MW, Ellis AW, Hunt EB, Johnson-Laird PNE (1994) The Blackwell dictionary of cognitive psychology. Basil Blackwell

Fan Q, Li Y, Xin Y, et al (2024) Leveraging contrastive learning and self-training for multimodal emotion recognition with limited labeled samples. In: Proceedings of the 2nd international workshop on multimodal and responsible affective computing. association for computing machinery, New York, pp 72–77

Ferguson HJ, Wimmer L (2023) A psychological exploration of empathy. In: Conversations on Empathy. Routledge, pp 60–77

Fleury A, Vacher M, Noury N (2010) SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. IEEE Trans Inf Technol Biomed 14:274–283. https://doi.org/10.1109/TITB.2009.2037317

Franaszczuk PJ, Bergey GK, Durka PJ, Eisenberg HM (1998) Time–frequency analysis using the matching pursuit algorithm applied to seizures originating from the mesial temporal lobe. Electroencephalogr Clin Neurophysiol 106:513–521. https://doi.org/10.1016/S0013-4694(98)00024-8

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139. https://doi.org/10.1006/jcss.1997.1504

Fu B, Gu C, Fu M et al (2023) A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. Front Neurosci. https://doi.org/10.3389/fnins.2023.1234162

Fu Z, Zhang B, He X et al (2022) Emotion recognition based on multi-modal physiological signals and transfer learning. Front Neurosci 16:1000716. https://doi.org/10.3389/fnins.2022.1000716

Gandhi T, Panigrahi B, Anand S (2011) A comparative study of wavelet families for EEG signal classification. Neurocomputing 74:3051–3057. https://doi.org/10.1016/j.neucom.2011.04.029

Gao J, Li P, Chen Z, Zhang J (2020a) A survey on deep learning for multimodal data fusion. Neural Comput 32:829–864. https://doi.org/10.1162/neco_a_01273

Gao Q, Wang C, Wang Z et al (2020b) EEG based emotion recognition using fusion feature extraction method. Multimed Tools Appl 79:27057–27074. https://doi.org/10.1007/s11042-020-09354-y

Gatti E, Calzolari E, Maggioni E, Obrist M (2018) Emotional ratings and skin conductance response to visual, auditory and haptic stimuli. Sci Data 5:180120. https://doi.org/10.1038/sdata.2018.120

Ghoniem RM, Algarni AD, Shaalan K (2019) Multi-modal emotion aware system based on fusion of speech and brain information. Information 10:239. https://doi.org/10.3390/info10070239

Giannakakis G, Grigoriadis D, Giannakaki K et al (2019) Review on psychological stress detection using biosignals. IEEE Trans Affect Comput 13:440–460

Gjoreski M, Kiprijanovska I, Stankoski S et al (2022) Facial EMG sensing for monitoring affect using a wearable device. Sci Rep 12:16876. https://doi.org/10.1038/s41598-022-21456-1

Gong L, Li M, Zhang T, Chen W (2023) EEG emotion recognition using attention-based convolutional transformer neural network. Biomed Signal Process Control 84:104835. https://doi.org/10.1016/j.bspc.2023.104835

Goshvarpour A, Goshvarpour A (2018) Poincaré's section analysis for PPG-based automatic emotion recognition. Chaos, Solitons Fractals 114:400–407. https://doi.org/10.1016/j.chaos.2018.07.035

Goshvarpour A, Goshvarpour A (2020) The potential of photoplethysmogram and galvanic skin response in emotion recognition using nonlinear features. Phys Eng Sci Med 43:119–134. https://doi.org/10.1007/s13246-019-00825-7

Guo H, Jiang N, Shao D (2020) Research on multi-modal emotion recognition based on speech, EEG and ECG signals. In: Qian J, Liu H, Cao J, Zhou D (eds) Robotics and rehabilitation intelligence. Springer, Singapore, pp 272–288

Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: a survey. IEEE Access 7:63373–63394

Harris FJ (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE 66:51–83. https://doi.org/10.1109/PROC.1978.10837

Hasnul MA, Aziz NAA, Alelyani S et al (2021) Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. Sensors 21:5015. https://doi.org/10.3390/s21155015

