Exploration of Medicine

Open Exploration

Open Access Review

Artificial intelligence strategies for emotion recognition in cancer pain research

Marco Cascella[1*], Adele Zarrella[1], Valeria Conti[1], Amelia Filippelli[1], Maria Pia Bruno[1], Dalila Esposito[1], Rosario De Feo[1], Valentina Cerrone[1], Cosimo Guerra[1], Flavio Di Lisio[2], Francesco Sabbatino[1], Ornella Piazza[1], Stefano Cirillo[3], Giuseppe Polese[3]

[1]Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, Baronissi, 84081 Salerno, Italy

[2]Interdisciplinary Center for Health Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy

[3]Department of Computer Science, University of Salerno, Fisciano, 84084 Salerno, Italy

***Correspondence:** Marco Cascella, Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, Via Salvatore Allende, Baronissi, 84081 Salerno, Italy. mcascella@unisa.it

Academic Editor: Hua Su, University of California, USA

Received: February 28, 2026 **Accepted:** April 20, 2026 **Published:** May 19, 2026

Cite this article: Cascella M, Zarrella A, Conti V, Filippelli A, Bruno MP, Esposito D, et al. Artificial intelligence strategies for emotion recognition in cancer pain research. Explor Med. 2026;7:1001404. https://doi.org/10.37349/emed.2026.1001404

Abstract

Although emotions play a fundamental role in modulating pain perception, their objective assessment in clinical contexts remains challenging. Recent advances in artificial intelligence (AI) have opened new opportunities to measure emotional states through facial expression analysis, physiological signal modeling, natural language processing (NLP), and multimodal data integration. In affective computing, the field that focuses on technologies designed to recognize, interpret, process, and simulate human emotions, facial expression-based emotion recognition has progressed from traditional machine learning methods to advanced deep learning approaches, including convolutional neural networks (CNNs), attention-based hybrid models, and transformer architectures. Similarly, recurrent neural networks and self-supervised learning methods have been implemented for developing models from physiological signals such as electrocardiography, photoplethysmography, galvanic skin response, and related biosignals. Additionally, NLP systems can extract affective information from naturalistic text, using both lexicon-based and transformer-based models. Finally, multimodal fusion and alignment techniques allow the integration of heterogeneous data streams, providing richer and more ecologically valid emotion representations. Collectively, these strategies offer powerful tools for advancing automatic pain assessment (APA) in cancer care, with the potential to support personalized, emotion-aware therapeutic approaches. However, from an AI perspective, several open challenges remain, including multimodal representation learning under weak supervision, robustness to missing or degraded modalities, limited explainability of affective inference models, lack of standardized benchmarking protocols, and the presence of bias and domain shift in emotion datasets. Given the inherently subjective, context-dependent, and culturally mediated features of the emotional experience, further research is needed to address these technical limitations, integrating technological advances with the intrinsic complexity of emotion interpretation.

Keywords

artificial intelligence (AI), emotional computing, automatic pain assessment, machine learning, computer vision, natural language processing, multimodal fusion, explainable AI

Introduction

Emotion is a brief, integrated psychophysiological state that combines subjective feeling, evaluative appraisal, autonomic bodily arousal, expressive behaviors, and readiness to act, triggered by events deemed relevant to the person's goals [1]. This multifaceted set of states helps individuals cope with stressful and non-stressful events. Their adaptive value became scientifically relevant in the second half of the 19th century, when Charles Darwin demonstrated that facial and bodily expressions serve communicative functions common to all species [2]. A century later, affective science converged on two complementary perspectives, including theories of discrete basic emotions and continuous dimensional models such as Russell's circumplex, which maps feelings along the axes of valence and arousal [3].

Similarly, pain is more than just a sensory signal, and the International Association for the Study of Pain (IASP) defines it as "an unpleasant sensory and emotional experience" [4]. Therefore, emotions modulate pain in a bidirectional manner; specifically, anxiety and sadness can amplify nociceptive processing, while positive perceptions of safety can dampen it through descending inhibitory pathways [5]. Furthermore, in chronic conditions, maladaptive emotions can worsen catastrophizing and increase disability. These key aspects underscore the clinical need to assess affective and nociceptive function. Nevertheless, this assessment is far from straightforward. Briefly, affective states can rapidly fluctuate, are filtered by cultural rules of expression, and can be blunted or masked by analgesics or sedatives [1–3, 6]. Moreover, different populations, such as infants, individuals with cognitive deficits, or those on ventilatory support, are unable to communicate their feelings and pain experiences reliably [7].

Moving beyond subjective assessments, efforts to establish objective and quantifiable indicators of affectivity have led researchers toward algorithmic measurement of emotions. Rosalind Picard's [8] manifesto Affective Computing reframed these theoretical insights as computational challenges, proposing that machines capable of recognizing and responding to emotion would transform human-computer interaction [9, 10]. Importantly, artificial intelligence (AI) strategies such as deep learning (DL) and machine learning (ML) models can integrate vision, speech, and verbal features, as well as physiological signals, to detect intonation shifts, but also the subtle, context-related signatures that elude human coders. This research area is commonly referred to as emotion AI or affective computing [11].

Traditionally, cancer pain assessment relies on patient-reported outcome (PRO) measures such as the Visual Analog Scale (VAS) and Numeric Rating Scale (NRS), as well as observational tools used when self-report is not feasible. While these approaches are widely adopted in clinical practice, they are inherently subjective, intermittent, and may be influenced by cognitive, emotional, and contextual factors, particularly in complex oncology settings. These limitations highlight the need for more objective and continuous assessment strategies, motivating the development of AI-based automatic pain assessment (APA) systems. This emerging field infers a patient's pain intensity from facial, vocal, and different physiological cues (i.e., biosignals) to precisely and objectively assess pain [12]. In a clinical landscape that requires multidisciplinary collaboration between clinicians, AI researchers mostly in the fields of computer vision (CV) and natural language processing (NLP), psychologists, linguists, and researchers in pain medicine, it is also necessary to identify the most suitable methods for APA [13].

From an AI perspective, APA can be formalized as a multimodal inference problem under uncertainty [14]. Therefore, emotional and nociceptive states are latent variables that must be inferred from heterogeneous data streams, including facial expressions, physiological signals, speech, and text. Nevertheless, several core AI challenges emerge. They include representation learning across modalities, temporal dependency modeling, domain adaptation from controlled datasets to real-world environments, and robustness to missing or degraded modalities, as well as multimodal alignment.

Since integrating emotional recognition methods into APA systems may help overcome current limitations of subjective pain ratings and improve the personalization of cancer pain management, we aim to provide a structured overview of AI strategies for emotion recognition in cancer pain research, organizing existing approaches into methodological categories and highlighting current limitations and future research directions. On the other hand, emotion recognition remains inherently complex due to its interdisciplinary nature, involving psychological processes, physiological responses, and computational analysis, which must be jointly considered to achieve reliable automated assessment. In this context, a clear distinction between different types of cancer-related pain is essential, particularly when considering acute and chronic manifestations. These pain types differ in temporal dynamics, behavioral expression, and physiological correlates, with important implications for the development and validation of APA models [12–14]. This review primarily focuses on chronic cancer-related pain, which represents the most prevalent and clinically complex condition in oncology. However, acute and procedural pain, such as postoperative pain or pain related to diagnostic and therapeutic interventions, are also considered where relevant, particularly in relation to the applicability of AI-based emotion recognition systems.

Although this work is not intended as a systematic review, a structured literature selection process was adopted to ensure transparency and methodological rigor. A targeted search was conducted across major scientific databases, including PubMed/MEDLINE, Scopus, Web of Science, and IEEE Xplore, using combinations of keywords related to APA, emotion recognition, and pain. To better focus on advances and perspectives in the field of APA, particular attention was given to studies addressing one or more of the core domains explored in this review, namely, facial expression analysis, physiological signal modeling, NLP, multimodal fusion, and foundation models.

