Exploration of Medicine

Open Exploration

Open Access Systematic Review

Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review

Marco Cascella[1*], Dalila Esposito[1], Maria Rosaria Muzio[2], Vincenzo Cascella[3], Valentina Cerrone[1]

[1]Department of Medicine, University of Salerno, 84081 Salerno, Italy
[2]Division of Infantile Neuropsychiatry, UOMI-Maternal and Infant Health, Asl Napoli 3 Sud, Torre del Greco, 80059 Naples, Italy
[3]School of Medicine, University of Pavia, 27100 Pavia, Italy

***Correspondence:** Marco Cascella, Department of Medicine, University of Salerno, 84081 Salerno, Italy. mcascella@unisa.it
Academic Editor: Min Cheol Chang, Yeungnam University, Republic of Korea
Received: August 27, 2025 **Accepted:** October 10, 2025 **Published:** November 12, 2025

Cite this article: Cascella M, Esposito D, Muzio MR, Cascella V, Cerrone V. Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review. Explor Med. 2025;6:1001370. https://doi.org/10.37349/emed.2025.1001370

Abstract

Background: Although accurate pain assessment is crucial in clinical care, pain evaluation is traditionally based on self-report or observer-based scales. Artificial intelligence (AI) applied to facial expression recognition is promising for objective, automated, and real-time pain assessment.

Methods: The study followed PRISMA guidelines. We searched PubMed/MEDLINE, Scopus, Web of Science, Cochrane Library, and the IEEE Xplore databases for the literature published between 2015 and 2025 on the applications of AI for pain assessment via facial expression analysis. Eligible studies included original articles in English applying different AI techniques. Exclusion criteria were neonatal/pediatric populations, non-facial approaches, reviews, case reports, letters, and editorials. Methodological quality was assessed using the RoB 2 tool (for RCTs) and adapted appraisal criteria for AI development studies. This systematic review was registered in PROSPERO (https://doi.org/10.17605/OSF.IO/N9PZA).

Results: A total of 25 studies met the inclusion criteria. Sample sizes ranged from small experimental datasets ($n < 30$) to larger clinical datasets ($n > 500$). AI strategies included machine learning models, convolutional neural networks (CNNs), recurrent neural networks such as long short-term memory (LSTM), transformers, and multimodal fusion models. The accuracy in pain detection varied between ~70% and > 90%, with higher performance observed in deep learning and multimodal frameworks. The risk of bias was overall moderate, with frequent concerns related to small datasets and lack of external validation. No meta-analysis was performed due to heterogeneity in datasets, methodologies, and outcome measures.

Discussion: AI-based facial expression recognition shows promising accuracy for automated pain assessment, particularly in controlled settings and binary classification tasks. However, evidence remains limited by small sample sizes, methodological heterogeneity, and scarce external validation. Large-scale

© The Author(s) 2025. This is an Open Access article licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

multicenter studies are required to confirm clinical applicability and to strengthen the certainty of evidence for use in diverse patient populations.

Keywords

artificial intelligence, facial expression recognition, pain assessment, action units

Introduction

Pain assessment is a crucial aspect of healthcare. Since it traditionally relies on subjective assessments and observable symptoms, potential inaccuracies and delays in effective intervention can occur [1]. In recent years, automatic pain assessment (APA) has emerged as a research area of significant interest [1, 2]. This complex set of approaches is aimed at objectively evaluating pain. Therefore, APA can enable individualized, patient-centered care, helping healthcare providers and caregivers to develop timely and appropriate interventions for improving pain management and quality of life [3].

Interestingly, different strategies have been implemented for APA. They encompass biosignal-based investigations and behavior-based approaches. Recognition of facial expressions is the most investigated behavior for APA [4]. Given the significant advancements in the field of automatic facial image analysis through computer vision models, research on pain detection from facial expressions is encouraged [5]. Specifically, many approaches focus on the analysis of action units (AUs). They are the smallest visually observable facial muscle movements codified by the Facial Action Coding System (FACS). It is a standardized tool where each AU corresponds to the activation of one or more facial muscles [3–5]. Furthermore, in the context of multimodal strategies for APA research, AUs are often combined with other behaviors or biosignals, such as electrocardiography (ECG)-derived parameters, electrodermal activity (EDA), photoplethysmography (PPG), respiratory rate, and vocal features [6, 7].

Nevertheless, despite their interesting promises, the reliable application of these methods in pain evaluation is still an open challenge. These unresolved research questions concern the selection and standardization of datasets, the choice of artificial intelligence (AI) models and architectures, the design of processing pipelines, and the need for both internal and external validation in real-world conditions. Moreover, their applicability across different clinical contexts, such as acute versus chronic pain, and in diverse care settings, from emergency departments to palliative care, should be carefully investigated [8, 9].

Objective

The objective was to summarize and critically evaluate the evidence of published studies between 2015 and 2025 on the application of AI for pain assessment through the analysis of facial expressions. The review addressed the following question: “What AI-based methods have been applied to detect or assess pain from facial expressions, and what is their accuracy and applicability in clinical or experimental settings?”.

Materials and methods

Literature search strategy

The search strategy was developed in line with the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines [10].

The systematic search was conducted in the main biomedical and computer databases, including PubMed/MEDLINE, Scopus, Cochrane Library, and Web of Science, integrated by IEEE Xplore for the literature in engineering and computer science. To ensure completeness, references of the selected articles and relevant citations were also manually checked. Only studies published in the period 1 January 2015–31 July 2025 were considered. The latest electronic search was completed in July 2025. Search records were managed in the Rayyan software [11].

Eligibility criteria

We included articles addressing techniques for APA through facial expression analysis. Studies conducted on adult patients of different clinical conditions, without restrictions related to the care setting (e.g., hospital, intensive care, outpatient, or experimental setting), were considered eligible. A mandatory comparator was not required; however, when present, traditional pain assessment tools (e.g., visual analogue scale, numeric rating scale, clinical observation) were evaluated.

Studies not related to the assessment of pain, opinion-only articles, editorials, conference abstracts without complete data, works that used non-AI-based approaches, or that employed AI in contexts unrelated to facial recognition were excluded.

Only peer-reviewed studies published in English were included. Unpublished articles, not peer-reviewed manuscripts, technical reports, or grey literature were not considered.

Search strategy

The search strategy was developed according to the PCC (Population, Concept, Context) framework and tailored for each database.

