Journal of Pain Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Application of Intelligent Neonatal Pain Assessment Tools Based on Multimodal Data Fusion: A Systematic Review

Qiaojia Zhou[1], Xuanni Huang[1], Yuanhong Lv[1], Zhitian Xiao[1], Shuli Luo[1], Zhiyong Wang[2], Yuan Li[3,4], Yingxin Li[3,4], Qiong Chen[3,4], Zhangbin Yu[5], Queyun Zhou[1]

[1]Department of Neonatology, Shenzhen Children's Hospital, Shenzhen, People's Republic of China; [2]School of Biomedical Engineering, Harbin Institute of Technology, Shenzhen, People's Republic of China; [3]Department of Neonatal Nursing, West China Second University Hospital, Sichuan University, Chengdu, People's Republic of China; [4]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, People's Republic of China; [5]Department of Neonatology, Shenzhen People's Hospital, Shenzhen, People's Republic of China

Correspondence: Queyun Zhou, Department of Neonatology, Shenzhen Children's Hospital, No. 7019 Yitian Road, Futian District, Shenzhen, Guangdong, 518038, People's Republic of China, Email 18938690108@189.cn; Zhangbin Yu, Department of Neonatology, Shenzhen People's Hospital, No. 1017 Dongmen North Road, Luohu District, Shenzhen, Guangdong, 518020, People's Republic of China, Email zhangbinyu@njmu.edu.cn

Objective: Effective neonatal pain assessment is crucial for optimal analgesic management; however, it remains challenging, as traditional pain scales and single-modal intelligent assessment approaches continue to face substantial methodological and clinical limitations. Intelligent systems integrating multimodal data offer promising alternatives for enhancing objectivity and continuity of assessment. This systematic review aimed to identify, evaluate, and synthesize current intelligent neonatal pain assessment methods based on multimodal data fusion.

Methods: Two investigators independently searched PubMed, Embase, Cochrane Library, and Web of Science databases for relevant studies published up to September 12, 2025. Studies reporting the use of a multimodal approach to neonatal pain assessment were included. The methodological quality of the selected studies was independently assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2. A systematic review of the findings was then conducted.

Results: Nine studies met the inclusion criteria and were included in this review. Seven studies compared the accuracy of single-modal versus multimodal assessments of neonatal pain, with five also reporting the Area Under the Curve (AUC) values. All studies demonstrated that multimodal approaches achieved higher accuracy than single-modal methods. Specifically, 77.78% (7/9) of the studies successfully assessed pain intensity and distinguished between neonates with and without pain. However, only 44.44% (4/9) addressed methods for handling missing data, and merely 33.33% (3/9) utilized external validation sets.

Conclusion: Multimodal pain assessment demonstrates superior accuracy in neonatal pain evaluation. However, the heterogeneity in modalities, outcome indicators, and performance statistics across existing studies limits the ability to identify an optimal combination of modalities. Further research focusing on standardization, clinical applicability, and robust validation is required to strengthen the evidence base in this field and facilitate clinical translation.

Keywords: pain, multimodal, artificial intelligence, neonates, systematic review

Introduction

Pain, defined as an aversive sensory and emotional experience typically caused by, or resembling that caused by, actual or potential tissue injury, is a complex physiological and psychological phenomenon.[1] Evidence indicates that fetuses during the second trimester are already capable of perceiving pain. Neonates exhibit a lower pain threshold than adults and are more vulnerable to the effects of noxious stimuli.[2,3] Compared with adults, neonates exhibit more pronounced and sustained physiological responses to pain and are unable to articulate or describe their pain experience in the manner that adults can.[4] Behavioral expressions also differ: underdeveloped facial muscles and limited motor skills produce

https://doi.org/10.2147/JPR.S578721
Received: 4 November 2025
Accepted: 11 March 2026
Published: 27 March 2026

movements distinct from those of adults, and pain behaviors often overlap with non-pain stress responses. These differences limit the direct application of adult pain assessment methods to neonates.