He B, Liu Z (2008) Multimodal functional neuroimaging: integrating functional MRI and EEG/MEG. IEEE Rev Biomed Eng 1:23–40. https://doi.org/10.1109/RBME.2008.2008233

He Z, Li Z, Yang F et al (2020) Advances in multimodal emotion recognition based on brain-computer interfaces. Brain Sci 10:687. https://doi.org/10.3390/brainsci10100687

Hjorth B (1970) EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol 29:306–310

Hua Y, Guo J, Zhao H (2015) Deep belief networks and deep learning. In: Proceedings of 2015 international conference on intelligent computing and internet of things. pp 1–4

Huang Y, Yang J, Liao P, Pan J (2017) Fusion of facial expressions and EEG for multimodal emotion recognition. Comput Intell Neurosci 2017:e2107451. https://doi.org/10.1155/2017/2107451

Hwang S, Hong K, Son G, Byun H (2020) Learning CNN features from DE features for EEG-based emotion recognition. Pattern Anal Applic 23:1323–1335. https://doi.org/10.1007/s10044-019-00860-w

Izard CE (1994) Innate and universal facial expressions: evidence from developmental and cross-cultural research. Psychol Bull 115:288–299. https://doi.org/10.1037/0033-2909.115.2.288

Jaswal RA, Dhingra S (2023) Empirical analysis of multiple modalities for emotion recognition using convolutional neural network. Meas: Sens 26:100716. https://doi.org/10.1016/j.measen.2023.100716

Katsigiannis S, Ramzan N (2018) DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform 22:98–107. https://doi.org/10.1109/JBHI.2017.2688239

Kawala-Sterniuk A, Browarska N, Al-Bakri A et al (2021) Summary of over fifty years with brain-computer interfaces—a review. Brain Sci 11:43. https://doi.org/10.3390/brainsci11010043

Khalili Z, Moradi MH (2009) Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of EEG. In: 2009 International joint conference on neural networks. IEEE, pp 1571–1575

Kheirkhah M, Brodoehl S, Leistritz L et al (2021) Automated emotion classification in the early stages of cortical processing: an MEG study. Artif Intell Med 115:102063. https://doi.org/10.1016/j.artmed.2021.102063

Kim J, André E, Rehm M, et al (2005) Integrating information from speech and physiological signals to achieve emotional sensitivity. In: Ninth European conference on speech communication and technology

Kirar JS, Agrawal RK (2016) Optimal spatio-spectral variable size subbands filter for motor imagery brain computer interface. Procedia Comput Sci 84:14–21. https://doi.org/10.1016/j.procs.2016.04.060

Klem GH, Lüders HO, Jasper HH, Elger C (1999) The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 52:3–6

Koelstra S, Muhl C, Soleymani M et al (2012) DEAP: a database for emotion analysis using physiological signals. IEEE Trans Affect Comput 3:18–31. https://doi.org/10.1109/T-AFFC.2011.15

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Koles ZJ, Lazar MS, Zhou SZ (1990) Spatial patterns underlying population differences in the background EEG. Brain Topogr 2:275–284. https://doi.org/10.1007/BF01129656

Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. In: Bergadano F, De Raedt L (eds) Machine learning: ECML-94. Springer, Berlin, Heidelberg, pp 171–182

Kumar S, Yadava M, Roy PP (2019) Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. Inform Fusion 52:41–52. https://doi.org/10.1016/j.inffus.2018.11.001

Lan Y-T, Liu W, Lu B-L (2020) Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In: 2020 International joint conference on neural networks (IJCNN). IEEE, pp 1–6

Lee MS, Lee YK, Pae DS et al (2019) Fast emotion recognition based on single pulse PPG signal with convolutional neural network. Appl Sci 9:3355. https://doi.org/10.3390/app9163355

Li C, Li P, Jiang L, et al (2019a) Emotion recognition with the feature extracted from brain networks. In: 2019 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA). pp 1–4

Li C, Li P, Zhang Y et al (2024a) Effective emotion recognition by learning discriminative graph topologies in EEG brain networks. IEEE Trans Neural Netw Learn Syst 35:10258–10272. https://doi.org/10.1109/TNNLS.2023.3238519