The search terms combined controlled vocabulary (e.g., MeSH) and free-text words related to artificial intelligence ("artificial intelligence", "machine learning", "deep learning", "neural networks", "computer vision"), pain assessment ("pain assessment", "pain detection", "pain evaluation", "pain monitoring", "pain recognition", "pain quantification", "pain scoring"), and facial expressions ("facial expression recognition", "facial expressions", "face recognition", "emotion recognition", "facial coding", "nonverbal communication", "visual perception").

Filters were applied to restrict results to:

- Years of publication: 2015–2025;
- Language: English;
- Study design: primary research articles (excluding reviews, case reports, letters, and editorials);
- Population: adult humans (excluding neonatal and adolescent populations).

Selection process and data collection

After removing duplicates, titles, and abstracts of all retrieved searches were independently screened by two reviewers (V Cerrone, MC) to assess if they met the inclusion criteria. Full texts of the included studies were then retrieved and reviewed for final selection. Disagreements were resolved through consultation with a third reviewer (DE). For each included study, two reviewers (V Cascella, MRM) independently extracted data using a standardized form developed for this review. Extracted information included study characteristics (authors, year, setting, population), methodological details, AI approaches applied, datasets used, and main outcomes. Discrepancies in data extraction were resolved through consensus.

Data items

For each included study, we extracted information on study characteristics (authors, year of publication, country, and setting), study population (sample size, age range, clinical condition), methodological design, AI approach (machine learning or deep learning algorithm used), type of dataset employed (public or clinical), and main outcomes. Outcomes of interest included accuracy, sensitivity, specificity, F1-score, area under the receiver operating characteristic curve (AUROC), and other reported performance metrics. Additional variables collected were funding sources, validation strategy (e.g., cross-validation, external validation), and whether the study involved real-world clinical implementation.

Study risk of bias assessment

Two reviewers (MRM, V Cascella) independently rated the methodological quality of all included studies using the Revised Cochrane risk-of-bias tool for randomized trials (RoB 2) [12]. As recommended by the

guidelines, the risk of bias was described and assessed for all primary outcomes evaluated in each study. Any disagreements about the methodological quality were resolved through consensus.

Effect measures

The primary effect measures considered were accuracy, sensitivity, specificity, F1-score, AUROC, and other key AI metrics, as reported in the original studies.

Synthesis methods

Given the methodological heterogeneity across studies (e.g., different AI algorithms, datasets, and outcomes), a narrative synthesis was conducted. Studies were grouped by type of dataset (clinical vs. benchmark/public datasets) and by AI method applied (machine learning vs. deep learning). Results were tabulated to allow for structured comparison of study characteristics and main findings. No quantitative meta-analysis was conducted due to variability in outcome measures and study designs.

Reporting bias assessment

Asymmetry or reporting bias was not formally assessed through statistical tests, since no meta-analysis was performed. However, we noted whether studies selectively reported only favorable performance metrics or failed to provide confidence intervals.

Certainty assessment

Given the methodological nature of the review and the lack of homogeneity across outcomes, the certainty of the body of evidence was not graded using GRADE. Instead, emphasis was placed on highlighting recurring strengths and limitations of the included studies, as well as identifying areas of consistency versus heterogeneity.

Results

Study selection

The literature search generated 79 references (Scopus = 16, PubMed = 31, Web of Science = 4, Cochrane Library = 2, IEEE = 26), of which 25 studies met our inclusion criteria. Numbers and reasons for exclusion at each stage are reported in Figure 1. The full search strategy for each database is reported in Table S1.

During the full-text review, four studies were excluded despite initially appearing to meet the inclusion criteria. The study by Chan et al. [13] was excluded because it was only an abstract (conference proceeding) and did not provide sufficient methodological details for inclusion. Other manuscripts were excluded [4, 14–16]. Specifically, they failed to focus on facial expression-based pain assessment [4, 14, 16], or we found a lack of AI strategies [15].

Study characteristics

In this systematic review, a total of 25 studies published between 2015 and 2025 were included [17–41]. Most of this scientific output concerned experimental or feasibility studies. The number of participants varied widely, ranging from small experimental samples of fewer than 30 healthy volunteers exposed to controlled pain stimuli (e.g., cold pressor or heat pain tests) to large clinical datasets exceeding 500 postoperative patients.

The population covered different contexts, such as healthy adults in laboratory conditions, surgical and perioperative patients, intensive care unit (ICU) patients, older adults with dementia, and oncology patients. While the common focus of these studies was the use of AI applied to facial expressions for the automatic detection or quantification of pain, some studies also integrated multimodal data (e.g., speech, audio, physiological signals), though facial analysis remained the primary input modality.

About the AI methods, researchers employed different AI modalities. They ranged from classical machine learning algorithms (e.g., random forests and support vector machines) to deep learning approaches. Artificial neural network models included convolutional neural networks (CNNs), often