Infants in neonatal intensive care units (NICUs) undergo an average of 14 painful procedures daily.[5] Common clinical interventions cause severe pain in a significant proportion (70.37%) of hospitalized neonates.[6] Beyond procedural pain, neonates in NICUs also experience acute postoperative pain and chronic pain from interventions such as endotracheal intubation and indwelling drains.[7] Persistent pain stress can disrupt physiological homeostasis, and the cumulative effects of repeated pain exposure can lead to adverse short-term and long-term outcomes; these include apnea, feeding difficulties, behavioral alterations, heightened pain sensitivity, and impairments in cognitive, motor, and neurological development,[3,8] all of which can significantly impact neonatal survival and quality of life. Given that neonates cannot verbally articulate their pain, its timely and accurate assessment is paramount for effective pain management.[9,10]

Current neonatal pain management largely relies on observational pain assessment scales administered by nurses. Numerous scales have been developed and validated for specific neonatal populations and clinical contexts.[7] This necessitates that neonatal nurses be proficient in multiple scales and select appropriate tools based on gestational age and pain type (eg., acute, prolonged).[11–17] However, this approach is often time-consuming, subjective, lacks continuity, and can exhibit significant inter-observer variability.[18] Such inconsistencies and intermittent assessments may lead to under-recognition of pain and difficulties in titrating sedative and analgesic medications, thereby constraining effective neonatal pain management.

In the era of smart healthcare, Artificial Intelligence (AI) has found widespread application in clinical medicine. Leveraging advancements in pain research and AI technology, researchers have developed automated pain assessment techniques for neonates. These systems aim to provide consistent, objective, and dynamic pain evaluation by analyzing behavioral cues (eg., facial expressions,[19–25] crying,[26,27] limb activity[28]) and neurophysiological responses (eg., heart rate variability,[29,30] cerebral hemodynamic changes[31]) to painful stimuli. These signals are automatically extracted and analyzed using computational methods. However, many early AI-based studies focused on a single modality. Pain expression is inherently diverse, complex, and susceptible to confounding factors such as gestational age, environmental noise, and sedation. Single-modality assessments may therefore struggle to accurately capture the entirety of the neonatal pain experience and are prone to bias if data from that one modality is lost or distorted due to clinical interferences.[32]

Recognizing these limitations, researchers have begun to develop and investigate multimodal approaches to neonatal pain assessment, creating datasets that integrate various pain indicators.[32–35] By applying AI techniques and computer algorithms to preprocess and analyze collected video, audio, and physiological information, these multimodal systems aim for more comprehensive data integration through feature extraction and classification. Preliminary findings suggest that multimodal assessment can achieve higher accuracy than single-modality automated methods.[34,36–40] However, there has been no systematic synthesis qualitatively summarizing or quantitatively comparing the performance of these emerging multimodal approaches, nor has there been an evaluation of the strengths and weaknesses of different modality combinations. Furthermore, the computational algorithms underpinning these systems are often published in computer science venues rather than medical journals, which can limit their clinical translation and interpretability by healthcare professionals.[41]

This systematic review aims to address this gap by comprehensively identifying, evaluating, and synthesizing studies on intelligent neonatal pain assessment methods based on multimodal data fusion. The objective is to analyze the current landscape, performance, and potential applications of these multimodal predictive methods in neonatal care.

Materials and Methods

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[42] As this study exclusively involved a review of previously published literature, ethical approval was not required. The systematic review protocol was registered with PROSPERO, registration number CRD42025619936. Appendix 1 of the Supplementary Materials contains the PRISMA 2020 Checklist.

Search Strategy and Study Identification

Two investigators (Q.Z., X.H.) with expertise in evidence-based medicine independently conducted a comprehensive literature search across four electronic databases: PubMed, Embase, Cochrane Library, and Web of Science. The search covered the period from database inception to September 12, 2025. To ensure comprehensive retrieval and minimize publication bias, forward and backward citation tracking was performed by examining the reference lists of included articles and relevant reviews. The search strategy incorporated keywords and subject headings related to three core concepts: neonates, pain, and multimodal intelligent assessment. Each concept included multiple synonyms to maximize search sensitivity. Detailed search strategies for each database are provided in Appendix 2 of the Supplementary Materials. All retrieved records were managed using EndNote X9 (Clarivate Analytics).

Eligibility Criteria

Eligibility criteria were established using the PICO/PECO (Population, Intervention/Exposure, Comparison, Outcome) framework.[43]

Original research articles were included if they reported on automated or intelligent neonatal pain assessment methods that used data from two or more modalities, such as facial expressions, body movements, crying, physiological signals (eg., heart rate, oxygen saturation), electroencephalography (EEG) or near-infrared spectroscopy (NIRS). The study population had to consist of neonates, defined as infants from birth to 28 days postpartum, including preterm (gestational age < 37 weeks), full-term (gestational age 37^{+0} to 41^{+6} weeks), post-term (gestational age ≥ 42 weeks) infants, and corrected premature infants with a gestational age of less than 44 weeks. Furthermore, studies were required to assess acute procedural pain, such as that from heel pricks or vaccinations, or postoperative pain. Studies were excluded if they focused on pain assessment using only a single modality, or if algorithm accuracy or a comparable performance metric (as the primary outcome) was not reported. Additionally, abstracts, reviews, meta-analyses, study protocols, letters to the editor, animal studies, and studies not published in English were excluded from this review.