Facial-expression-based emotion recognition

In different settings, facial expression analysis represents a valuable approach for pain assessment, particularly in patients who are unable to provide reliable self-reports due to advanced disease, cognitive impairment, or treatment-related factors [15]. In such contexts, automatic analysis of facial cues may offer an objective and continuous alternative to traditional assessment methods [12]. Consequently, it represents a promising non-invasive modality for APA, enabling continuous monitoring of pain-related affective responses even in patients with communication barriers. Notably, facial expression recognition (FER) is fundamental to understanding human emotions. Moreover, it is also a key part of non-verbal communication. On the other hand, the inherent complexity of emotions, together with individual variations in how they are perceived and processed, still represents a significant obstacle for automatic recognition systems [16]. Facial muscle activation patterns can be used from a theoretical perspective to analyze facial expressions, as formally stated in Facial Action Coding System (FACS), which encodes facial movements into action units (AUs), thus facilitating orderly and repeatable inference of underlying emotional states [17]. Building on the relevance of AUs for objective facial coding, several APA approaches have been proposed [12, 13]. For example, researchers developed a binary classifier grounded in extended FACS features, using an artificial neural architecture to discriminate between pain and no-pain states in a cohort of oncology patients undergoing video recordings during clinical procedures. The model, trained on facial AUs extracted using OpenFace from datasets annotated according to the FACS, achieved a validation accuracy of 94.48%, with a precision of 0.95, a recall (sensitivity) of 0.97, and an AUROC of approximately 0.98 [18].

Driven primarily by the introduction of DL methods, automatic FER has undergone radical changes over the past decade. Research has shown that some emotions, such as fear, are processed differently than others, producing distinct patterns in their classification and recognition [19]. In this context, DL, particularly convolutional neural networks (CNNs), has proven useful for continuous prediction tasks and emotion classification [20]. CNNs, commonly used in image recognition applications, are powerful neural networks. They operate on a series of convolutional and pooling layers that gradually extract meaningful features from the input images. One or more fully connected layers then process these features to produce the final prediction. To perform image recognition, a CNN must be trained using a large, annotated dataset

containing examples of the desired elements. The network learns the connections between the input features and their appropriate labels by adjusting its parameters using backpropagation and optimization techniques during the training phase. Once trained, CNNs can infer labels for previously unseen images [21].

From a methodological perspective, CNNs are particularly effective in this domain because they can capture spatial hierarchies of facial features, enabling the detection of subtle muscle activations associated with pain expressions. These models use reference data leveraging psychological principles, exploiting emotional recognition based on classification and regression processes. Specifically, classification assigns emotional data (e.g., facial expressions, physiological signals) to discrete categories such as "happy", "angry", or "in pain", while regression estimates the intensity or continuous level of an emotional response, allowing for more nuanced emotional profiling [17, 20]. Recent advances have led to interesting discoveries. One of the most notable examples is the development of techniques for estimating emotional content through spatial analysis of facial expressions. Computerized systems for advanced cognitive perception, particularly those based on neural networks such as DL models, rely on detailed, static representations of facial features. These systems use geometric and spatial analysis to improve the accuracy of facial recognition [22]. Beyond clinical and affective computing contexts, visual and behavioral analysis techniques have also been applied in intelligent video surveillance. In this domain, the goal is not limited to low-level feature extraction, but extends to the semantic interpretation of human activities and behaviors in order to detect relevant or abnormal events within complex scenes [23]. Knowledge representation frameworks have been proposed to model contextual elements, actions, and their temporal compositions, enabling higher-level reasoning over video data. Such approaches integrate visual analysis with structured representations of context and events, supporting both automatic event recognition and the summarization of relevant video segments for human monitoring. Although originally developed for intelligent video surveillance, knowledge representation frameworks for modeling contextual and temporal patterns can be adapted to APA systems to support the structured interpretation of pain-related behaviors in clinical settings.

Interestingly, CNN-based facial analysis can be embedded into APA pipelines to provide objective emotional markers that correlate with pain intensity. For example, in a recent feasibility study, the authors employed the YOLOv8 real-time object detection CNN to identify facial regions and extract pain-related expressions from live video streams of oncological patients. The system achieved an overall detection accuracy of 91.7%, with a mean inference time of 18.2 ms per frame, thus allowing real-time monitoring even in non-controlled clinical environments. Notably, the model demonstrated an F1-score of 0.90 for pain detection versus baseline states, and thus demonstrates the robustness of CNN-based facial recognition for dynamic bedside assessment [24].

Following this line, more complex architectures have been developed and fine-tuned. Wang and Jia [25] fine-tuned an extended neural network consisting of a hybrid dual-branch structure with an attention mechanism focused on the global and local facial features. The parallel configuration of the network helps to distinguish between similar facial expressions, contributing to the improved model's accuracy, even in the presence of noise or some change in the environment. The results obtained on popular datasets, such as RAF-DB and FER-Plus, heavily outperformed the traditional methods, which communicated the value of multimodal fusion and selective attention in the domain of automatic emotion recognition. These advanced architectures, such as hybrid attention and transformer models, could improve the robustness and accuracy of APA systems in real-world clinical environments.

In addition to hybrid models and attention-based architectures, further research is needed in the realm of automatic FER. For example, generalizability remains one of the primary challenges. In this context, while most systems are trained and validated using controlled and standardized datasets, they fail to encompass the diversity and complexity of spontaneous facial expressions [26]. This discrepancy results in a decline in performance when models are applied to data collected in natural conditions, which are characterized by numerous confounding factors. These include variations in head positioning, partial occlusions caused by items such as glasses, beards, or masks, uneven lighting conditions, and demographic

imbalances in the training data, which can negatively affect the system's ability to accurately recognize emotions in diverse populations [27, 28].

To overcome these challenges, several recent studies have developed advanced models designed to maintain high performance despite obstacles on the face and variations in position. For example, Gao and Zhao [28] developed a transformer-based model called transformer facial encoders (TFEs), which dynamically focuses attention on visible regions of the face while simultaneously reconstructing hidden parts. Validated on the RAF-DB and AffectNet [29] datasets, this approach showed better performance than traditional CNN methods, especially in the presence of partial facial coverings. Similarly, Li et al. [30] proposed a multi-angle feature extraction (MAFE) framework that leverages a hybrid backbone consisting of CNN and Swin Transformer. This model has been specifically optimized to preserve fine facial details, thus improving the robustness and accuracy of recognition in unfavorable environmental conditions. Another promising way to improve the performance of FER systems provides the ability to generalize to unseen domains. Zhang et al. [31] proposed an innovative approach that combines features derived from CLIP with sigmoid masks to isolate relevant expressive signals, enabling zero-shot expression recognition, without the need for retraining on the target domain. Their framework demonstrated superior performance compared to traditional methods on multiple benchmarks, offering greater robustness in heterogeneous and realistic contexts.

Recent trends in FER research emphasize the importance of modelling temporal dynamics and incorporating contextual information to more accurately capture the transient nature of emotional expressions. Methods leveraging static images often fail to detect microexpressions or subtle emotional transitions, leading to increasing interest in approaches leveraging video sequences and temporal modeling. In particular, architectures integrating convolutional backbones with temporal modules, such as Temporal Convolutional Networks (TCNs), have demonstrated strong performance in capturing dynamic facial patterns across benchmark datasets [32], highlighting their potential relevance for APA. Moreover, architectures such as three-dimensional CNNs (3D CNNs) and recurrent models, including long short-term memory (LSTM) networks, have demonstrated superior capabilities in capturing the temporal evolution of facial features, thus enabling more accurate emotion classification in dynamic environments [33, 34]. Temporal modeling of microexpressions and dynamic facial patterns is particularly valuable for APA, as it can facilitate the detection of subtle pain-related changes over time (Table 1).

Table 1. Deep learning applications for facial-expression-based emotion recognition.

Approach/Model[†]	Description	Strengths (relevance to APA)	Main limitations	Ref.
CNN	Extraction of spatial features from static facial images for emotion classification or regression	Effective detection of facial muscle activation patterns (AUs); suitable for baseline pain/no-pain discrimination	Limited ability to capture temporal dynamics; reduced performance in real-world conditions (occlusions, variability)	[24]
Hybrid CNN	A combination of convolutional feature extraction with attention mechanisms focusing on salient facial regions	Improved discrimination of subtle expressions; better robustness to noise and inter-individual variability	Increased architectural complexity; requires large annotated datasets	[25]
Transformer-based models (TFE, Swin)	Attention-based architectures model global dependencies and dynamically focus on informative regions	Strong robustness to occlusions and pose variations; improved generalization across datasets	High computational cost; data-intensive training	[28–30]
CNN + temporal models (TCN, LSTM, 3D CNN)	Integration of spatial feature extraction with temporal modeling of video sequences and facial dynamics	Capture of microexpressions and temporal evolution of pain-related facial patterns; critical for continuous and real-time APA	Requires temporally annotated datasets; higher computational burden	[32–34]

[†] These approaches differ in their ability to capture static versus dynamic emotional features, with temporal models being particularly relevant for continuous pain assessment in clinical settings. CNN: convolutional neural network; LSTM: long short-term memory; TFE: transformer facial encoder; 3D CNN: three-dimensional CNN; TCN: Temporal Convolutional Network; AUs: action units; APA: automatic pain assessment.