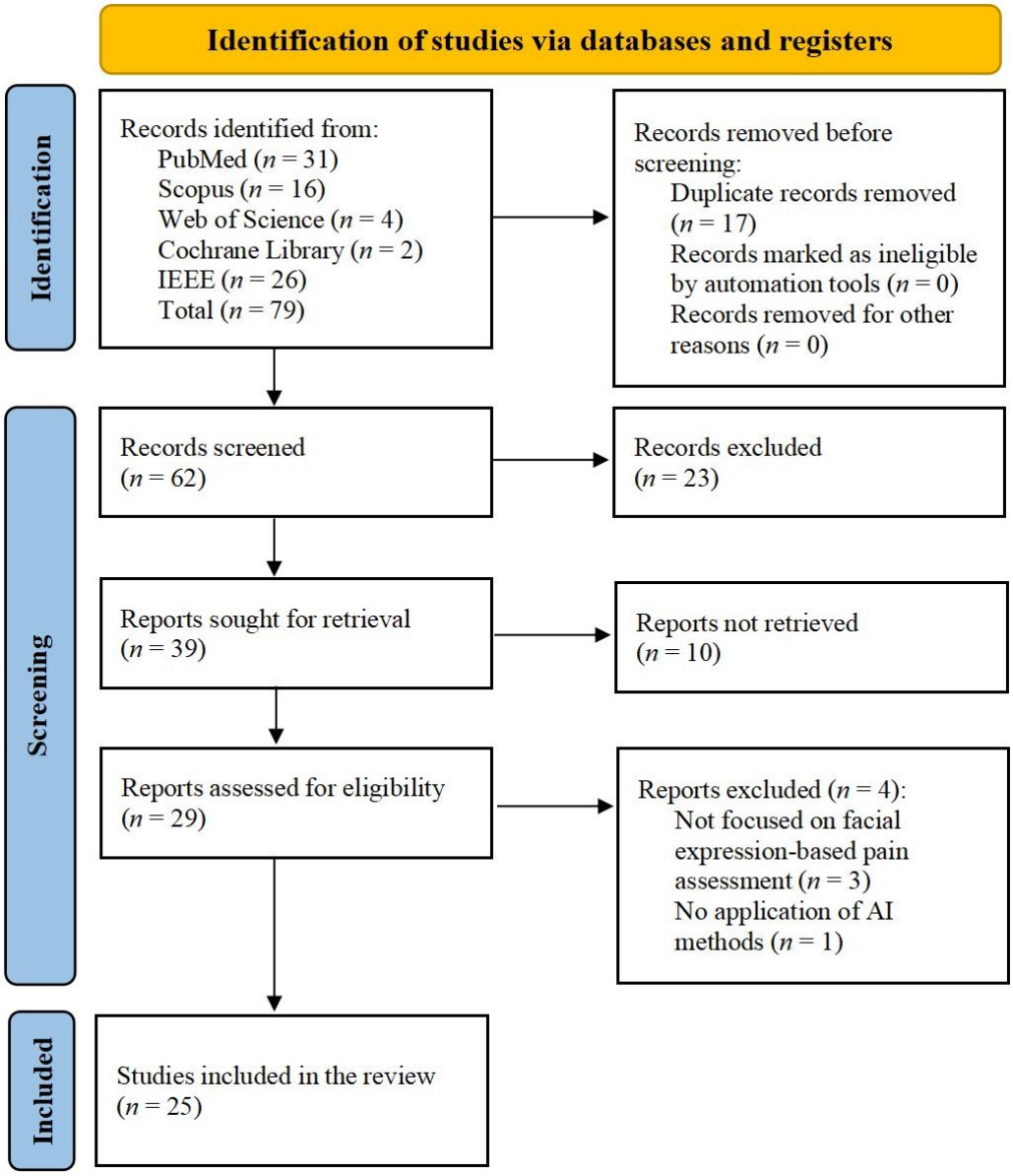

Figure 1. PRISMA flow diagram. Adapted from [10]. © 2019 The Authors. Licensed under a Creative Commons Attribution (CC BY 4.0).

combined with recurrent neural networks (RNNs) or long short-term memory (LSTM) units [18, 27, 29, 35, 37]. More recent contributions introduced vision transformers (ViTs). These architectures leverage self-attention mechanisms to process images as sequences of patches, thereby improving robustness in different challenging conditions such as ICU scenarios [17, 31]. In parallel, attention-based deep learning architectures, such as enhanced residual attention-based subject-specific network (ErAS-Net), were specifically designed to enhance feature selection and improve classification performance across datasets [33]. A distinct subgroup of studies focused on multimodal approaches, integrating facial data with additional modalities such as speech, audio, or physiological signals [20, 22, 28, 38]. These systems consistently outperformed unimodal facial analysis, particularly in challenging populations such as ICU patients with partial facial occlusion [40] or older adults with dementia [36], although at the expense of increased computational and methodological complexity. Other innovative approaches included the application of transfer entropy to landmark time-series analysis [25], surface electromyographic (sEMG) signals to capture subtle muscle activation [32], and binary classifiers trained on facial AUs derived from the FACS [21, 26].

The reported outcomes were generally expressed as accuracy, area under the curve (AUC), sensitivity, specificity, and F1-score, with many studies achieving high performance on benchmark datasets, such as Delaware [42], University of Northern British Columbia Pain Expression dataset (UNBC)-McMaster dataset [43], or on locally collected clinical datasets.

The summary of the individual study characteristics, including setting and number of patients included, population, approach, and inclusion criteria, AI method, and reported outcomes, is presented in Table 1.

Table 1. Summary of the included studies.

Source/Year	Dataset/*n*	Population	Approach/Inclusion criteria	AI method	Outcome (as reported)
Bargshady et al. [17], 2024	Lab datasets (AI4PAIN: 51; BioVid: 87)	Adults	Acute Pain Datasets (video-based)	Video vision transformers (ViViTs)	Accuracy 66.9% (AI4PAIN), 79.9% (BioVid), outperforming ResNet baselines
Bargshady et al. [18], 2020	UNBC-McMaster, MIntPAIN	Adults	Benchmark datasets	Ensemble DL model (CNN + RNN hybrid, EDLM)	Accuracy > 89%, ROC 0.93; robust vs. single-stream CNN
Bellal et al. [19], 2024	ICU, 30 patients	Critically ill, non-communicative adults	NEVVA® pilot device calibration	AI-based computer vision integrated in devices	Feasible, device calibrated against expert assessment
Benavent-Lledo et al. [20], 2023	UNBC, BioVid	Adults	Public pain expression datasets	Transformer-based computer vision	Accuracy > 96% (UNBC), > 94% (BioVid); high precision, recall
Cascella et al. [21], 2024	Oncology + public datasets (Delaware, UNBC)	Cancer patients, adults	Binary classifier using AUs	Neural network (17 AUs, OpenFace)	Accuracy ~94%; AUROC 0.98
Cascella et al. [22], 2024	Oncology	Adult cancer patients	Video + audio (facial + speech)	Multimodal AI (speech emotion + facial expression)	Feasibility shown; early accuracy promising
Cascella et al. [23], 2023	Clinical feasibility (real-time)	Adults	Real-time pain detection from facial videos	YOLOv8 object detection	Feasible, metrics reported with good accuracy (JPR)
Casti et al. [24], 2019	Clinical/Lab setting	Adults	Automatic pain detection calibration	DL-based system (CNN)	Benchmarked; addressed inter-/intra-observer variability
Casti et al. [25], 2021	Public dataset (video pain sequences)	Adults	Landmark time-series analysis	Transfer entropy (TE) + ML classifiers	TE-based approach improved accuracy, robust to noise
Chen et al. [26], 2022	UNBC + lung cancer dataset	Adults, including patients with lung cancer	Pain-related AUs	Weakly supervised MIL/MCIL	Accuracy 87%, AUC 0.94 (UNBC); validated also on clinical lung cancer data
Dutta and M [27], 2018	UNBC + live video	Adults	Real-time video-based pain recognition	Hybrid DL model	Validated in real-time; high accuracy reported
Ghosh et al. [28], 2025	UNBC, BioVid + VIVAE (audio)	Adults	Multimodal (facial + audio)	Ensemble DL with CNN + fusion	Accuracy up to 99.5% (3-class), 87.4% (5-class); audio peak 98%
Guo et al. [29], 2021	Cold pressor experiment; 29 subjects	Adults	Cold pain induction	CNN (Inception V3, VGG-LSTM, ConvLSTM)	F1 score 79.5% (personalized ConvLSTM)