Study Selection

Two investigators (Q.Z., X.H.) independently screened the titles and abstracts of all retrieved records to identify potentially eligible studies. Subsequently, the full texts of these potentially relevant articles were retrieved and assessed for final inclusion by the same two investigators. Given the interdisciplinary nature of the topic, an expert in computer science (Z.W.) was consulted for technical input during the full-text screening phase but did not act as an independent reviewer. Reasons for excluding studies at the full-text stage were documented in Appendix 3 of the Supplementary Materials. Any disagreements between the two primary investigators at either the initial screening or full-text review stage were resolved through discussion and consensus. If consensus could not be reached, a third investigator (Q.Y.Z.) arbitrated the decision.

Data Extraction

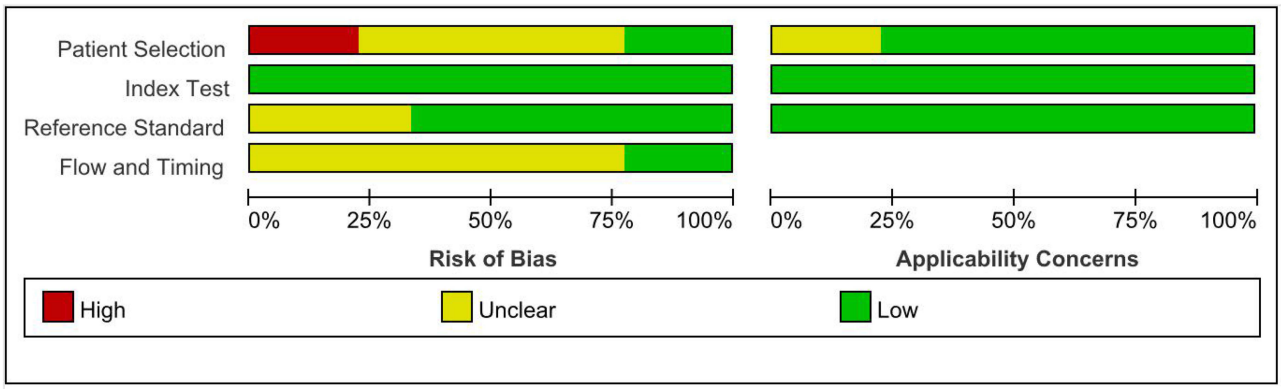
Two investigators (Q.Z., X.H.) independently extracted data manually from the included studies using a standardized data extraction form. The extracted information included: (1) Study Information: first author, year of publication, country of origin, author disciplinary background; (2) Population Characteristics: sample size, gestational age, gender, ethnicity (if reported), type of pain stimulus; (3) Dataset Information (if applicable): database name, video/signal capture details (eg., timing, duration), methods for video/signal labeling (eg., scales used, number of raters), inter-rater reliability measures; (4) Multimodal Assessment Method: modalities included, feature extraction techniques, classification algorithms, methods for data fusion (eg., feature-level, decision-level), handling of missing data, and use of validation sets (internal or external); (5) Outcome Measures: such as accuracy, AUC, or sensitivity, specificity, and any reported comparisons between single-modal and multimodal approaches or between different multimodal combinations.

Discrepancies in extracted data were resolved through discussion and consensus between the two investigators. The computer science expert (Z.W.) was available for consultation regarding technical details of algorithms or computational methods during data extraction. A third investigator (Q.Y.Z.) acted as an arbitrator if consensus could not be reached.

Assessment of Methodological Quality

Two investigators (Q.Z., X.H.) independently evaluated the full-text articles using the established QUADAS-2 tool.[44] The assessment focused on four key domains, including patient selection, index test, reference standard, and flow and timing, to evaluate the risk of bias and applicability of the included studies. Any discrepancies arising during the quality assessment were resolved through discussion. If consensus could not be reached, the third investigator (Q.Y.Z.) arbitrated the decision. The results of this evaluation are presented in Figure 1.