As emotion recognition systems strive for greater realism and accuracy, the integration of multimodal data, such as vocal cues, physiological measurements, and contextual information, has emerged as a critical direction for research. This multimodal integration seeks to improve the interpretability of emotional states by addressing the limitations of relying solely on facial signals, which can often be ambiguous or insufficient when considered in isolation. Although these systems inevitably involve increased computational requirements, they provide a richer and more ecologically valid representation of emotional behavior. Mirroring the multisensory integration that characterises human emotional perception, multimodal approaches can substantially improve the accuracy, generalisability, and contextual sensitivity of emotion recognition models [35].

A growing number of studies are devoted to the issue of fairness and bias in FER recognition systems. Empirical evidence has shown that these models often exhibit inconsistent performance across different demographic groups, such as gender, age, and ethnicity, mainly due to unbalanced representation within widely used FER datasets. To overcome these limitations, current research is increasingly exploring approaches such as data balancing, domain adaptation, and the integration of fairness-aware learning frameworks. These strategies are becoming fundamental to the development of more equitable and generalisable FER systems [36].

Physiological and wearable signal modelling

The modelling of physiological signals obtained through wearable technologies has attracted considerable interest in both academic and applied research, particularly for its potential in real-time, continuous, and non-invasive monitoring of human physical conditions. Wearable devices can capture a wide range of biosignals, including electrocardiography (ECG), photoplethysmography (PPG), galvanic skin response (GSR), skin temperature, respiratory rate, accelerometric data, and other complementary physiological modalities. These signals are key indicators of autonomic nervous system activity and provide insights into various psychophysiological processes, such as stress, fatigue, emotional arousal, and cognitive load [37, 38].

Physiological signals offer an objective window into affective and nociceptive states, making them essential components of multimodal APA frameworks [39–41]. In AI terms, physiological signal modeling constitutes a non-stationary multivariate time-series learning problem, characterized by high inter-subject variability and context-dependent dynamics. Recurrent and convolutional architectures implicitly encode inductive biases related to temporal continuity and local dependencies, while self-supervised objectives aim to learn invariant representations across subjects and recording conditions.

However, modelling biosignals presents multiple challenges, including high susceptibility to motion artefacts, inter-individual variability, and the non-linear and non-stationary nature of physiological data. Recent work has applied DL models, particularly recurrent architectures such as LSTM networks, to address these issues, demonstrating greater robustness to noise and motion when using multisensory fusion [42]. LSTM networks are an advanced type of recurrent neural network designed to capture long-term dependencies in sequential data. They use a gated memory mechanism, consisting of input, forget, and output gates, to regulate the flow of information and preserve relevant temporal features over time. This makes them particularly well-suited to the analysis of physiological signal time series, where past states contain information essential for accurate modelling [43]. Moreover, by learning temporal autonomic patterns, LSTM-based models can support APA in continuously detecting pain responses without patient self-report.

Transfer learning techniques have also been employed to improve model generalisability across heterogeneous sensing modalities by leveraging pretrained representations and cross-modal adaptation strategies, particularly when transferring knowledge between physiologically related biosignals or integrating multimodal affective cues in data-limited settings [44, 45]. Hybrid architectures integrating CNNs, bidirectional LSTMs, wavelet-based feature extraction, and attention mechanisms have shown promising performance in reconstructing ECG signals from PPG inputs, improving signal fidelity while enhancing model interpretability [46].

Recent innovations in physiological signal modelling have focused on improving signal quality and learning generalisable representations from wearable device data [47]. One promising direction involves the use of advanced signal processing and representation learning techniques to improve the quality and robustness of PPG signals in real-world conditions [48]. These models aim to remove motion artefacts and noise by learning a compressed representation of the clean signal, exploiting the sparsity hypothesis of physiological signals. This is particularly useful for real-world scenarios, where PPG recordings from wrist-worn wearables often suffer from low signal-to-noise ratios due to motion or ambient light interference. At the same time, the development of base models, i.e., large-scale models pre-trained on diverse, multimodal physiological datasets, represents a step towards more flexible and reusable architectures. Recent advances have also explored multimodal foundation models for physiological signals, aiming to learn transferable representations across tasks such as emotion recognition, stress detection, and sleep analysis. For example, NormWear has been trained on signals such as ECG, PPG, GSR, and electroencephalography (EEG) across various populations and contexts [49]. These models can generate robust representations that transfer across tasks and domains, enabling zero-shot or few-shot learning for applications such as emotion recognition, stress detection, sleep stage classification, and even automotive risk estimation.

Furthermore, self-supervised learning (SSL) approaches, particularly those based on reconstruction objectives, have emerged as a key technique for exploiting large amounts of unlabelled physiological data, especially in wearable biosignal analysis contexts [50]. Unlike supervised approaches, SSL learns meaningful representations by solving pretext tasks, such as signal reconstruction, temporal order prediction, or contrastive learning, without relying on manual annotations. This is particularly advantageous in biomedical contexts, where labelled data is often scarce or costly to obtain. The resulting models can then be optimised for downstream tasks with minimal supervision, improving scalability and applicability in health monitoring via wearable devices [50]. In this context, transfer and SSL can further enhance APA scalability by exploiting large amounts of wearable data without the need for extensive manual labeling (Table 2).

Table 2. Physiological and wearable signal modeling for emotion recognition.

Approach/Model	Input signals	Main function	Typical applications	Key challenges	Ref.
Recurrent and transfer learning models (LSTM, CNN-BiLSTM, transfer learning)	ECG, PPG, GSR, temperature, respiration, accelerometry	Temporal modeling and cross-modal generalization of physiological time series	Emotion recognition, stress detection, cognitive load monitoring, wearable health inference	Motion artifacts, inter-individual variability, non-stationarity, domain shift	[42–47]
Signal processing and feature extraction approaches	Primarily PPG and related wearable biosignals	Signal denoising, compression, and improvement of signal quality	Preprocessing for wearable-based monitoring and downstream classification tasks	Sensitivity to real-world noise, reduced robustness under motion, and low signal-to-noise conditions	[48]
Multimodal foundation models (e.g., NormWear)	Multivariate physiological signals (e.g., ECG, PPG, GSR, EEG)	Learning transferable representations across tasks and populations	Emotion recognition, stress detection, sleep analysis, zero-shot/few-shot wearable sensing	High training cost, data heterogeneity, potential bias, and limited clinical validation	[49]
Self-supervised learning (SSL)	Unlabeled physiological and wearable biosignals, especially PPG	Learning latent representations through reconstruction-based or related pretext tasks	Scalable wearable biosignal modeling with limited annotation requirements	Defining suitable pretext objectives, interpretability, and downstream transferability	[50]

LSTM: long short-term memory; CNN: convolutional neural network; BiLSTM: bidirectional LSTM; ECG: electrocardiography; PPG: photoplethysmography; GSR: galvanic skin response; EEG: electroencephalography.

Natural-language approaches to affective text

In the context of cancer pain, NLP offers a valuable opportunity to extract affective and pain-related information from textual data sources routinely generated in clinical practice, including PROs, clinician notes, and semi-structured interviews. Importantly, recent evidence suggests that patients' verbal expressions encode not only pain intensity but also emotional, cognitive, and pragmatic dimensions of the

pain experience. For instance, linguistic analyses of oncological patients' utterances have shown that pain-related discourse frequently includes expressive speech acts, metaphorical descriptions, and narrative structures reflecting psychological states such as distress, adaptation, and coping mechanisms [51]. At the same time, accurately identifying emotional states from everyday language remains a fundamental challenge in psychology and computational linguistics [1, 3, 11]. Additionally, understanding linguistic variability is crucial for APA, as language often encodes subtle emotional and pain-related cues [11, 24, 51]. From a computational perspective, affective text analysis extends beyond traditional sentiment classification and can be framed as an inference task over latent emotional states expressed through language [52]. While transformer-based models learn contextual semantic representations, they remain sensitive to individual linguistic habits, pragmatic context, and domain-specific language use, posing crucial challenges for generalization and interpretability, particularly in clinical narratives, where emotional expressions may be implicit, metaphorical, or shaped by coping strategies.