Table 1. Summary of the included studies. (*continued*)

Source/Year	Dataset/*n*	Population	Approach/Inclusion criteria	AI method	Outcome (as reported)
Heintz et al. [30], 2025	Perioperative, multicenter (503 pts)	Adults perioperative	Computer vision nociception detection	CNN-based	Strong AUROC, external validation, and feasibility proven
Mao et al. [31], 2025	UNBC	Adults	Pain intensity estimation	Conv-Transformer (multi-task joint optimization)	Outperformed SOTA; improved regression + classification
Mieronkoski et al. [32], 2020	31 healthy volunteers, experimental	Adults	Pain induction + sEMG	ML (supervised on muscle activation)	Modest c-index 0.64; eyebrow/lip muscles most predictive
Morsali and Ghaffari [33], 2025	UNBC, BioVid	Adults	Public Pain Datasets	ErAS-Net (attention-based DL)	Accuracy 98.8% (binary, UNBC); 94.2% (4-class); cross-dataset BioVid 78%
Park et al. [34], 2024	155 pts post-gastrectomy	Postoperative adults	Clinical recordings	ML models (facial, ANI, vitals)	AUROC 0.93 (facial); better than ANI/vitals
Pikulkaew et al. [35], 2021	UNBC dataset	Adults	Sequential facial images	CNN (DL motion detection)	Precision: 99.7% (no pain), 92.9% (becoming pain), 95.1% (pain)
Rezaei et al. [36], 2021	Dementia patients, LTC setting	Older adults, dementia	Unobtrusive video dataset	Deep learning + pairwise/contrastive training	Outperformed baselines; validated on dementia cohort
Rodriguez et al. [37], 2022	UNBC + CK	Adults	Raw video frames	CNN + LSTM	Outperformed SOTA AUC (UNBC); competitive on CK
Semwal and Londhe [38], 2024	Multimodal dataset	Adults	Facial + multimodal integration	Multi-stream spatio-temporal network	Showed robust multiparametric pain assessment
Tan et al. [39], 2025	200 patients	Adults perioperative/interventional	Video recording (STA-LSTM)	STA-LSTM DL network	Accuracy, sensitivity, recall, F1 ≈ 0.92; clinical feasibility
Yuan et al. [40], 2024	ICU, public + 2 new datasets	Critically ill adults (ventilated)	Facial occlusion management	AU-guided CNN framework	Superior performance in binary, 4-class, regression tasks
Zhang et al. [41], 2025	503 postop patients + volunteers	Adults postoperative	Clinical Pain Dataset (CPD; 3,411 images) + Simulated Pain Dataset (CD)	VGG16 pretrained	AUROC 0.898 (CPD severe pain), 0.867 (CD); software prototype developed

AI: artificial intelligence; ResNet: Residual Network; UNBC: University of Northern British Columbia Pain Expression dataset; MIntPAIN: Multimodal International Pain dataset; DL: deep learning; CNN: convolutional neural network; RNN: recurrent neural network; EDLM: ensemble deep learning model; ROC: receiver operating characteristic; ICU: intensive care unit; NEVVA: Non-Verbal Visual Analog device; AUs: action units; AUROC: area under the receiver operating characteristic curve; YOLOv8: You Only Look Once version 8; JPR: Journal of Pain Research; ML: machine learning; AUC: area under the curve; MIL: multiple instance learning; MCIL: multiple clustered instance learning; VIVAE: Visual and Vocal Acute Expression dataset; VGG: visual geometry group; LSTM: long short-term memory; ConvLSTM: convolutional long short-term memory; SOTA: state-of-the-art; sEMG: surface electromyography; ErAS-Net: enhanced residual attention-based subject-specific network; ANI: analgesia nociception index; LTC: long-term care; CK: Cohn-Kanade dataset; STA-LSTM: Spatio-Temporal Attention Long Short-Term Memory; CD: Control Dataset.

Risk of bias in studies

Since most included studies were observational, experimental, or methodological (not RCTs), the RoB 2 tool was adapted to the specific study designs. Specifically, the category "Low risk" was assigned when the

methods, datasets, and validation were clearly reported; "Some concerns" was used when limitations such as small samples, lack of external validation, or simulation-only data were present (Table 2).

Table 2. Risk of bias of included studies.

Author/Year	Country	Intervention/AI approach	Timing	Outcomes measurement	Validation of tool (Y/N)	Quality assessment (RoB 2 overall)
Bargshady et al. [17], 2024	Australia/USA	Vision transformer	Acute pain datasets	Accuracy, comparison with baselines	Y	Low risk (well-reported external datasets)
Bargshady et al. [18], 2020	Australia/Netherlands	Ensemble CNN + RNN	Lab datasets	Accuracy, ROC	Y	Some concerns (no external clinical validation)
Bellal et al. [19], 2024	France	NEVVA® device (AI facial)	ICU pilot	Device calibration vs. experts	Y	Some concerns (small sample, feasibility only)
Benavent-Lledo et al. [20], 2023	Spain	Transformer-based CV	Lab datasets	Accuracy, F1	Y	Low risk (robust datasets, transparent methods)
Cascella et al. [21], 2024	Italy	Binary AU-based classifier	Oncology outpatient	Accuracy, AUROC	Y	Some concerns (limited clinical cohort)
Cascella et al. [22], 2024	Italy	Multimodal (speech + facial)	Clinical trial NCT04726228	Classification accuracy	Y	Low risk (registered trial, multimodal)
Cascella et al. [23], 2023	Italy	YOLOv8	Lab/clinical feasibility	Detection metrics	Y	Some concerns (pilot, limited validation)
Casti et al. [24], 2019	Italy	DL pain intensity system	Lab	Accuracy, calibration	Y	Low risk (strong methodological rigor)
Casti et al. [25], 2021	Italy	Transfer entropy + ML	Lab	Accuracy, robustness	Y	Low risk
Chen et al. [26], 2022	USA	AU combinations + MIL	Clinical + lab	Accuracy, AUC	Y	Low risk
Dutta and M [27], 2018	India	Hybrid DL	Lab + simulated	Accuracy, computational metrics	Y	Some concerns (older methods, limited clinical data)
Ghosh et al. [28], 2025	India/Switzerland	Multimodal (facial + audio)	Lab datasets	Accuracy (2–5 classes)	Y	Low risk
Guo et al. [29], 2021	China	CNN/LSTM	Cold pressor	F1 score	Y	Some concerns (small sample)
Heintz et al. [30], 2025	USA multicenter	CNN-based	Perioperative	AUROC, Brier score	Y	Low risk (robust clinical dataset)
Mao et al. [31], 2025	China	Conv-Transformer multitask	Lab	Regression + classification	Y	Low risk
Mieronkoski et al. [32], 2020	Finland	sEMG + ML	Experimental pain	c-index, features	Y	Some concerns (small sample, modest accuracy)
Morsali and Ghaffari [33], 2025	Iran/UK	ErAS-Net	Lab datasets	Accuracy, cross-dataset	Y	Low risk
Park et al. [34], 2024	Korea	ML (facial, ANI, vitals)	Postoperative	AUROC	Y	Low risk (clinical real-world)
Pikulkaew et al. [35], 2021	Thailand	CNN	Lab	Precision, accuracy	Y	Low risk
Rezaei et al. [36], 2021	Canada	DL	Long-term care	Sensitivity, specificity	Y	Low risk (validated on target population)
Rodriguez et al. [37], 2022	Spain/Denmark	CNN + LSTM	Lab	AUC, accuracy	Y	Low risk