A



B

Inclusion studies	Risk of Bias				Applicability Concerns		
Study	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Zamzmi G, et.al (2016) [36]	High	Low	Unclear	Unclear	Unclear	Low	Low
Zamzmi G, et.al (2017) [37]	High	Low	Unclear	Unclear	Unclear	Low	Low
Egede J, et.al (2019) [32]	Unclear	Low	Low	Unclear	Low	Low	Low
van der Vaart M,et.al (2019) [38]	Low	Low	Unclear	Low	Low	Low	Low
Salekin MS, et.al (2019) [45]	Unclear	Low	Low	Unclear	Low	Low	Low
Salekin MS, et.al (2021) [9]	Unclear	Low	Low	Unclear	Low	Low	Low
Zamzmi G, et.al (2022) [34]	Unclear	Low	Low	Unclear	Low	Low	Low
Salekin MS, et.al (2022) [39]	Unclear	Low	Low	Unclear	Low	Low	Low
Zhu H, et.al (2024) [40]	Low	Low	Low	Low	Low	Low	Low

=Low risk =High risk =Unclear risk

Figure 1 Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2. (**A**) Overall Quality Assessment Results Graph; (**B**) Quality Assessment Results.

https://doi.org/10.2147/JPR.S578721

Results

Study Selection

The systematic search across PubMed, Embase, Cochrane Library, and Web of Science databases initially identified 1,122 records. After removing 375 duplicates, the titles and abstracts of the remaining 747 articles were screened. Based on the predefined eligibility criteria, 700 articles were excluded at this stage. The full texts of the remaining 47 articles were then thoroughly assessed for eligibility. Of these, 41 articles were excluded for various reasons (Appendix 3 of the Supplementary Materials). This process resulted in six original studies meeting the inclusion criteria. Additionally, 26 articles in the field of computer science were identified through citation retrieval. Following screening of the titles and abstracts, 17 articles were excluded at this stage. An additional five articles were excluded after full-text review (Appendix 3 of the Supplementary Materials). This process resulted in three original studies meeting the inclusion criteria. Ultimately, a total of nine studies were included in this systematic review (Figure 2).

Characteristics of Included Studies

The nine included studies[9,32,34,36–40,45] were all published within the last decade, between 2016 and 2024. Geographically, six studies originated from the United States,[9,34,36,37,39,45] two from the United Kingdom,[32,38] and one from China.[40] The first authors of eight studies were computer scientists,[9,32,34,36,37,39,40,45] while only one study had a physician as the first author.[38] The total number of neonates across all studies was 1,453, with individual study sample sizes ranging from 18 to 1,091. Four studies provided detailed information on the race and ethnicity of their participants.[9,34,37,45] Notably, one study by Egede et al[32] utilized data (APN-DB) collected in Nigeria, although the study itself was UK-based; the authors also noted that the "primitive pain face" is considered consistent across ethnicities.[46] All included studies involved neonates with a gestational age of up to 42 weeks, comprising both preterm and full-term infants, with two studies including extremely preterm infants (<28 weeks' gestation).[32,40] Five studies reported the sex distribution of the participants.[9,32,34,38,40] The pain stimuli investigated were predominantly acute