Recent developments in NLP have improved the detection of emotional content in text by leveraging datasets collected in natural contexts and by using models that account for both linguistic context and individual differences in emotional expression. Recent research conducted by Fisher et al. [53] sought to use pioneering tools in negative emotion monitoring therapy by employing NLP to determine its utility in adolescents. The authors used transcripts from Ecological Momentary Assessment (EMA), a real-time emotional capturing tool, with a dataset that comprised 7,680 open-ended texts obtained from 97 subjects. The primary aim of the research was to determine the possibility of negative emotion classifiers identifying affective models drawn from language used in daily self-reporting annotations [53]. This work is notable for the direct contrast made between nomothetic and idiographic modelling approaches. Nomothetic models use aggregate data and ascribe the same predictive elements to every individual. Idiographic models are constructed for everyone, personally mapping language use and emotional states. This distinction in methodology rests on the assumption that emotional expression is deeply personal. The same linguistic or syntactic cues can mean vastly different emotional things depending on the individual, i.e., the speaker. Some expressions that, in the case of an adolescent, suggest some sort of distress, are neutral or even positive for another.

This level of variability makes it difficult to apply generalizable models to clinical psychology and youth mental health, where emotional signals are often misinterpreted, potentially placing vulnerable individuals at risk and substantially affecting clinical outcomes [53]. Findings in this regard are pivotal to effective text analysis. They show how ineffective blanket approaches are in any emotionally diverse population. In recent years, NLP has increasingly focused on models that move beyond basic sentiment or valence detection toward richer representations of affective and contextual meaning. Specifically, researchers are attempting to understand the subtleties, variability, and context-dependence in naturalistic language. For instance, transformer-based models, such as Robustly Optimized BERT Pretraining Approach (RoBERTa) [54], provide powerful contextual embeddings that could be fine-tuned on domain-specific datasets, including those targeting affective states, to improve the detection of subtle emotional patterns. This is particularly relevant in light of recent large-scale studies leveraging social media corpora, where emotional meaning is inferred from millions of real-world text instances and enriched through human-annotated lexicons, enabling a more nuanced and multidimensional representation of affective content [55, 56].

Moreover, as affective computing applications move toward personalized mental health monitoring, the ability to model individual baselines and detect deviations over time becomes critical. This adaptability is particularly relevant in digital mental health tools where passive, language-based monitoring can offer early indicators of emotional distress without requiring active clinical input [57]. However, these advancements also raise significant methodological and ethical questions, such as the reliability of emotion inference over time, the interpretability of DL models in sensitive settings, and the need to safeguard user privacy when analyzing emotionally laden text data [58].

Continued progress in this field will likely depend on the development of multimodal, ethically responsible systems that combine linguistic, behavioral, and contextual signals. Such systems should not only predict affective states with precision but also contribute to actionable outcomes, such as timely

intervention or supportive feedback. As NLP tools are increasingly integrated into clinical and educational technologies, the challenge will be to ensure that affective text analysis remains both scientifically rigorous and aligned with human-centered values [59]. This aspect is of pivotal importance for APA research as it can provide a reliable extraction of affective information from patient reports, clinical notes, or digital health records (Table 3).

Table 3. NLP approaches for affective text analysis.

Approach/Model	Description	Strengths (relevance to APA)	Main limitations	Ref.
Transformer-based models (BERT, RoBERTa)	Pre-trained language models fine-tuned on emotion-specific datasets to capture contextual semantic representations	High sensitivity to subtle affective cues; effective modeling of context-dependent emotional language	Limited interpretability; sensitive to domain shift and individual linguistic variability	[52, 54]
Lexicon-based and hybrid approaches (e.g., NRC, emoji lexicons)	Combination of emotion lexicons and data-driven representations to infer affective content from text	Interpretable and robust across domains; useful for capturing multidimensional emotional signals	Limited ability to model complex, implicit, or metaphorical language	[55, 56]
Idiographic versus nomothetic modeling approaches	Comparison between population-level models and personalized language-emotion mappings	Enables modeling of individual variability in emotional expression; critical for personalized APA	Requires longitudinal and subject-specific data; limited generalizability	[53]
Clinical NLP and longitudinal monitoring approaches	Analysis of patient-generated text (e.g., PROs, clinical notes, EMA) to track emotional states over time	Supports early detection of emotional distress; enables continuous and real-world monitoring	Ethical concerns (privacy); variability in data quality; challenges in temporal consistency	[53, 58, 59]

APA: automatic pain assessment; NLP: natural language processing; BERT: Bidirectional Encoder Representations from Transformers; RoBERTa: Robustly Optimized BERT Pretraining Approach; NRC: National Research Council (Emotion Lexicon); EMA: Ecological Momentary Assessment; PROs: patient-reported outcomes.

Multimodal fusion and alignment

In the context of cancer pain assessment, multimodal fusion is particularly relevant because pain is inherently multidimensional, involving behavioral, physiological, and cognitive components [4]. Integrating heterogeneous signals such as facial expressions, speech, and physiological responses allows a more comprehensive representation of the patient's affective and nociceptive state, reducing the uncertainty associated with single-modality approaches. This distinction is particularly relevant when integrating multimodal data, as the relative contribution of facial, physiological, and behavioral signals may differ between acute and chronic pain conditions [12, 13].

Multimodal fusion strategies in AI can be broadly categorized into early fusion, late fusion, and hybrid approaches. Early fusion integrates raw or low-level features from different modalities into a shared representation, while late fusion combines modality-specific predictions at the decision level. Hybrid and representation-level fusion methods aim to learn joint latent spaces through cross-modal attention or shared embeddings [60].

Notably, multimodal fusion can provide the foundation for next-generation APA systems by combining complementary emotional and physiological signals, consistent with prior multimodal pain assessment frameworks integrating facial and physiological signals [61]. Furthermore, multimodal AI systems can integrate speech analysis and FER, supporting continuous and real-time pain assessment. In cancer pain, researchers implemented an automatic emotion recognition system trained on the EMOVO dataset, a validated Italian corpus designed to simulate the six prototypical emotions ("Big Six": happiness, anger, fear, sadness, surprise, disgust) [62]. A Multi-Layer Perceptron Neural Network, trained on 181 prosodic features (including pitch, intensity, formants, jitter, shimmer, and speech rate), achieved an overall classification accuracy of 84%, with balanced precision, recall, and F1-scores across emotion classes. This emotional speech model was then applied to real-world audio recordings obtained from clinical interviews with oncology patients, allowing for the continuous annotation of emotional states during naturally occurring communicative contexts. In parallel, facial expressions were analyzed through an AUs-based classifier, and using continuous video streams, facial expressions were processed in real-time and

categorized into binary pain/no-pain states. The FER model alone yielded an accuracy of 82.4%. Nevertheless, the integration of SER features and facial emotion markers through feature-level fusion significantly enhanced the system's discriminative power. The combined multimodal model achieved an overall accuracy of 89.3% and an AUC of 0.91, substantially outperforming the unimodal models. Furthermore, facial emotional markers exhibited a statistically significant positive correlation with patients' self-reported pain intensity scores ($r = 0.62$, $p < 0.01$). Finally, speech annotation and facial expression analysis were synchronized using the Eudico Linguistic Annotator (ELAN) v6.7, enabling time-aligned multimodal labeling and exploration of emotional trajectories during the interviews. This methodological framework allowed the identification of predominant emotional states over time and their dynamic relationship with pain expressions [63]. Importantly, similar AI-based approaches for pain assessment have been explored by multiple independent research groups, particularly in multimodal and physiological signal-based frameworks, supporting the generalizability of these methodologies [64, 65].

However, in clinical APA systems, multimodal integration requires not only combining data streams but also ensuring their meaningful alignment. For instance, facial expressions, physiological responses, and speech signals must be temporally synchronized to reflect the same underlying pain episode. Misalignment between modalities may lead to inconsistent or misleading interpretations of the patient's state. Therefore, multimodal alignment plays a critical role in ensuring that heterogeneous signals are coherently integrated in real-world clinical scenarios. This process requires not only the fusion of data, but also sensible structuring and valid correlation between information and modalities (alignment) [66, 67]. The integration of heterogeneous modalities such as visual, physiological, and audio signals poses significant challenges in clinical environments, where data may be incomplete, noisy, or acquired at different sampling rates.