Li C, Tang T, Pan Y et al (2024b) An efficient graph learning system for emotion recognition inspired by the cognitive prior graph of EEG brain network. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2024.3405663

Li D, Liu J, Yang Y et al (2022) Emotion recognition of subjects with hearing impairment based on fusion of facial expression and EEG topographic map. IEEE Trans Neural Syst Rehabil Eng 31:437–445

Li D, Wang Z, Wang C et al (2019b) The fusion of electroencephalography and facial expression for continuous emotion recognition. IEEE Access 7:155724–155736. https://doi.org/10.1109/ACCESS.2019.2949707

Li P, Liu H, Si Y et al (2019c) EEG based emotion recognition by combining functional connectivity network and local activations. IEEE Trans Biomed Eng 66:2869–2881. https://doi.org/10.1109/TBME.2019.2897651

Li Z, Zhang G, Dang J, et al (2021) Multi-modal emotion recognition based on deep learning Of EEG and audio signals. In: 2021 International joint conference on neural networks (IJCNN). pp 1–6

Lin Y-P, Wang C-H, Jung T-P et al (2010) EEG-based emotion recognition in music listening. IEEE Trans Biomed Eng 57:1798–1806

Liu K, Li Y, Xu N, Natarajan P (2018) Learn to combine modalities in multimodal deep learning. https://doi.org/10.48550/arXiv.1805.11730

Liu W, Zheng W-L, Lu B-L (2016) Emotion recognition using multimodal deep learning. In: Hirose A, Ozawa S, Doya K et al (eds) Neural information processing. Springer, Cham, pp 521–529

Liu Z-T, Hu S-J, She J et al (2023) Electroencephalogram emotion recognition using combined features in variational mode decomposition domain. IEEE Trans Cognit Dev Syst 15:1595–1604. https://doi.org/10.1109/TCDS.2022.3233858

Loveys K, Sagar M, Billinghurst M et al (2022) Exploring empathy with digital humans. In: 2022 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW). pp 233–237

Lu Y, Zhang H, Shi L et al (2021) Expression-EEG bimodal fusion emotion recognition method based on deep learning. Comput Math Methods Med 2021:1–10

Lu Y, Zheng W-L, Li B, Lu B-L (2015) Combining eye movements and EEG to enhance emotion recognition. In: Proceedings of the 24th International conference on artificial intelligence. AAAI Press, Buenos Aires, Argentina, pp 1170–1176

Ma J, Tang H, Zheng W-L, Lu B-L (2019) Emotion recognition using multimodal residual LSTM network. In: Proceedings of the 27th ACM international conference on multimedia. Association for computing machinery, New York, pp 176–183

Masci J, Meier U, Cireşan D, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela T, Duch W, Girolami M, Kaski S (eds) Artificial neural networks and machine learning– ICANN 2011. Springer, Berlin, Heidelberg, pp 52–59

Michalowicz JV, Nichols JM, Bucholtz F (2013) Handbook of differential entropy. CRC Press

Miranda-Correa JA, Abadi MK, Sebe N, Patras I (2021) AMIGOS: a dataset for affect, personality and mood research on individuals and groups. IEEE Trans Affect Comput 12:479–493. https://doi.org/10.1109/TAFFC.2018.2884461

Moher D, Liberati A, Tetzlaff J, Altman DG (2010) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 8:336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Moin A, Aadil F, Ali Z, Kang D (2023) Emotion recognition framework using multiple modalities for an effective human–computer interaction. J Supercomput 79:9320–6349. https://doi.org/10.1007/s11227-022-05026-w

Morales S, Bowers ME (2022) Time-frequency analysis methods and their application in developmental EEG data. ScienceDirect. 54:101067

Motamedi-Fakhr S, Moshrefi-Torbati M, Hill M et al (2014) Signal processing techniques applied to human sleep EEG signals—a review. Biomed Signal Process Control 10:21–33

Muhammad F, Hussain M, Aboalsamh H (2023) A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences. Diagnostics 13:977. https://doi.org/10.3390/diagnostics13050977

Mumenthaler C, Sander D, Manstead ASR (2020) Emotion recognition in simulated social interactions. IEEE Trans Affect Comput 11:308–312. https://doi.org/10.1109/TAFFC.2018.2799593