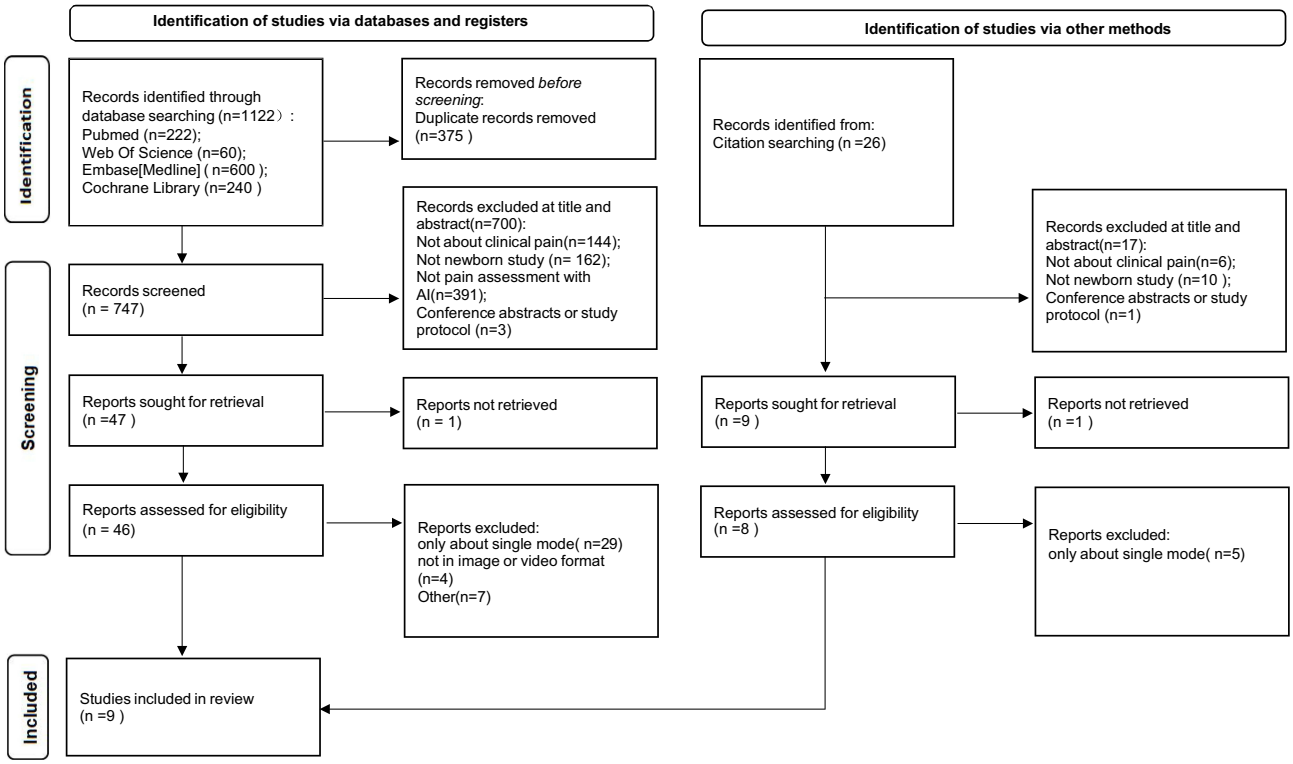


Figure 2 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

https://doi.org/10.2147/JPR.S578721

procedural pain events, such as heel pricks, with three studies also including postoperative pain.[9,34,39] Detailed characteristics of the included studies are presented in Table 1.

Quality Evaluation

Figure 1A and B present the overall risk of bias graph and summary for the included studies. The methodological quality of the nine studies included in this systematic review was assessed using the QUADAS-2 tool.[9,32,34,36–40,45] Among them, two studies were rated as high quality,[38,40] five as moderate,[9,32,34,39,45] and two as low quality.[36,37] In the patient selection domain, 77.78% of studies did not report whether participants were consecutively or randomly enrolled, indicating potential selection bias. Two studies with small sample sizes (n=18) were rated as high risk in this domain,[36,37] but were retained after consultation with the computer science expert (Z.W.) due to their pioneering work on multimodal intelligent pain assessment algorithms for neonates. In the index test domain, all studies provided sufficient details of their assessment methods and predetermined thresholds. In the reference standard domain, 33.33% of the studies did not conduct inter-rater reliability testing, suggesting possible bias in the application of the reference standard. In the flow and timing domain, 77.78% of studies did not clearly report whether all participants were included in the analysis.

Types of Modalities Included

The included studies varied in the number and combination of modalities used for pain assessment. Eight studies developed automated multimodal assessment methods incorporating between two and five modalities.[9,32,34,36–38,40,45] Details of the specific modalities are provided in the third column of Table 2. These studies employed diverse preprocessing, feature extraction, and classification techniques. One study focused on validating the performance of multimodal pain assessment using different computational methods and algorithms on established multimodal datasets.[39]

The pioneering work by Zamzmi et al[36] in 2016 introduced an automated multimodal analysis combining facial expressions, body movements, and vital signs. This group later expanded their approach to include cry features and developed the NPAD dataset.[34,37] In 2019, Salekin et al[45] collected video recordings capturing facial expressions and limb movements from 31 neonates undergoing heel lance or immunization procedures. This research represented the first attempt to investigate neonatal pain using a multi-channel Convolutional Neural Network (CNN). In 2021, Salekin et al[33] introduced the first publicly available multimodal neonatal pain dataset, USF-MNPADi, which encompassed behavioral responses (facial expressions, body movements, crying), and physiological signals (vital signs) during procedural and postoperative pain states. Two subsequent studies[9,39] by Salekin et al utilized a subset of the USF-MNPADi dataset to develop a multimodal spatio-temporal deep learning approach and a Transformer-based attention model for neonatal postoperative pain assessment, focusing on facial expression, body movement, and crying. Egede et al[32] constructed the APN-db dataset using facial expression and limb movement. Notably, this was the first work to provide pain intensity annotations on a neonate dataset based on clinical neonatal pain assessment parameters. Zhu et al[40] incorporated facial expression and limb movement, with a specific focus on evaluating pain in scenarios with partially obscured or distorted modal information. The study by van der Vaart, M. et al[38] included a comprehensive set of five modalities: facial expressions, heart rate, oxygen saturation, EEG, and bilateral EMG.