The joint processing of heterogeneous information sources is a central challenge in contemporary AI, particularly in the field of multimodal learning [60]. In application contexts where data no longer appear as homogeneous entities but as parallel streams of different types (such as images, audio signals, text, three-dimensional data, sensor measurements, and time recordings), it becomes essential not only to aggregate these sources but also to structure them according to a logic that preserves semantic consistency and informational relevance [68–70]. Simply superimposing disparate signals is not enough to generate usable knowledge; to prevent the operation from degenerating into a disorganised aggregation of data, it is necessary to implement two complementary and interdependent processes: multimodal fusion and multimodal alignment [71].

Fusion can be defined as the process through which representations from different domains are combined into a unified form, allowing for the synergistic exploitation of the entire wealth of information. Essentially, this involves moving beyond unimodal analysis to obtain joint representations capable of capturing intermodal relationships, thereby improving the system's ability to perform complex inferences [72]. However, this fusion operation is only effective if preceded or accompanied by a precise alignment phase, which aims to establish relevant correspondences—temporal, spatial, or conceptual—between entities represented in different modes but referring to the same event, object, or concept [73]. In clinical APA applications, alignment is primarily temporal (e.g., synchronizing audio tracks with lip movements in videos) [74] and spatial (e.g., associating RGB data with depth information in images) [73]. Conceptual alignment is also relevant, as it links observed signals to higher-level representations or target variables within the model [75].

The complexity of these tasks is compounded by several challenges: different sampling frequencies, temporal discrepancies, incomplete data, modality-specific sounds, and asymmetries in semantic abstraction levels [76]. Furthermore, modalities vary significantly in their intrinsic structure: text is discrete and symbolic, images are continuous and spatially distributed, and audio is sequential and context dependent [77]. This heterogeneity requires advanced representation techniques such as shared latent spaces, multimodal DL architectures based on cross-modal attention mechanisms, and contrastive learning approaches that facilitate associations between modalities without imposing a reductive common format [78].

To address these issues, recent research has focused heavily on models that can not only integrate different modalities but also do so contextually, considering the dynamic relationships and semantic context in which these modalities manifest themselves [79]. State-of-the-art multimodal architectures, including multimodal transformers and graph-based neural networks, do not merely map heterogeneous inputs into a common space but rather aim to model the complex interactions between elements of different modalities [80]. As a result, fusion becomes an adaptive and context-driven process, while alignment takes on an epistemological role by anchoring representations to shared and interpretable semantic structures [81].

Although significant progress has been made, several theoretical and practical challenges remain. In particular, handling cases where one or more modalities are missing or degraded (mode dropout) [82], reducing dependence on large manually annotated datasets through SSL methods [83], and ensuring the interpretability of multimodal decision-making processes [84] remain active areas of research. Furthermore, the issue of scalability poses significant obstacles, as the computational complexity associated with simultaneously handling multiple modalities can quickly compromise system efficiency and reliability, especially in real-time applications such as autonomous driving or surgical robotics [85]. These challenges are particularly critical in oncology, where real-time and reliable pain monitoring is required to support clinical decision-making.

Foundation models and self-supervised emotion understanding

Foundation models have recently emerged as a transformative paradigm in AI, enabling the learning of rich and transferable representations through large-scale self-supervised training on heterogeneous and predominantly unlabeled data [86]. This paradigm is particularly relevant for emotion understanding, a domain traditionally constrained by limited annotated datasets and by the inherently subjective, context-dependent, and multimodal nature of affective signals [87]. Unlike supervised approaches relying on explicit emotion labels, which are often noisy or inconsistent, self-supervised strategies exploit intrinsic data properties, such as masked prediction, temporal coherence, and cross-modal correspondence, to learn robust representations without direct annotation [88].

Recent developments have demonstrated the applicability of these approaches to emotion recognition tasks across different modalities. In particular, self-supervised and multimodal architectures have been applied to video, speech, and physiological signals, showing improved generalization across datasets and robustness to inter-subject variability [89–91]. These characteristics are especially relevant for affective computing and APA, where emotional and nociceptive expressions are dynamic, heterogeneous, and often partially observable.

Emotional expression rarely emerges from a single modality; rather, it results from the interaction of facial dynamics, vocal features, physiological responses, and linguistic content, each contributing complementary information across time and context [89]. Foundation models trained on multimodal data are therefore well suited to capture these interactions by learning shared latent representations that preserve temporal, spatial, and semantic relationships across modalities [90, 91]. This capability enhances robustness to noise, missing data, and variability in acquisition conditions, which are common in real-world scenarios [92]. In the context of APA, such properties are particularly valuable for integrating heterogeneous clinical signals and improving the reliability of pain inference.

From a methodological perspective, foundation models support transferable and scalable learning. Pretrained representations can be adapted to new populations, tasks, or environments with limited fine-tuning, helping to address challenges such as inter-individual variability, cultural diversity, and domain shift in emotion recognition [93]. This is especially important in clinical contexts, where collecting large, well-annotated datasets is often impractical, and where models must generalize across diverse patient populations.

However, despite these advantages, the application of foundation models to APA in oncology remains limited. Most existing studies are conducted in controlled environments or non-clinical settings, and robust validation in real-world cancer populations is still lacking. This represents a significant gap in the current literature and highlights the need for prospective, clinically grounded investigations to assess the translational potential of these approaches [12, 13, 40, 94].

In addition, foundation models introduce several critical challenges. The interpretability of learned representations remains limited, as the internal mechanisms of large deep architectures are often opaque, hindering the identification of spurious correlations and reducing trust in sensitive applications [95]. Training on large-scale, weakly curated datasets may also amplify demographic and cultural biases, leading to uneven performance across population subgroups and raising fairness concerns [96]. Furthermore, the multidimensional nature of emotions, including intensity, temporal evolution, and contextual modulation, poses difficulties for evaluation, as current benchmarks often rely on simplified categorical labels that do not capture the continuous and dynamic structure of affective states [34].

Recent methodological advances aim to address these limitations by integrating relational and contextual reasoning into emotion modeling. Approaches combining foundation models with graph-based representations, attention mechanisms, and knowledge-driven reasoning have been proposed to better capture long-range dependencies, conversational context, and social interactions [97–99]. These developments suggest that future foundation models for affective computing and APA will increasingly combine structured knowledge with data-driven learning to enhance both interpretability and performance.

Overall, foundation models trained with self-supervised objectives represent a powerful methodological advance for emotion understanding, enabling scalable, multimodal, and transferable representation learning [87, 88, 100]. However, their effective adoption requires careful attention to interpretability, bias mitigation, and evaluation rigor, particularly when applied to clinically sensitive domains such as cancer pain assessment [101, 102]. Interestingly, early work on unified normalization frameworks for heterogeneous multimedia data provides a conceptual basis for consistent cross-modal reasoning, which remains relevant for contemporary multimodal foundation models [103].

Future directions

Despite significant advances, no single sensor or method is capable of fully capturing the complexity of human emotional states, due to overlapping physiological patterns and strong context dependency [104]. Facial expressions, physiological responses, and language-based cues each provide partial and context-dependent information and are subject to modality-specific limitations, including noise, inter-individual variability, and environmental constraints [12, 13, 39–41]. Consequently, multimodal approaches that integrate complementary signals through robust fusion and alignment strategies are essential to achieve reliable and clinically meaningful APA systems. Nevertheless, a major limitation in affective computing, particularly in clinical contexts, is the scarcity of large, high-quality emotion-labeled datasets, which constrains model generalizability and reproducibility [105].

Another central insight concerns the gap between methodological innovation and real-world clinical applicability. Although advanced ML and DL models have demonstrated promising performance in controlled experimental settings, their translation into oncology practice remains limited by issues of generalizability, interpretability, and bias. These challenges are particularly relevant in cancer pain management, where AI-driven inferences may influence monitoring strategies and therapeutic decisions [95, 96].