Table 2. Risk of bias of included studies. (*continued*)

Author/Year	Country	Intervention/AI approach	Timing	Outcomes measurement	Validation of tool (Y/N)	Quality assessment (RoB 2 overall)
Semwal and Londhe [38], 2024	India	Spatio-temporal network	Lab	Accuracy	Y	Some concerns (no external validation)
Tan et al. [39], 2025	Singapore	STA-LSTM	Clinical	Accuracy, F1	Y	Low risk
Yuan et al. [40], 2024	China	AU-guided CNN	ICU, ventilated pts	Accuracy, regression	Y	Low risk
Zhang et al. [41], 2025	China	VGG16 pretrained	Postoperative	AUROC, F1	Y	Low risk

AI: artificial intelligence; CNN: convolutional neural network; RNN: recurrent neural network; ROC: receiver operating characteristic; NEVVA: Non-Verbal Visual Analog device; ICU: intensive care unit; CV: computer vision; AU: action unit; AUROC: area under the receiver operating characteristic curve; YOLOv8: You Only Look Once version 8; ML: machine learning; MIL: multiple instance learning; AUC: area under the curve; DL: deep learning; LSTM: long short-term memory; sEMG: surface electromyography; ErAS-Net: enhanced residual attention-based subject-specific network; ANI: analgesia nociception index; STA-LSTM: Spatio-Temporal Attention Long Short-Term Memory.

Overall, we assigned low risk of bias to studies with robust datasets, transparent methodology, external validation, or those conducted within the framework of a registered clinical trial. Studies judged as having some concerns were typically characterized by small sample sizes, absence of external validation, or approaches tested only in laboratory-controlled settings, which can limit their generalizability. None of the studies included clearly fell into the category of high risk of bias.

Results of individual studies

The included studies reported heterogeneous outcomes reflecting the performance of AI applied to facial expression recognition for pain assessment. Most studies evaluated models on either publicly available datasets [e.g., UNBC-McMaster Shoulder Pain, BioVid Heat Pain, Multimodal International Pain dataset (MIntPAIN)] or on original clinical cohorts (perioperative, postoperative, oncology, or ICU patients). Outcomes were mainly expressed as accuracy, AUROC, sensitivity, specificity, recall, and F1-scores. Confidence intervals were rarely reported.

Experimental and dataset-based studies demonstrated very high performance. Bargshady et al. [18] proposed an ensemble deep learning framework integrating CNN and RNN, reaching > 89% accuracy and AUC 0.93 on MIntPAIN and UNBC. Chen et al. [26] introduced a weakly supervised approach based on combinations of AUs, achieving 87% accuracy with AUC 0.94 on UNBC. Morsali and Ghaffari [33] developed ErAS-Net, an attention-based deep learning model, achieving 98.7% accuracy for binary and 94.2% for four-class pain classification, with cross-dataset validation on BioVid still showing robust results (78.1%). Rodriguez et al. [37] combined CNNs and LSTMs for temporal analysis, outperforming state-of-the-art methods on UNBC. Pikulkaew et al. [35] applied deep learning to 2D motion and expressions, achieving > 95% accuracy across three pain classes.

Transformer-based methods showed promise. Bargshady et al. [17] employed video ViTs (ViViTs) on AI4PAIN and BioVid datasets, with accuracies of 66.9% and 79.9%, outperforming CNN baselines. Mao et al. [31] refined Conv-Transformer architectures with multi-task learning, improving estimation of continuous pain intensities on UNBC.

Clinical studies confirmed feasibility in real-world settings. Zhang et al. [41] applied VGG16 to > 3,000 images from 503 postoperative patients, reporting AUROC 0.898 for severe pain detection and consistent F1-scores. Heintz et al. [30] validated CNN-based nociception recognition perioperatively, reporting robust AUROC and calibration (Brier score) across multicenter datasets. Park et al. [34] compared models in 155 gastrectomy patients, showing that facial-expression machine learning achieved an AUROC 0.93 for severe postoperative pain, outperforming analgesia nociception index (ANI) and vital signs. Bellal et al. [19] tested the NEVVA© device in ICU patients, showing the feasibility of automated detection in non-communicative patients.

Other methodological innovations included electromyography of facial muscles [32], showing modest predictive capacity (c-index 0.64), and transfer entropy applied to facial landmarks [25], which demonstrated robustness to uncertainty. Earlier, Casti et al. [24] benchmarked multi-expert calibration, confirming reproducibility of automated systems. Moreover, Dutta and M [27] provided proof-of-concept evidence for real-time video analysis.

Special populations were investigated. Rezaei et al. [36] validated a deep learning system in older adults with dementia, showing reliable detection where self-report is not possible. Yuan et al. [40] developed an AU-guided CNN for ICU patients with occluded faces (ventilation), reporting strong performance across binary, multiclass, and regression tasks. On the other hand, oncology-related contributions are limited. In 2023, Cascella et al. [23] developed a binary classifier on cancer patients' videos, achieving ~94% accuracy and an AUROC of 0.98. Later, the same authors integrated facial and vocal features in oncologic pain monitoring, confirming feasibility in clinical use [21] and subsequently also tested the You Only Look Once version 8 (YOLOv8) architecture for real-time pain detection, reporting strong detection performance [22].