Video Capture and Labeling

Seven studies specified the time points for video recording in relation to the pain procedures and documented the pain-free baseline state prior to the initiation of the painful procedure.[9,32,34,36–39] One study reported that each video was one minute in length and captured the entire painful procedure.[40] Another study, while dividing data recording into eight time periods—including the baseline period (T0) and the procedural preparation period (T1)—did not specify the exact time points or the duration of video acquisition.[45] Neither of these two studies provided an explicit definition of no pain. Postoperative pain was recorded for the longest duration, with video capture extending from 15 minutes before surgery to 3 hours post-surgery.[9,34,39] Video labeling in all studies was conducted by trained nurses or researchers. The USF-MNPADi and NPAD datasets employed blind annotation methods to ensure rater independence.[9,34,39] Five studies assessed and reported inter-rater reliability for video labeling.[9,32,34,39,45]

Table 1 Basic Characteristics of Included Studies

Reference	Region	Disciplinary Background of First Author	Race and Ethnicity	Sample Size of Population (n)	GA	Gender	Crowd Classification	Circumstances	Several Modality Indicators
Zamzmi G et.al (2016)[36]	USA	Computer scientist	NA	18	32–41W (36W)	NA	NA	Heel lancing	3
Zamzmi G et.al (2017)[37]	USA	Computer scientist	White,Caucasian, Hispanic, African American, Asian, and others.	18	28–41W (36W)	NA	NA	Heel lancing, immunization	4
Egede J et.al (2019)[32]	UK	Computer scientist	NA	101	<28W; 28–31W; 32–36W; ≥37W	Male: 59; Female: 42	NA	Heel lancing, intravenous injection, intramuscular injection, lumbar puncture	2
van der Vaart M et.al (2019)[38]	UK	Physician	NA	109	34–42W	Male: 56; Female: 53	Preterm:21 full-term:40	Heel lancing	5
Salekin MS et.al (2019)[45]	USA	Computer scientist	Asian, African American, White, Caucasian.	31	32–40W (35W)	NA	NA	Heel lancing, immunization	2
Salekin MS et.al (2021)[9]	USA	Computer scientist	Asian, African, American, Caucasian	Total=45; n=36 (procedural); n= 9 (postoperative)	30–41W	Male: 19; Female: 17 (procedural) Male: 5; Female: 4 (postoperative)	NA	Procedural pain, postoperative pain	3
Zamzmi G et.al (2022)[34]	USA	Computer scientist	White, Caucasian, African American, Asian	Total=40; n=31 (procedural); n=9 (postoperative)	32–40W (35.9W)	Male: 20; Female: 20	Preterm:18 full-term:22	Heel lancing, immunization, postoperative pain	4
Salekin MS et.al (2022)[39]	USA	Computer scientist	NA	USF-MNPAD-I*	NA	NA	NA	Procedural pain, postoperative pain	3
Zhu Het.al (2024)[40]	China	Computer scientist	NA	1091	26–42W (36W)	Male: 656; Female:435	NA	Clinical procedures (17 types): arterial blood collection; fingertip blood et. al	2

Note: NA, no report.

Abbreviations: GA, gestational age; USF-MNPAD-I, University of South Florida Multimodal Neonatal Pain Assessment Dataset.