Murugappan M, Ramachandran N, Sazali Y (2010) Classification of human emotion from EEG using discrete wavelet transform. J Biomed Sci Eng 3:390

Mutawa AM, Hassouneh A (2024) Multimodal real-time patient emotion recognition system using facial expressions and brain EEG signals based on machine learning and log-sync methods. Biomed Signal Process Control 91:105942. https://doi.org/10.1016/j.bspc.2023.105942

Nakisa B, Rastgoo MN, Rakotonirainy A et al (2020) Automatic emotion recognition using temporal multimodal deep learning. IEEE Access 8:225463–225474. https://doi.org/10.1109/ACCESS.2020.3027026

Newson JJ, Thiagarajan TC (2019) EEG frequency bands in psychiatric disorders: a review of resting state studies. Front Hum Neurosci 12:521

Noroozi F, Marjanovic M, Njegus A et al (2019) Audio-visual emotion recognition in video clips. IEEE Trans Affect Comput 10:60–75. https://doi.org/10.1109/TAFFC.2017.2713783

Ogawa S, Tank DW, Menon R et al (1992) Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc Natl Acad Sci USA 89:5951–5955. https://doi.org/10.1073/pnas.89.13.5951

Pan J, Fang W, Zhang Z et al (2023) Multimodal emotion recognition based on facial expressions, speech, and EEG. IEEE Open J Eng Med Biol. https://doi.org/10.1109/OJEMB.2023.3240280

Pan J, Yang F, Qiu L, Huang H (2022) Fusion of EEG-based activation, spatial, and connection patterns for fear emotion recognition. Comput Intell Neurosci 2022:3854513. https://doi.org/10.1155/2022/3854513

Panda D, Chakladar DD, Dasgupta T (2020) Multimodal system for emotion recognition using EEG and customer review. In: Mandal JK, Mukhopadhyay S (eds) Proceedings of the Global AI congress 2019. Springer, Singapore, pp 399–410

Pant S, Yang H-J, Lim E et al (2023) PhyMER: physiological dataset for multimodal emotion recognition with personality as a context. IEEE Access 11:107638–107656. https://doi.org/10.1109/ACCESS.2023.3320053

Park B-J, Jang E-H, Chung M-A, Kim S-H (2013) Design of prototype-based emotion recognizer using physiological signals. ETRI J 35:869–879. https://doi.org/10.4218/etrij.13.0112.0751

Park CY, Cha N, Kang S et al (2020) K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci Data 7:293. https://doi.org/10.1038/s41597-020-00630-y

Patel P, Annavarapu RN (2021) EEG-based human emotion recognition using entropy as a feature extraction measure. Brain Inf 8:20. https://doi.org/10.1186/s40708-021-00141-5

Paul A, Chakraborty A, Sadhukhan D et al (2023) A simplified PPG based approach for automated recognition of five distinct emotional states. Multimed Tools Appl 83:30697–30718. https://doi.org/10.1007/s11042-023-16744-5

Pei G, Li T (2021) A literature review of EEG-based affective computing in marketing. Front Psychol 12:602843

Peizhuang W (1983) Pattern recognition with fuzzy objective function algorithms (James C. Bezdek). SIAM Rev 25:442. https://doi.org/10.1137/1025116

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Pereira L, Brás S, Sebastião R (2022) Characterization of emotions through facial electromyogram signals. In: Pinho AJ, Georgieva P, Teixeira LF, Sánchez JA (eds) Pattern recognition and image analysis. Springer, Cham, pp 230–241

Picard RW (1997) Affective computing. MIT Press, Cambridge, Mass

Pincus SM (1991) Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 88:2297–2301. https://doi.org/10.1073/pnas.88.6.2297

Plutchik R (1982) A psychoevolutionary theory of emotions. Soc Sci Inf 21:529–553. https://doi.org/10.1177/053901882021004003

Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17:715–734

Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Inf Fusion 37:98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Qiu J-L, Liu W, Lu B-L (2018) Multi-view emotion recognition using deep canonical correlation analysis. In: Cheng L, Leung ACS, Ozawa S (eds) Neural information processing. Springer, Cham, pp 221–231