Several multimodal approaches combined facial expression with additional modalities. Benavent-Lledo et al. [20] used computer vision and multimodal signals, achieving > 96% accuracy and > 94% F1-score on UNBC and BioVid. Ghosh et al. [28] proposed an IoT-enabled system integrating facial and audio data, reaching near-perfect accuracy (> 99%) in binary and three-class classification. Semwal and Londhe [38] designed a spatio-temporal behavioral system with multimodal integration, confirming the added value of combining sources.

In summary, across the 25 included studies [17–41], AI-based facial expression recognition for pain consistently demonstrated high performance in controlled datasets and increasing feasibility in clinical populations (postoperative, perioperative, oncology, ICU, dementia). The overall result ranged from modest (with an accuracy c-index of 0.64) [32] to excellent (with an accuracy > 98%) [33]. Figure 2 illustrates performance (i.e., accuracy) ranges across different AI methods.

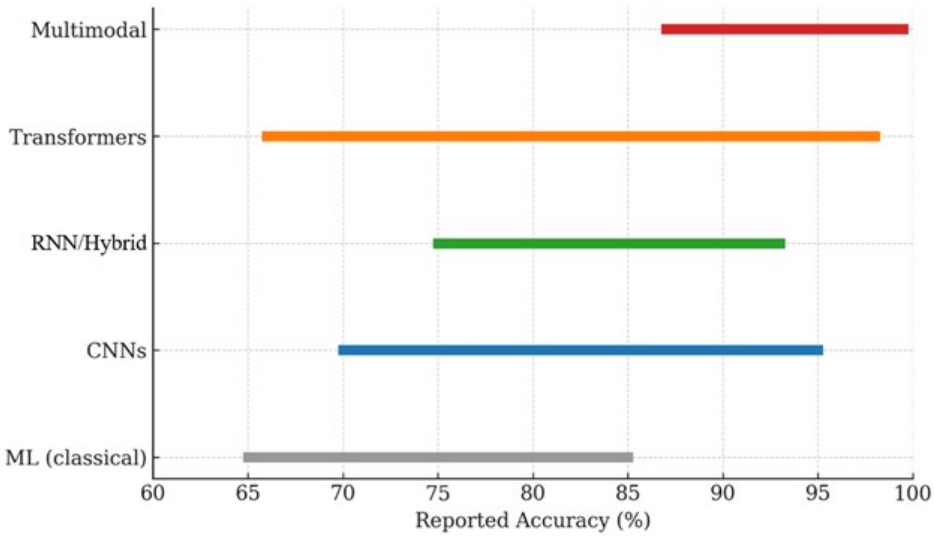

Figure 2. AI methods used for facial expression-based pain assessment (2015–2025) and their reported performance ranges (i.e., accuracies). Traditional machine learning (ML) approaches showed moderate performance (65–85%). Convolutional neural networks (CNNs) and RNN/hybrid (recurrent neural network) models achieved higher accuracy (70–95% and 75–93%, respectively). Transformer-based models reached accuracies ranging from 66% to 98%. Multimodal approaches (facial + audio and/or physiological signals) consistently outperformed unimodal systems, achieving accuracies up to 99.5%.

Discussion

This systematic review highlights both the promise and the limitations of AI applied to facial expression analysis for APA. Across 25 studies published between 2015 and 2025, AI-based systems demonstrated consistently strong performance in controlled settings, with reported accuracies often exceeding 90% and

AUROC values ranging between 0.87 and 0.98 in clinical studies [21, 30, 34, 41]. Collectively, these results support the feasibility of APA recognition and underscore the potential of AI to provide objective, real-time assessments in contexts where traditional self-report is unreliable or unavailable. On the other hand, it is important to note that not all approaches yielded successful outcomes. Some studies, for instance, reported modest or even poor performance. Specifically, performance varied markedly depending on study design, population, and dataset source. For example, models trained and validated exclusively on benchmark datasets (e.g., UNBC-McMaster, BioVid, MIntPAIN) achieved very high accuracies, sometimes exceeding 95% [18, 20, 26, 33, 35, 37], although their generalizability to clinical cohorts was limited. In contrast, studies conducted in real-world clinical populations, including perioperative patients [30, 39], oncology cohorts [21–23], ICU patients [19, 40], and older adults with dementia [36], reported slightly lower but more clinically relevant performance. Importantly, this gap is more evident for analyses relying on small datasets, limited facial visibility (e.g., ICU settings), or when applying methods such as sEMG-based models that achieved only moderate predictive value (c-index 0.64) [32]. These negative or suboptimal results highlight the fragility of certain approaches and underscore the need for robust, diverse, and clinically validated datasets. Moreover, when comparing studies using benchmark datasets with those relying on clinical populations, it emerges that while benchmark-based models frequently reported very high accuracies (> 90%) [18, 20, 26, 33, 35, 37], their clinical generalizability was often limited. Conversely, clinical studies, although achieving slightly lower performance, provided more realistic insights into real-world feasibility and robustness [19, 21–23, 30, 34, 36, 40, 41].

Multimodal approaches integrating facial expressions with speech, audio, or physiological signals generally outperformed unimodal systems [20, 22, 28, 38], particularly in challenging scenarios such as ICU patients [40] or dementia patients [36]. Nevertheless, these models also introduced greater complexity, raising issues of computational burden, implementation feasibility, and interpretability in clinical practice.

Concerning the AI-based strategies, methodological evolution across the last decade reflects a shift from classical CNN- and RNN-based pipelines to more sophisticated approaches involving transformer-based and multimodal systems. Specifically, compared to CNNs, ViTs, and hybrid Conv-Transformer models have demonstrated superior ability to capture long-range dependencies and subtle spatio-temporal dynamics in facial expressions [17, 31]. Similarly, attention-based networks such as ErAS-Net provide efficient feature selection and cross-dataset robustness, achieving accuracies close to 99% in binary pain classification [33]. In parallel, multimodal fusion architectures that integrate facial cues with audio, speech emotion recognition, or physiological signals have consistently improved F1-scores and AUROC values compared with unimodal models [20, 22, 28, 38]. These advances highlight a trend toward increasingly complex frameworks capable of addressing challenging clinical conditions, such as facial occlusion in ICU patients [40] or atypical expressions in dementia [36]. Furthermore, the integration of explainable AI (XAI) techniques and standardized multimodal datasets may represent key steps to improve transparency, scalability, and cross-institutional generalizability.