https://doi.org/10.2147/JPR.S578721

Table 2 Description of the Database in the Selected Study

Reference	Dateset Name	Type of Included Mode	Video and Audio Collection Duration	Tagged Video Scale			Quality
				The Person Responsible for Marking	Measurement Consistency Method	Scale	
Zamzmi G et.al (2016)[36]	NA	Facial expression; body motion; vital signs	Begins 5 minutes before the painful procedure and ends 5 minutes after its completion	Trained nurses	NA	NIPS	Lower
Zamzmi G et.al (2017)[37]	NA	Facial expression; body motion; crying sound; vital signs	Begins 5 minutes before the painful procedure and ends 5 minutes after its completion	Trained nurses	NA	NIPS	Lower
Egede J et. al (2019)[32]	APN-db (NFLAPS)	Facial expression; limb movement	Started 2 minutes before the pain stimulus and continued at least 1 minute	Two nurses who had over 25 years of NICU experience	Cohen's Kappa Coefficien (0.67)	NIPS, NFCS	Moderate
van der Vaart M. et. al (2019)[38]	NA	Facial expressions; heart rate; Oxygen saturation; EEG; Ipsilateral EMG; Contralateral EMG	15 seconds before and 30 seconds after the heel lance	Two researchers trained	NA	PIPP, PIPP-R	Higher
Salekin MS et.al (2019)[45]	NA	Facial expression; body movement	NA	Two caregivers	Kappa coefficient (0.85) Pearson correlation (0.9)	NIPS	Moderate
Salekin MS et.al (2021)[9]	USF-MNPAD-I	Facial expression; body movement; crying sound	a.baseline, during a procedural pain stimulus and immediately after the completion of the stimulus; b.baseline state and monitored for three hours after the surgery	Trained nurses independently perform manual marking.	Kappa coefficient (0.85); Pearson correlation (0.89)	NIPS, N-PASS	Moderate
Zamzmi G et.al (2022)[34]	NPAD	Facial expression, body motion, crying sound, vital signs	a.3–6min; b.15 minutes before surgery to 3 hours after	Two trained nurses independently perform manual marking.	Kappa coefficient (0.85)	NIPS, N-PASS	Moderate
Salekin MS et.al (2022)[39]	NA	Facial expression; body movement; crying sound	a.baseline state and during a procedural pain stimulus and immediately after the completion of the stimulus; b.baseline state and monitored for three hours after the surgery	Trained nurses independently perform manual marking.	Kappa coefficient (0.85) and Pearson correlation (0.89)	NIPS, N-PASS	Moderate
Zhu H et.al (2024)[40]	NA	Facial expression; body movement	1min	Two nurses individually assessed	When the assessment results are inconsistent	NIPS	Higher

Notes: a, procedural pain; b, postoperative pain.

Abbreviations: NIPS, the Neonatal Infant Pain Scale; PIPP, the Premature Infant Pain Profile; PIPP-R, the premature infant pain profile-revised; NFCS, the Neonatal Facial Coding System; N-PASS, Neonatal Pain, Agitation and Sedation Scale; EEG,electroencephalogram; EMG, electromyography.

Pain Assessment Scales Used for Ground Truth Labeling

Several established neonatal pain scales were used to provide ground truth labels for the video data. The Neonatal Infant Pain Scale (NIPS)[12] was utilized in four studies for labeling acute procedural pain.[36,37,40,45] One study employed the Premature Infant Pain Profile (PIPP)[11] and its revised version, PIPP-R,[47] for acute pain stimuli.[38] Three studies used both NIPS and the Neonatal Pain, Agitation, and Sedation Scale (N-PASS)[13] to label procedural and postoperative pain.[9,34,39] Egede et al[32] reported using the Neonatal Facial Coding System (NFCS)[48] for facial expression changes and NIPS for body movement markers. Combining these two scales generated an 11-point Neonatal Face and Limb Acute Pain Scale (NFLAPs). These scales are recognized for their validity and reliability in assessing neonatal pain.[49] Details regarding dataset names, included modality types, and pain annotation information are provided in Table 2.

Identification of Pain Intensity Levels

An important capability of automated pain assessment systems lies not only in detecting pain but also in quantifying its intensity, enabling clinicians to implement appropriate nonpharmacological or pharmacological interventions based on the determination of mild, moderate, or severe pain levels. This quantification process relies on the rigorous alignment of multimodal data with clinical "gold standards". Specifically, expert clinicians perform granular annotation by synchronizing behavioral and physiological indicators with validated pain scales, ensuring that the machine learning models learn to map complex feature patterns—such as specific facial clusters or heart rate variability changes—to precise intensity scores. This meticulous labeling approach transforms subjective clinical observations into objective, trainable data, which is essential for the transition from simple binary detection to nuanced intensity grading. Five of the included studies successfully distinguished between different levels of pain intensity, such as moderate versus severe pain.[9,32,34,36,39] Zamzmi et al (2017),[37] due to a limited number of moderate pain instances in their dataset, focused on a binary classification (no pain vs. severe pain). Zhu et al[40] advanced this by classifying pain into three categories: mild, moderate, and severe.

Handling of Missing Modalities

The clinical environment can make complete data collection for all modalities challenging. Four studies did not explicitly mention how missing modality data were handled.[9,32,36,45] One study reported excluding instances with missing modalities.[37] Four studies described specific methods to address missing data.,[34,38–40] Salekin et al (2022)[39] proposed an attentional model to reconstruct missing modalities. Zhu et al[40] developed a dataset with interference annotations (eg., facial occlusion, motion interference) to train their model for robustness under uncontrolled conditions. Van der Vaart, M. et al[38] selected a machine learning algorithm (Random Forest) inherently capable of handling missing data, while Zamzmi et al (2022)[34] trained separate classifiers for each scenario of missing modalities.