Rahate A, Walambe R, Ramanna S, Kotecha K (2022) Multimodal Co-learning: challenges, applications with datasets, recent advances and future directions. Inform Fusion 81:203–239. https://doi.org/10.1016/j.inffus.2021.12.003

Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34:96–108. https://doi.org/10.1109/MSP.2017.2738401

Reske M, Habel U, Kellermann T et al (2009) Differential brain activation during facial emotion discrimination in first-episode schizophrenia. J Psychiatr Res 43:592–599. https://doi.org/10.1016/j.jpsychires.2008.10.012

Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol-Heart Circ Physiol 278:H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.H2039

Rojas GM, Alvarez C, Montoya C, et al (2017) Multimodal study of resting-state functional connectivity networks using EEG electrodes position as seed. bioRxiv 167585

Ruiz-Padial E, Ibáñez-Molina AJ (2018) Fractal dimension of EEG signals and heart dynamics in discrete emotional states. Biol Psychol 137:42–48. https://doi.org/10.1016/j.biopsycho.2018.06.008

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Russell JA (2003) Core affect and the psychological construction of emotion. Psychol Rev 110:145–172. https://doi.org/10.1037/0033-295X.110.1.145

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Saffaryazdi N, Wasim ST, Dileep K et al (2022) Using facial micro-expressions in combination with EEG and physiological signals for emotion recognition. Front Psychol 13:864047

Saha PK, Rahman MdA, Alam MK et al (2021) Common spatial pattern in frequency domain for feature extraction and classification of multichannel EEG signals. SN Comput Sci 2:149. https://doi.org/10.1007/s42979-021-00586-9

Salankar N, Mishra P, Garg L (2021) Emotion recognition from EEG signals using empirical mode decomposition and second-order difference plot. Biomed Signal Process Control 65:102389. https://doi.org/10.1016/j.bspc.2020.102389

Samal P, Hashmi MF (2024) Role of machine learning and deep learning techniques in EEG-based BCI emotion recognition system: a review. Artif Intell Rev 57:50. https://doi.org/10.1007/s10462-023-10690-2

Sebe N, Cohen I, Huang TS (2005) Multimodal emotion recognition. Handbook of pattern recognition and computer vision. World Scientific, pp 387–409

Sepúlveda A, Castillo F, Palma C, Rodriguez-Fernandez M (2021) Emotion recognition from ECG signals using wavelet scattering and machine learning. Appl Sci 11:4945. https://doi.org/10.3390/app11114945

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma R, Pachori RB, Sircar P (2020) Automated emotion recognition based on higher order statistics and deep learning algorithm. Biomed Signal Process Control 58:101867. https://doi.org/10.1016/j.bspc.2020.101867

Siddharth PAN, Jung T-P, Sejnowski TJ (2019) A wearable multi-modal bio-sensing system towards real-world applications. IEEE Trans Biomed Eng 66:1137–1147. https://doi.org/10.1109/TBME.2018.2868759

Soleymani M, Koelstra S, Patras I, Pun T (2011) Continuous emotion detection in response to music videos. In: 2011 IEEE international conference on automatic face & gesture recognition (FG). pp 803–808

Soleymani M, Lichtenauer J, Pun T, Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3:42–55. https://doi.org/10.1109/T-AFFC.2011.25

Somol P, Novovicová J, Pudil P (2010) Efficient feature subset selection and subset size optimization. Pattern recognition recent advances. IntechOpen Rijeka, Croatia

Sriramprakash S, Prasanna VD, Murthy OVR (2017) Stress detection in working people. Procedia Comput Sci 115:359–366. https://doi.org/10.1016/j.procs.2017.09.090

Steriade M, Gloor P, Llinas RR et al (1990) Basic mechanisms of cerebral rhythmic activities. Electroencephalogr Clin Neurophysiol 76:481–508

Su Y, Li W, Bi N, Lv Z (2019) Adolescents environmental emotion perception by integrating EEG and eye movements. Front Neurorobot 13:46

Subasi A (2007) EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl 32:1084–1093. https://doi.org/10.1016/j.eswa.2006.02.005

Subramanian R, Wache J, Abadi MK et al (2018) ASCERTAIN: emotion and personality recognition using commercial sensors. IEEE Trans Affect Comput 9:147–160. https://doi.org/10.1109/TAFFC.2016.2625250