An additional critical dimension for the clinical translation of emotion-aware AI systems concerns regulatory approval and governance. Systems designed for APA may be classified as Software as a Medical Device (SaMD) and must therefore comply with regulatory frameworks such as FDA clearance in the United States and CE marking under the European Medical Device Regulation (MDR). These pathways require rigorous evidence of safety, performance, and clinical benefit, as well as transparency in the model design

and validation processes. In this context, several challenges emerge. AI models must demonstrate robustness across diverse patient populations and clinical settings, ensuring generalizability and minimizing bias. Continuous performance monitoring and post-market surveillance are also essential, particularly for adaptive or continuously learning systems. Furthermore, the integration of explainable AI approaches may support regulatory acceptance by improving interpretability and clinician trust. Addressing these regulatory requirements is a key step toward the safe and effective deployment of AI-based emotion recognition systems in oncology practice [106].

From a translational perspective, emotion-aware AI systems should be conceived as decision-support tools that augment, rather than replace, clinical judgment. When rigorously validated and appropriately integrated into clinical workflows, such systems may support continuous pain monitoring, enable earlier detection of pain exacerbations, and contribute to more personalized and adaptive management strategies, especially in vulnerable populations [12, 13, 107].

Looking forward, future research should prioritize longitudinal and real-world validation studies, the development of explainable and bias-aware multimodal models, and the seamless integration of APA frameworks into digital health infrastructures. In this complex scenario, strengthening interdisciplinary collaboration among clinicians, affective scientists, and AI researchers will be critical to ensure that methodological advances translate into tangible clinical benefits.

Conclusions

Taken together, the reviewed evidence confirms that affective processes constitute a fundamental yet under-assessed dimension of cancer pain, with a significant impact on symptom perception, coping strategies, and treatment outcomes. Therefore, automated emotion recognition represents both a major opportunity and a methodological challenge for AI-driven cancer pain assessment. Continued progress in this field will depend not only on algorithmic performance but also on clinical relevance, transparency, and ethical robustness, ultimately shaping the role of affective computing in personalized cancer pain care. Finally, future research should address existing technical limitations by integrating technological advances with the intrinsic complexity of emotion interpretation, acknowledging the subjective, context-dependent, and culturally mediated nature of emotional experience.

Abbreviations

AI: artificial intelligence

APA: automatic pain assessment

AUs: action units

CNNs: convolutional neural networks

DL: deep learning

ECG: electrocardiography

FACS: Facial Action Coding System

FER: facial expression recognition

GSR: galvanic skin response

LSTM: long short-term memory

ML: machine learning

NLP: natural language processing

PPG: photoplethysmography

PRO: patient-reported outcome

SSL: self-supervised learning

Declarations

Author contributions

MC: Resources, Data curation, Formal analysis, Software. AZ: Resources, Data curation, Formal analysis, Software. OP: Conceptualization, Formal analysis, Data curation. CG: Investigation, Visualization. FDL: Formal analysis, Supervision. V Cerrone and AZ: Validation, Formal analysis, Writing—original draft, Writing—review & editing. AF: Methodology, Investigation. V Conti: Software, Formal analysis. GP and SC: Methodology, Software. FS, RDF, and DE: Conceptualization, Formal analysis. MPB: Validation, Investigation. All authors read and approved the submitted version.

Conflicts of interest

Marco Cascella, who is the Editorial Board Member and Guest Editor of *Exploration of Medicine*, had no involvement in the decision-making or the review process of this manuscript. The other authors declare no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

This study does not involve original data; all data analyzed are publicly available and have been appropriately cited.

Funding

This research received no external funding.

Publisher's note

References

1. Scherer KR, Moors A. The Emotion Process: Event Appraisal and Component Differentiation. Annu Rev Psychol. 2019;70:719–45. [DOI] [PubMed]
2. Jack RE, Garrod OG, Yu H, Caldara R, Schyns PG. Facial expressions of emotion are not culturally universal. Proc Natl Acad Sci U S A. 2012;109:7241–4. [DOI] [PubMed] [PMC]
3. Posner J, Russell JA, Peterson BS. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol. 2005;17:715–34. [DOI] [PubMed] [PMC]

4. Raja SN, Carr DB, Cohen M, Finnerup NB, Flor H, Gibson S, et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. Pain. 2020;161: 1976–82. [DOI] [PubMed] [PMC]
5. Wiech K. Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. Science. 2016;354:584–7. [DOI] [PubMed]
6. Murphy SE, Downham C, Cowen PJ, Harmer CJ. Direct effects of diazepam on emotional processing in healthy volunteers. Psychopharmacology (Berl). 2008;199:503–13. [DOI] [PubMed] [PMC]
7. Cascella M, Bimonte S, Saettini F, Muzio MR. The challenge of pain assessment in children with cognitive disabilities: Features and clinical applicability of different observational tools. J Paediatr Child Health. 2019;55:129–35. [DOI] [PubMed]
8. Picard RW. Affective computing. Cambridge (MA): The MIT Press; 1997. [DOI]
9. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc Natl Acad Sci U S A. 2017;114:E7900–9. [DOI] [PubMed] [PMC]
10. Schuller B, Steidl S, Batliner A, Epps J, Eyben F, Ringeval F, et al. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. Proc Interspeech. 2014:427–31. [DOI]
11. Brigham TJ. Merging Technology and Emotions: Introduction to Affective Computing. Med Ref Serv Q. 2017;36:399–407. [DOI]
12. Cascella M, Schiavo D, Cuomo A, Ottaiano A, Perri F, Patrone R, et al. Artificial Intelligence for Automatic Pain Assessment: Research Methods and Perspectives. Pain Res Manag. 2023;2023: 6018736. [DOI] [PubMed] [PMC]
13. Gkikas S, Tsiknakis M. Automatic assessment of pain based on deep learning methods: A systematic review. Comput Methods Programs Biomed. 2023;231:107365. [DOI]
14. Gkikas S, Tachos NS, Andreadis S, Pezoulas VC, Zaridis D, Gkois G, et al. Multimodal automatic assessment of acute pain through facial videos and heart rate signals utilizing transformer-based architectures. Front Pain Res (Lausanne). 2024;5:1372814. [DOI] [PubMed] [PMC]
15. Kunz M, Lautenbacher S, LeBlanc N, Rainville P. Are both the sensory and the affective dimensions of pain encoded in the face? Pain. 2012;153:350–8. [DOI] [PubMed]
16. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychol Sci Public Interest. 2019; 20:1–68. [DOI] [PubMed] [PMC]
17. Colares WG, Costa MGF, Costa Filho CFF. Enhancing Emotion Recognition: A Dual-Input Model for Facial Expression Recognition Using Images and Facial Landmarks. Annu Int Conf IEEE Eng Med Biol Soc. 2024;2024:1–5. [DOI] [PubMed]
18. Cascella M, Vitale VN, Mariani F, Iuorio M, Cutugno F. Development of a binary classifier model from extended facial codes toward video-based pain recognition in cancer patients. Scand J Pain. 2023;23: 638–45. [DOI] [PubMed]
19. Saarimäki H, Glerean E, Smirnov D, Mynttinen H, Jääskeläinen IP, Sams M, et al. Classification of emotion categories based on functional connectivity patterns of the human brain. NeuroImage. 2022;247:118800. [DOI]
20. Li S, Deng W. Deep Facial Expression Recognition: A Survey. IEEE Trans Affect Comput. 2022;13: 1195–215. [DOI]
21. Krichen M. Convolutional Neural Networks: A Survey. Computers. 2023;12:151. [DOI]
22. Vinay BV, M R, Math S. Comprehensive Study of Low Light Facial Recognition Based on Conventional and Deep Learning Models. In: 2025 3rd International Conference on Data Science and Network Security (ICDSNS). 2025. pp. 1–4. [DOI]
23. Caruccio L, Polese G, Tortora G, Iannone D. EDCAR: A knowledge representation framework to enhance automatic video surveillance. Expert Syst Appl. 2019;131:190–207. [DOI]