The risk of bias was overall moderate. Studies based on larger, multicentre clinical cohorts with transparent methodology and external validation were considered low risk [21–23, 30, 34, 41], while smaller experimental works without external validation raised concerns [27, 32]. Importantly, no study was deemed consistently at high risk of bias. Furthermore, selective reporting remains an issue as many experimental studies emphasized accuracy, F1-score, or AUROC [18, 26, 29, 33, 37], while failing to report misclassification rates, calibration statistics, or subgroup-specific results. The lack of registered protocols in most works further increases the risk of reporting bias and reduces comparability across studies [22, 30]. Moreover, a recurrent limitation is the reliance on small, non-representative datasets and the frequent absence of confidence intervals in reporting. These factors increase the risk of overestimating performance and reduce the robustness of the reported findings. From the analysis, other key issues emerged. For example, we underline that while systematic use of calibration metrics and external validation is crucial for clinical translation, their absence in most studies represents a major barrier to clinical adoption and should be prioritized in future research.

Multimodal models integrating facial data with physiological or audio signals outperformed unimodal approaches, particularly in challenging clinical scenarios. Moreover, the certainty of the evidence is best described as moderate for binary classification tasks (pain vs. no pain), where results are consistent across different datasets and study designs [18, 26, 33, 36], but low for more advanced applications such as pain intensity estimation [29, 31, 32] or deployment in fragile populations [19, 36, 41]. This downgrading is mainly due to small sample sizes, reliance on controlled datasets, and the absence of precision measures such as confidence intervals. Consequently, dataset-related limitations strongly affected model performance. Thus, restricted sample sizes, lack of demographic diversity, and heterogeneous annotation protocols limited reproducibility and generalizability, especially across populations with diverse ethnic and clinical backgrounds.

Despite these interesting results, several challenges should be addressed. The ethical implications of AI-based pain assessment cannot be overlooked. Patient privacy in facial video datasets, the absence of clear data sharing policies, and the limited interpretability of most deep learning models raise concerns regarding fairness, transparency, and responsible use in clinical care.

In conclusion, the findings confirm that AI can complement traditional pain assessment, particularly in patients unable to self-report. Nevertheless, key challenges remain before large-scale clinical adoption can be realized. These include the need for standardized datasets reflecting real-world heterogeneity, transparent reporting practices, and multicentre trials for evaluating AI performance across diverse populations and settings.

Abbreviations

AI: artificial intelligence

ANI: analgesia nociception index

APA: automatic pain assessment

AUC: area under the curve

AUROC: area under the receiver operating characteristic curve

AUs: action units

CNNs: convolutional neural networks

ErAS-Net: enhanced residual attention-based subject-specific network

FACS: Facial Action Coding System

ICU: intensive care unit

LSTM: long short-term memory

MIntPAIN: Multimodal International Pain dataset

RNN: recurrent neural network

sEMG: surface electromyography

UNBC: University of Northern British Columbia Pain Expression dataset

ViTs: vision transformers

ViViTs: video vision transformers

Supplementary materials

The supplementary table for this article is available at: https://www.explorationpub.com/uploads/Article/file/1001370_sup_1.pdf.

Declarations

Author contributions

MC: Writing—original draft, Formal analysis, Conceptualization, Methodology, Supervision, Writing—review & editing. DE: Writing—review & editing. MRM: Writing—review & editing. V Cascella: Writing—review & editing. V Cerrone: Writing—review & editing, Conceptualization, Methodology, Supervision. 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

The primary data for this review were sourced online from databases listed in the methods. Referenced articles are accessible on PubMed/MEDLINE, Scopus, Cochrane Library, Web of Science, and IEEE Xplore. All data are available from the corresponding author on reasonable request.

Funding

No external funding was received for this study.

Copyright

© The Author(s) 2025.

Publisher's note

Open Exploration maintains a neutral stance on jurisdictional claims in published institutional affiliations and maps. All opinions expressed in this article are the personal views of the author(s) and do not represent the stance of the editorial team or the publisher.

References

1. Albahdal D, Aljebreen W, Ibrahim DM. PainMeter: Automatic Assessment of Pain Intensity Levels From Multiple Physiological Signals Using Machine Learning. IEEE Access. 2024;12:48349–65. [DOI]
2. Cascella M, Leoni MLG, Shariff MN, Varrassi G. Artificial Intelligence-Driven Diagnostic Processes and Comprehensive Multimodal Models in Pain Medicine. J Pers Med. 2024;14:983. [DOI] [PubMed] [PMC]
3. Cao S, Fu D, Yang X, Wermter S, Liu X, Wu H. Pain recognition and pain empathy from a human-centered AI perspective. iScience. 2024;27:110570. [DOI] [PubMed] [PMC]
4. Miao S, Xu H, Han Z, Zhu Y. Recognizing Facial Expressions Using a Shallow Convolutional Neural Network. IEEE Access. 2019;7:78000–11. [DOI]
5. Hassan T, Seus D, Wollenberg J, Weitz K, Kunz M, Lautenbacher S, et al. Automatic Detection of Pain from Facial Expressions: A Survey. IEEE Trans Pattern Anal Mach Intell. 2021;43:1815–31. [DOI] [PubMed]

6. Zen G, Porzi L, Sangineto E, Ricci E, Sebe N. Learning Personalized Models for Facial Expression Analysis and Gesture Recognition. IEEE Trans. Multimedia. 2016;18:775–88. [DOI]
7. Gkikas S, Tsiknakis M. Automatic assessment of pain based on deep learning methods: A systematic review. Comput Methods Programs Biomed. 2023;231:107365. [DOI] [PubMed]
8. Cascella M, Shariff MN. Why is applying artificial intelligence to pain so challenging? Curr Med Res Opin. 2024;40:2021–4. [DOI] [PubMed]
9. Cascella M, Ponsiglione AM, Santoriello V, Romano M, Cerrone V, Esposito D, et al. Expert consensus on feasibility and application of automatic pain assessment in routine clinical use. J Anesth Analg Crit Care. 2025;5:29. [DOI] [PubMed] [PMC]
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. [DOI] [PubMed] [PMC]
11. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:210. [DOI] [PubMed] [PMC]
12. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020;126:37–44. [DOI] [PubMed]
13. Chan DXH, Tan CW, Du T, Teo JC, Wong J, Tan YR, et al. LP001 Automated pain detection via facial expression for adult patients using artificial intelligence. Reg Anesth Pain Med. 2024;49:A410. [DOI]
14. López M, Palacios-Arias C, Romeu J, Jofre-Roca L. 3D Pain Face Expression Recognition Using a ML-MIMO Radar Profiler. IEEE Access. 2024;12:48266–76. [DOI]
15. Pu L, Coppieters MW, Smalbrugge M, Jones C, Byrnes J, Todorovic M, et al. Associations between facial expressions and observational pain in residents with dementia and chronic pain. J Adv Nurs. 2024;80:3846–55. [DOI] [PubMed]
16. Uddin MT, Zamzmi G, Canavan S. Cooperative Learning for Personalized Context-Aware Pain Assessment From Wearable Data. IEEE J Biomed Health Inform. 2023;27:5260–71. [DOI] [PubMed]
17. Bargshady G, Joseph C, Hirachan N, Goecke R, Rojas RF. Acute Pain Recognition from Facial Expression Videos using Vision Transformers. Annu Int Conf IEEE Eng Med Biol Soc. 2024;2024:1–4. [DOI] [PubMed]
18. Bargshady G, Zhou X, Deo RC, Soar J, Whittaker F, Wang H. Ensemble neural network approach detecting pain intensity from facial expressions. Artif Intell Med. 2020;109:101954. [DOI] [PubMed]
19. Bellal M, Lelandais J, Chabin T, Heudron A, Gourmelon T, Bauduin P, et al. Calibration trial of an innovative medical device (*NEVVA*©) for the evaluation of pain in non-communicating patients in the intensive care unit. Front Med (Lausanne). 2024;11:1309720. [DOI] [PubMed] [PMC]
20. Benavent-Lledo M, Mulero-Pérez D, Ortiz-Perez D, Rodriguez-Juan J, Berenguer-Agullo A, Psarrou A, et al. A Comprehensive Study on Pain Assessment from Multimodal Sensor Data. Sensors (Basel). 2023;23:9675. [DOI] [PubMed] [PMC]
21. 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]
22. Cascella M, Shariff MN, Bianco GL, 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]
23. 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]