Taha B, Hwang DY, Hatzinakos D (2023) EEG emotion recognition via ensemble learning representations. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp 1–5

Thakor NV, Sherman DL (2012) EEG signal processing: theory and applications. Neural engineering. Springer, pp 259–303

Ting W, Guo-zheng Y, Bang-hua Y, Hong S (2008) EEG feature extraction based on wavelet packet decomposition for brain computer interface. Measurement 41:618–625. https://doi.org/10.1016/j.measurement.2007.07.007

Tuncer T, Dogan S, Subasi A (2021) EEG-based driving fatigue detection using multilevel feature extraction and iterative hybrid feature selection. Biomed Signal Process Control 68:102591. https://doi.org/10.1016/j.bspc.2021.102591

Übeyli ED (2010) Lyapunov exponents/probabilistic neural networks for analysis of EEG signals. Expert Syst Appl 37:985–992. https://doi.org/10.1016/j.eswa.2009.05.078

Vairamani AD (2024) Advancements in multimodal emotion recognition: integrating facial expressions and physiological signals. In: Garg M, Prasad RS (eds) Affective computing for social good: enhancing well-being, empathy, and equity. Springer, Cham, pp 217–240

Van Den Broek EL, Lisý V, Janssen JH et al (2010) Affective man-machine interface: unveiling human emotions through biosignals. In: Fred A, Filipe J, Gamboa H (eds) Biomedical engineering systems and technologies. Springer, Berlin, Heidelberg, pp 21–47

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999. https://doi.org/10.1109/72.788640

Wang M, Huang Z, Li Y et al (2021) Maximum weight multi-modal information fusion algorithm of electroencephalographs and face images for emotion recognition. Comput Electr Eng 94:107319. https://doi.org/10.1016/j.compeleceng.2021.107319

Wang Q, Wang M, Yang Y, Zhang X (2022) Multi-modal emotion recognition using EEG and speech signals. Comput Biol Med 149:105907. https://doi.org/10.1016/j.compbiomed.2022.105907

Wang S, Qu J, Zhang Y, Zhang Y (2023) Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access 11:33061–33068. https://doi.org/10.1109/ACCESS.2023.3263670

Wang S-H, Li H-T, Chang E-J, Wu A-Y (2018) Entropy-assisted emotion recognition of valence and arousal using XGboost classifier. In: Iliadis L, Maglogiannis I, Plagianakos V (eds) Artificial intelligence applications and innovations. Springer, Cham, pp 249–260

Wang X-W, Nie D, Lu B-L (2011) EEG-based emotion recognition using frequency domain features and support vector machines. In: Lu B-L, Zhang L, Kwok J (eds) Neural information processing. Springer, Berlin, Heidelberg, pp 734–743

Wen T, Zhang Z (2017) Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification. Medicine (Baltimore) 96:e6879. https://doi.org/10.1097/MD.0000000000006879

Wu D, Zhang J, Zhao Q (2020) Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. IEEE Access 8:133180–133189

Wu X, Zheng W-L, Li Z, Lu B-L (2022) Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. J Neural Eng 19:016012. https://doi.org/10.1088/1741-2552/ac49a7

Xefteris V-R, Tsanousa A, Georgakopoulou N et al (2022) Graph theoretical analysis of EEG functional connectivity patterns and fusion with physiological signals for emotion recognition. Sensors 22:8198. https://doi.org/10.3390/s22218198

Xing B, Zhang H, Zhang K et al (2019) Exploiting EEG signals and audiovisual feature fusion for video emotion recognition. IEEE Access 7:59844–59861. https://doi.org/10.1109/ACCESS.2019.2914872

Xingyuan W, Chao L, Juan M (2009) Nonlinear dynamic research on EEG signals in HAI experiment. Appl Math Comput 207:63–74. https://doi.org/10.1016/j.amc.2007.10.064

Yadav SP, Zaidi S, Mishra A, Yadav V (2022) Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Arch Comput Methods Eng 29:1753–1770. https://doi.org/10.1007/s11831-021-09647-x

Yakovyna V, Khavalko V, Sherega V, et al (2021) Biosignal and Image processing system for emotion recognition applications. In: IT&AS. pp 181–191