24. Cascella M, Shariff MN, Lo Bianco G, Monaco F, Gargano F, Simonini A, et al. Employing the Artificial Intelligence Object Detection Tool YOLOv8 for Real-Time Pain Detection: A Feasibility Study. J Pain Res. 2024;17:3681–96. [DOI] [PubMed] [PMC]
25. Wang W, Jia M. A facial expression recognition network based on attention double branch enhanced fusion. PeerJ Comput Sci. 2024;10:e2266. [DOI] [PubMed] [PMC]
26. Liang X, Xu L, Zhang W, Zhang Y, Liu J, Liu Z. A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. Vis Comput. 2022;39:2277–90. [DOI]
27. Kim J, Lee D. Facial expression recognition robust to occlusion and to intra-similarity problem using relevant subsampling. Sensors. 2023;23:2619. [DOI] [PubMed] [PMC]
28. Gao J, Zhao Y. TFE: A transformer architecture for occlusion aware facial expression recognition. Front Neurorobot. 2021;15:763100. [DOI] [PubMed] [PMC]
29. Huang ZY, Chiang CC, Chen JH, Chen YC, Chung HL, Cai YP, et al. A study on computer vision for facial emotion recognition. Sci Rep. 2023;13:8425. [DOI] [PubMed] [PMC]
30. Li Y, Liu H, Liang J, Jiang D. Occlusion-robust facial expression recognition based on multi-angle feature extraction. Appl Sci. 2025;15:5139. [DOI]
31. Zhang Y, Zheng X, Liang C, Hu J, Deng W. Generalizable facial expression recognition. In: Computer Vision – ECCV 2024: 18th European Conference; 2024 Sep 29–Oct 4; Milan, Italy. Berlin, Heidelberg: Springer-Verlag; 2024. pp. 231–48. [DOI]
32. Aly M. Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. Multimed Tools Appl. 2025;84:12575–614. [DOI]
33. Duraj A, Szczepaniak PS, Sadok A. Detection of Anomalies in Data Streams Using the LSTM-CNN Model. Sensors (Basel). 2025;25:1610. [DOI] [PubMed] [PMC]
34. Wu Y, Mi Q, Gao T. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics (Basel). 2025;10:418. [DOI] [PubMed] [PMC]
35. Pordoy J, Farman H, Dicheva NK, Anwar A, Nasralla MM, Khilji N, et al. Multi-frame transfer learning framework for facial emotion recognition in e-learning contexts. IEEE Access. 2024;12:151360–81. [DOI]
36. Elgendi M, Galli V, Ahmadizadeh C, Menon C. Dataset of psychological scales and physiological signals collected for anxiety assessment using a portable device. Data. 2022;7:132. [DOI]
37. Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K. Deep PPG: large-scale heart rate estimation with convolutional neural networks. Sensors. 2019;19:3079. [DOI] [PubMed] [PMC]
38. Nooh S, Ragab M, Aboalela R, Al-Ghamdi AA, Abdulkader OA, Alghamdi G. An exploratory analysis of longitudinal artificial intelligence for cognitive fatigue detection using neurophysiological based biosignal data. Sci Rep. 2025;15:15736. [DOI] [PubMed] [PMC]
39. Cascella M, Di Gennaro P, Crispo A, Vittori A, Petrucci E, Sciorio F, et al. Advancing the integration of biosignal-based automated pain assessment methods into a comprehensive model for addressing cancer pain. BMC Palliat Care. 2024;23:198. [DOI] [PubMed] [PMC]
40. Moscato S, Orlandi S, Giannelli A, Ostan R, Chiari L. Automatic pain assessment on cancer patients using physiological signals recorded in real-world contexts. Annu Int Conf IEEE Eng Med Biol Soc. 2022;2022:1931–4. [DOI] [PubMed]
41. Cascella M, Vitale VN, D'Antò M, Cuomo A, Amato F, Romano M, et al. Exploring Biosignals for Quantitative Pain Assessment in Cancer Patients: A Proof of Concept. Electronics. 2023;12:3716. [DOI]
42. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst. 2017;28:2222–32. [DOI]
43. Banerjee R, Ghose A. Synthesis of realistic ECG waveforms using a composite generative adversarial network for classification of atrial fibrillation. In: 2021 29th European Signal Processing Conference (EUSIPCO). 2021. pp. 1145–9. [DOI]

44. Engel E, Hudy C, Li L, Schleusner R. Multi-modal transfer learning for dynamic facial emotion recognition in the wild. arXiv [Preprint]. 2025 [cited 2026 May 5]. Available from: https://arxiv.org/pdf/2504.21248v1
45. Li Q, Li Q, Cakmak AS, Da Poian G, Bliwise DL, Vaccarino V, et al. Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables. Physiol Meas. 2021;42:044004. [DOI]
46. Ezzat A, Omer OA, Mohamed US, Mubarak AS. ECG signal reconstruction from PPG using a hybrid attention-based deep learning network. EURASIP J Adv Signal Process. 2024;2024:95. [DOI]
47. Fadhlullah BR, Nuha HH, Putrada AG. Quality control for PPG-based hypertension detection: an imbalance-aware deep learning approach. In: 2025 International Seminar on Intelligent Technology and Its Applications (ISITIA); 2025. pp. 795–800. [DOI]
48. Zhang H, Wang Z, Zhuang Y, Yin S, Chen Z, Liang Y. Assessment of mental workload level based on PPG signal fusion continuous wavelet transform and cardiopulmonary coupling technology. Electronics. 2024;13:1238. [DOI]
49. Luo Y, Chen Y, Salekin A, Rahman T. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals. ArXiv [Preprint]. 2025 [cited 2026 Feb 24]. Available from: https://doi.org/10.48550/arXiv.2412.09758
50. Webster MB, Lee D, Lee J. Self-supervised autoencoder network for robust heart rate extraction from noisy photoplethysmogram: Applying blind source separation to biosignal analysis. Comput Biol Med. 2025;199:111319. [DOI]
51. Napoletano F, Cutugno F, Cascella M, Maffia M. Can you describe your pain? Combining psycho-emotional and pragmatic analysis on cancer patients' utterances. Int J Linguist. 2025;17:86–103. [DOI]
52. Calvo RA, D'Mello S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Trans Affect Comput. 2010;1:18–37. [DOI]
53. Fisher H, Jaffe N, Pidvirny K, Tierney A, Pizzagalli D, Webb C. Using Natural Language Processing to Track Negative Emotions in the Daily Lives of Adolescents. Res Sq [Preprint]. 2025 [cited 2026 Feb 24]. Available from: https://doi.org/10.21203/rs.3.rs-6414400/v1
54. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv [Preprint]. 2019 [cited 2026 Feb 24]. Available from: https://doi.org/10.48550/arXiv.1907.11692
55. Mohammad SM, Turney PD. Crowdsourcing a Word–Emotion Association Lexicon. Comput Intell. 2013;29:436–65. [DOI]
56. Godard R, Holtzman S. The multidimensional lexicon of emojis: a new tool to assess the emotional content of emojis. Front Psychol. 2022;13:921388. [DOI]
57. Shen J, Zhang S, Tong Y, Dong X, Wang X, Fu G, et al. Establishment and psychometric characteristics of emotional words list for suicidal risk assessment in speech emotion recognition. Front Psychiatry. 2022;13:1022036. [DOI]
58. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med. 2022;5:46. [DOI]
59. Sakai H, Lam SS. Large Language Models for Health Care Text Classification: Systematic Review. JMIR AI. 2026;5:e79202. [DOI]
60. Baltrušaitis T, Ahuja C, Morency LP. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41:423–43. [DOI]
61. Kächele M, Thiam P, Amirian M, Werner P, Walter S, Schwenker F, et al. Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity. In: Iliadis L, Jayne C, editors. Engineering Applications of Neural Networks. EANN 2015. Communications in Computer and Information Science. Springer, Cham; 2015. pp. 275–85. [DOI]

62. Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO corpus: an Italian emotional speech database. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 3501–4.
63. Cascella M, Cutugno F, Mariani F, Vitale VN, Iuorio M, Cuomo A, et al. AI-based cancer pain assessment through speech emotion recognition and video facial expressions classification. Signa Vitae. 2024;20;28–38. [DOI]
64. Limbrecht-Ecklundt K, Werner P, Traue HC, Al-Hamadi A, Walter S. Mimic activity of differentiated pain intensities: Correlation of characteristics of Facial Action Coding System and electromyography. Schmerz. 2016;30:248–56. German. [DOI] [PubMed]
65. Lopez-Martinez D, Picard R. Continuous Pain Intensity Estimation from Autonomic Signals with Recurrent Neural Networks. Annu Int Conf IEEE Eng Med Biol Soc. 2018;2018:5624–7. [DOI]
66. Chen Y, Qiu Z, Meng F, Li H, Xu L, Wu Q. Leveraging pre-trained models for multimodal class-incremental learning under adaptive fusion. In: ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025. pp. 1–5. [DOI]
67. Yang Z. Research on multi-source data fusion analysis model based on deep learning. In: 2025 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS). 2025; pp. 1178–83. [DOI]
68. Yue Y. Multimodal learning data fusion and analysis based on self-attention mechanism. In: 2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI). 2025. pp. 1040–7. [DOI]
69. Zhang J, Xue S, Wang X, Liu J. Survey of multimodal sentiment analysis based on deep learning. In: 2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS). 2023. pp. 446–50. [DOI]
70. Shivappa ST, Trivedi MM, Rao BD. Audiovisual information fusion in human–computer interfaces and intelligent environments: a survey. Proc IEEE. 2010;98:1692–715. [DOI]
71. Islam MM, Yasar MS, Iqbal T. MAVEN: a memory augmented recurrent approach for multimodal fusion. IEEE Trans Multimedia. 2023;25:3694–708. [DOI]
72. Yu J, Pu J, Cheng Y, Feng R, Shan Y. Learning music-dance representations through explicit-implicit rhythm synchronization. IEEE Trans Multimedia. 2024;26:8454–63. [DOI]
73. Wu Y, Jia T, Yang B, Li W, Yang T. Design of acquisition system based on RGB-D-T multi-modal images and research on alignment techniques. In: 2024 IEEE 14th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). 2024. pp. 260–5. [DOI]
74. Freeman ED, Ipser A, Palmbaha A, Paunoiu D, Brown P, Lambert C, et al. Sight and sound out of synch: fragmentation and renormalisation of audiovisual integration and subjective timing. Cortex. 2013;49:2875–87. [DOI]
75. Al-Dailami A, Kuang H, Wang J. Multimodal representation learning based on personalized graph-based fusion for mortality prediction using electronic medical records. Big Data Min Anal. 2025;8: 933–50. [DOI]
76. Cai L, Zeng W, Chen H, Zhang H, Li Y, Feng Y, et al. MM-GTUNets: unified multi-modal graph deep learning for brain disorders prediction. IEEE Trans Med Imaging. 2025;44:3705–16. [DOI]
77. Wang J, Zhang O, Jiang Y. Multimodal diffusion framework for collaborative text image audio generation and applications. Sci Rep. 2025;15:20604. [DOI] [PubMed] [PMC]
78. Ai W, Shou Y, Meng T, Li K. DER-GCN: dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Trans Neural Netw Learn Syst. 2025;36: 4908–21. [DOI]
79. Boie SD, Giesa N, Sekutowicz M, Zhumagambetov R, Haufe S, Grünewald E, et al. Multimodal data for predictive medicine: algorithmic fusion of clinical data in anesthesiology and intensive care. Front Med. 2026;13:1746867. [DOI]

80. Khader F, Kather JN, Müller-Franzes G, Wang T, Han T, Tayebi Arasteh S, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. Sci Rep. 2023;13:10666. [DOI]
81. Xiang H, Zhang H, Cheng Y, Quan X, Huang W. SMFusion: Semantic-Preserving Fusion of Multimodal Medical Images for Enhanced Clinical Diagnosis. IEEE J Biomed Health Inform. 2025;PP. [DOI]
82. Intriago JA, Estevez P, Cortes-Briones JA, Okuma CA, Henriquez F, Lillo P, et al. Detecting early risk of Alzheimer's disease using self-supervised multimodal representation learning. In: 2023 IEEE Conference on Artificial Intelligence (CAI). 2023. pp. 158–60. [DOI]
83. Shinde RK, Sodhi A, Mane PB, Mehta H. Adaptive multimodal learning for robot decision-making in dynamic environments. In: 2025 11th International Conference on Control, Automation and Robotics (ICCAR); 2025. pp. 157–62. [DOI]
84. Dimakatso T, Kuthadi V, Selvaraj R, Dinakenyane O. Pragmatic review on progressions in multimodal disease prediction with combination of machine learning, deep learning and electronic health records. In: 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG). 2024. pp. 1–7. [DOI]
85. Lee CH, Kim H, Yoon BC, Kim DJ. Toward foundational model for sleep analysis using a multimodal hybrid-self-supervised learning framework. IEEE Trans Cybern. 2025;55:5619–32. [DOI]
86. Li Y, Chen J, Li F, Fu B, Wu H, Ji Y, et al. GMSS: graph-based multi-task self-supervised learning for EEG emotion recognition. IEEE Trans Affect Comput. 2023;14:2512–25. [DOI]
87. Rajan V, Brutti A, Cavallaro A. Robust latent representations via cross-modal translation and alignment. In: ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021. pp. 4315–9. [DOI]
88. Pa Aung KP, Yin HL, Ma TF, Zheng WL, Lu BL. A multimodal Myanmar emotion dataset for emotion recognition. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2024. pp. 1–4. [DOI]
89. Gao P, Liu T, Liu JW, Lu BL, Zheng WL. Multimodal multi-view spectral-spatial-temporal masked autoencoder for self-supervised emotion recognition. In: ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. pp. 1926–30. [DOI]
90. Hou M, Zhang Z, Liu C, Lu G. Semantic alignment network for multi-modal emotion recognition. IEEE Trans Circuits Syst Video Technol. 2023;33:5318–29. [DOI]
91. Heo S, Kyung J, Chang JH. Multimodal emotion recognition with target speaker-based facial embeddings. In: ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025. pp. 1–5. [DOI]
92. Qiu F, Du H, Zhang W, Liu C, Li L, Guo T, et al. Learning transferable compound expressions from masked autoencoder pretraining. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024. pp. 4733–41. [DOI]
93. Hjuler MJ, Clemmensen LH, Das S. Exploring local interpretable model-agnostic explanations for speech emotion recognition with distribution-shift. In: ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025. pp. 1–5. [DOI]
94. Cascella M, Perri F, Ottaiano A, Santorsola M, Marciano ML, Rampetta FR, et al. Linking Cancer Pain Features and Biosignals for Automatic Pain Assessment. Cancers (Basel). 2026;18:646. [DOI] [PubMed] [PMC]
95. Xu Y, Pinkney JNM, Yang YL, Shao T, Zhou K. Emotion amplification of facial videos using a fine-tuned StyleGAN. Comput Vis Media. 2025;11:587–601. [DOI]
96. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health. 2024;3:e0000651. [DOI] [PubMed] [PMC]

97. Han B, Yau C, Lei S, Gratch J. Knowledge-based emotion recognition using large language models. In: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII); 2024. pp. 1–9. [DOI]
98. Wang H, Kim DH. Graph Neural Network-Based Speech Emotion Recognition: A Fusion of Skip Graph Convolutional Networks and Graph Attention Networks. Electronics. 2024;13:4208. [DOI]
99. Liu J, Li J, Dong J, Mo Z, Liu N, Li Q, et al. Adaptive Graph Learning with Multimodal Fusion for Emotion Recognition in Conversation. Biomimetics (Basel). 2025;10:414. [DOI] [PubMed] [PMC]
100. Sukumar A, Desai A, Singhal P, Gokhale S, Jain DK, Walambe R, et al. Training against disguises: addressing and mitigating bias in facial emotion recognition with synthetic data. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). 2024. pp. 1–6. [DOI]
101. Bodyanskiy Y, Kulishova N, Malysheva D. The multidimensional extended neo-fuzzy system and its fast learning for emotions online recognition. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP); 2018. pp. 473–7. [DOI]
102. Shehada D, Tawfik H, Bouridane A, Hussain A. An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition. Sensors (Basel). 2025;25:7320. [DOI] [PubMed] [PMC]
103. Chang SK, Deufemia V, Polese G, Vacca M. A Normalization Framework for Multimedia Databases. IEEE Trans Knowl Data Eng. 2007;19:1666–79. [DOI]
104. Dzedzickis A, Kaklauskas A, Bucinskas V. Human Emotion Recognition: Review of Sensors and Methods. Sensors (Basel). 2020;20:592. [DOI] [PubMed] [PMC]
105. Mohammad GB, Potluri S, Kumar A, A RK, P D, Tiwari R, et al. An Artificial Intelligence-Based Reactive Health Care System for Emotion Detections. Comput Intell Neurosci. 2022;2022:8787023. [DOI] [PubMed] [PMC]
106. Montomoli J, Bitondo MM, Cascella M, Rezoagli E, Romeo L, Bellini V, et al. Algor-ethics: charting the ethical path for AI in critical care. J Clin Monit Comput. 2024;38:931–9. [DOI]
107. Fang J, Wu W, Liu J, Zhang S. Deep learning-guided postoperative pain assessment in children. Pain. 2023;164:2029–35. [DOI] [PubMed] [PMC]