24. Casti P, Mencattini A, Comes MC, Callari G, Di Giuseppe D, Natoli S, et al. Calibration of Vision-Based Measurement of Pain Intensity With Multiple Expert Observers. IEEE Trans Instrum. Meas. 2019;68: 2442–50. [DOI]
25. Casti P, Mencattini A, Filippi J, D'Orazio M, Comes MC, Giuseppe DD, et al. Metrological Characterization of a Pain Detection System Based on Transfer Entropy of Facial Landmarks. IEEE Trans Instrum Meas. 2021;70:1–8. [DOI]
26. Chen Z, Ansari R, Wilkie DJ. Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning. IEEE Trans Affect Comput. 2022;13:135–46. [DOI] [PubMed] [PMC]
27. Dutta P, M N. Facial Pain Expression Recognition in Real-Time Videos. J Healthc Eng. 2018;2018: 7961427. [DOI] [PubMed] [PMC]
28. Ghosh A, Umer S, Dhara BC, Ali GGMN. A Multimodal Pain Sentiment Analysis System Using Ensembled Deep Learning Approaches for IoT-Enabled Healthcare Framework. Sensors (Basel). 2025; 25:1223. [DOI] [PubMed] [PMC]
29. Guo Y, Wang L, Xiao Y, Lin Y. A Personalized Spatial-Temporal Cold Pain Intensity Estimation Model Based on Facial Expression. IEEE J Transl Eng Health Med. 2021;9:4901008. [DOI] [PubMed] [PMC]
30. Heintz TA, Badathala A, Wooten A, Cu CW, Wallace A, Pham B, et al. Preliminary Development and Validation of Automated Nociception Recognition Using Computer Vision in Perioperative Patients. Anesthesiology. 2025;142:726–37. [DOI] [PubMed]
31. Mao S, Li A, Luo Y, Gou S, Qi M, Li T, et al. Multi-Task Hybrid Conv-Transformer With Emotional Localized Ambiguity Exploration for Facial Pain Assessment. IEEE J Biomed Health Inform. 2025;29: 5133–45. [DOI] [PubMed]
32. Mieronkoski R, Syrjälä E, Jiang M, Rahmani A, Pahikkala T, Liljeberg P, et al. Developing a pain intensity prediction model using facial expression: A feasibility study with electromyography. PLoS One. 2020;15:e0235545. [DOI] [PubMed] [PMC]
33. Morsali M, Ghaffari A. Enhanced residual attention-based subject-specific network (ErAS-Net): facial expression-based pain classification with multiple attention mechanisms. Sci Rep. 2025;15:19425. [DOI] [PubMed] [PMC]
34. Park I, Park JH, Yoon J, Na H, Oh A, Ryu J, et al. Machine learning model of facial expression outperforms models using analgesia nociception index and vital signs to predict postoperative pain intensity: a pilot study. Korean J Anesthesiol. 2024;77:195–204. [DOI] [PubMed] [PMC]
35. Pikulkaew K, Boonchieng W, Boonchieng E, Chouvatut V. 2D Facial Expression and Movement of Motion for Pain Identification With Deep Learning Methods. IEEE Access. 2021;9:109903–14. [DOI]
36. Rezaei S, Moturu A, Zhao S, Prkachin KM, Hadjistavropoulos T, Taati B. Unobtrusive Pain Monitoring in Older Adults With Dementia Using Pairwise and Contrastive Training. IEEE J Biomed Health Inform. 2021;25:1450–62. [DOI] [PubMed]
37. Rodriguez P, Cucurull G, Gonzalez J, Gonfaus JM, Nasrollahi K, Moeslund TB, et al. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. IEEE Trans Cybern. 2022;52:3314–24. [DOI] [PubMed]
38. Semwal A, Londhe ND. A multi-stream spatio-temporal network based behavioural multiparametric pain assessment system. Biomed Signal Process Control. 2024;90:105820. [DOI]
39. Tan CW, Du T, Teo JC, Chan DXH, Kong WM, Sng BL. Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (STA-LSTM) network. Sci Rep. 2025;15:13429. [DOI] [PubMed] [PMC]
40. Yuan X, Cui Z, Xu D, Zhang S, Zhao C, Wu X, et al. Occluded Facial Pain Assessment in the ICU using Action Units Guided Network. IEEE J Biomed Health Inform. 2024;28:438–49. [DOI] [PubMed]
41. Zhang J, Hu X, Duan W, Ji M, Yang J. Application of deep learning-based facial pain recognition model for postoperative pain assessment. J Clin Anesth. 2025;105:111898. [DOI] [PubMed]

42. Mende-Siedlecki P, Qu-Lee J, Lin J, Drain A, Goharzad A. The Delaware Pain Database: a set of painful expressions and corresponding norming data. Pain Rep. 2020;5:e853. [DOI] [PubMed] [PMC]
43. Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In: Proceedings of 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG); 2011 Mar 21-25; Santa Barbara, CA, USA. IEEE; 2011. pp. 57–64. [DOI]