Yang K, Wang C, Gu Y et al (2023) Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Trans Affect Comput 14:1082–1097. https://doi.org/10.1109/TAFFC.2021.3100868

Yang Y, Gao Q, Song Y et al (2022) Investigating of deaf emotion cognition pattern By EEG and facial expression combination. IEEE J Biomed Health Inform 26:589–599. https://doi.org/10.1109/JBHI.2021.3092412

Yang Y, Wu Q, Qiu M, et al (2018) Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. In: 2018 International joint conference on neural networks (IJCNN). pp 1–7

Yong X, Menon C (2015) EEG classification of different imaginary movements within the same limb. PLoS ONE 10:e0121896

Yu C, Wang M (2022) Survey of emotion recognition methods using EEG information. Cognit Robot 2:132–146. https://doi.org/10.1016/j.cogr.2022.06.001

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp 856–863

Zeng H, Shu X, Wang Y et al (2021) EmotionCues: emotion-oriented visual summarization of classroom videos. IEEE Trans vis Comput Graph 27:3168–3181. https://doi.org/10.1109/TVCG.2019.2963659

Zhang H (2020) Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. IEEE Access 8:164130–164143. https://doi.org/10.1109/ACCESS.2020.3021994

Zhang J, Chen M, Zhao S et al (2016a) ReliefF-based EEG sensor selection methods for emotion recognition. Sensors 16:1558. https://doi.org/10.3390/s16101558

Zhang J, Yin Z, Chen P, Nichele S (2020) Emotion recognition using multi-modal data and machine learning techniques: a tutorial and review. Inform Fusion 59:103–126. https://doi.org/10.1016/j.inffus.2020.01.011

Zhang X, Liu J, Shen J et al (2021a) Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. IEEE Trans Cybernet 51:4386–4399. https://doi.org/10.1109/TCYB.2020.2987575

Zhang Y, Cheng C, Zhang Y (2021b) Multimodal emotion recognition using a hierarchical fusion convolutional neural network. IEEE Access 9:7943–7951. https://doi.org/10.1109/ACCESS.2021.3049516

Zhang Y, Ji X, Liu B et al (2017a) Combined feature extraction method for classification of EEG signals. Neural Comput Appl 28:3153–3161. https://doi.org/10.1007/s00521-016-2230-y

Zhang Y, Ji X, Zhang S (2016b) An approach to EEG-based emotion recognition using combined feature extraction method. Neurosci Lett 633:152–157

Zhang Y, Liu B, Ji X, Huang D (2017b) classification of EEG signals based on autoregressive model and wavelet packet decomposition. Neural Process Lett 45:365–378. https://doi.org/10.1007/s11063-016-9530-1

Zhang Z (2019) Spectral and time-frequency analysis. EEG Signal Processing and feature extraction 89–116

Zhao L-M, Li R, Zheng W-L, Lu B-L (2019) Classification of five emotions from EEG and eye movement signals: complementary representation properties. In: 2019 9th International IEEE/EMBS conference on neural engineering (NER). pp 611–614

Zhao Z-W, Liu W, Lu B-L (2021) Multimodal emotion recognition using a modified dense co-attention symmetric network. In: 2021 10th International IEEE/EMBS conference on neural engineering (NER). pp 73–76

Zheng W-L, Liu W, Lu Y et al (2019) EmotionMeter: a multimodal framework for recognizing human emotions. IEEE Trans Cybernet 49:1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

Zheng X, Yu X, Yin Y et al (2021) Three-dimensional feature maps and convolutional neural network-based emotion recognition. Int J Intell Syst 36:6312–6336. https://doi.org/10.1002/int.22551

Zhou J, Wei X, Cheng C, et al (2018)Multimodal Emotion Recognition Method Based on Convolutional Auto-Encoder. Int J Comput Intell Syst 12:351–358. https://doi.org/10.2991/ijcis.2019.125905651

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Rajasekhar Pillalamarri[1] · **Udhayakumar Shanmugam**[1]

✉ Udhayakumar Shanmugam
s_udhayakumar@ch.amrita.edu

Rajasekhar Pillalamarri
p_rajasekhar@ch.students.amrita.edu

[1] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India