Validation and Subgroup Analysis of Datasets

The use of independent validation sets and subgroup analyses (eg., by gestational age or pain type) enhances the generalizability and robustness of findings. Three studies reported validation of their datasets using independent data.[9,32,38] Salekin et al (2021)[9] used a random 25% of their data for validation. Egede et al[32] validated their dataset with 17 videos from 13 participants. Van der Vaart, M. et al[38] used a validation set of 32 samples to verify their multimodal dataset and also performed a subgroup analysis comparing preterm and term infants. While their multimodal assessment achieved higher accuracy (88%) and AUC (0.94) in the preterm group for discriminating between injurious and non-injurious stimuli, the difference between the two gestational age groups was not statistically significant (P=0.22).

Performance of Computerized Multimodal Pain Assessment

The included studies exhibited considerable heterogeneity in terms of participant characteristics, included modalities, data preprocessing, feature extraction methods, classification algorithms, and reporting styles, precluding a formal meta-analysis. Key performance indicators, primarily accuracy and AUC, are summarized in Tables 3 and 4. The selection of

Table 3 Method and Performance of Modal Feature Extraction and Classification

Reference	Extracted Features	Classification	Performance
Zamzmi G, et.al (2016)[36]	Facial strain magnitude, body motion amount and vital signs (HR, RR, SpO2)	KNN, SVM, Random Forest, leave-one-subject-out cross validation	Accuracy: 95%
Zamzmi G et.al (2017)[37]	Strain magnitude, LPCC and MFCC, Total motion, and vital signs (HR, RR, SpO2)	SVM, LS-SVM, Random Forest, KNN, 10-fold cross-validation, Leave one-subject-out cross-validation	Accuracy: 96.7%
Egede J et.al (2019)[32]	HOG, shape features, data-learned features	RVM, CNN, leave-one-subject cross-validation	RMSE: 1.89, MAE: 1.71, PCC: 0.46
van der Vaart M et.al(2019)[38]	Behavioral features and vital signs (EEG, EMG, HR)	Random forests	Accuracy: 81%, AUC: 0.90
Salekin MS et.al (2019)[45]	Facial Feature, Body Feature	Multi-channel CNN,VGG-16, LSTM	Accuracy: 92.48%, AUC: 0.90
Salekin MS et.al (2021)[9]	Facial expression, Body movement, MFCC	VGG-16, Bilinear CNN, LSTM	Accuracy: 78.95%, AUC: 0.88
Zamzmi G et.al (2022)[34]	Facial expression, body motion, MFCC, LPCC and vital signs (HR, RR, SpO_2)	SVM, Multimodal fusion, 10-fold Cross-Validation, Leave-One-Subject-Out Cross-Validation	Decision: Accuracy: 95.56%, AUC: 0.87; Feature: Accuracy: 92.28%, AUC:0.91;
Salekin MS et.al (2022)[39]	Spatial and temporal features	CNN, LSTM, Transformer-based Attentional Model, leave-one-subject-out	Accuracy: 82.02%, AUC: 0.91
Zhu H et.al (2024)[40]	Facial regional features, Body skeleton sequences	Region-Channel-Attention, Bi-GRU, Keyframes-aware convolution	Accuracy: 91.04%

Abbreviations: HR, heart rate; RR, respiratory rate; SpO2, oxygen saturation; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; LS-SVM, Least Squares Support Vector Machine; LPCC, Linear Predictive Cepstral Coefficients; MFCC, Mel-Frequency Cepstral Coefficients; HOG, Histogram of Oriented Gradients; LBP, Local Binary Patterns; VGG-16, Visual Geometry Group 16-layer network; RVM, Relevance Vector Machine; CNN, Convolutiona Neural Network; EEG, Electroencephalogram; EMG, Electromyogram; LSTM, Long Short-Term Memory; Bi-GRU, Bidirectional Gated Recurrent Unit; AUC, Area Under the Curve.

these metrics reflects the current emphasis in the field on the discriminative power of algorithms. However, it is important to note that accuracy and AUC alone do not fully capture a model's clinical utility. Advanced metrics such as calibration, which evaluates the agreement between predicted probabilities and actual pain occurrences, and uncertainty quantification, which assesses the model's confidence in its predictions, were rarely reported in the included studies.

Seven studies directly compared the performance of single-modal versus multimodal pain assessments,[9,34,36–38,40,45] with several also reporting AUC values.[9,34,38,39,45] Across all these studies, multimodal approaches consistently achieved higher accuracy in neonatal pain assessment compared to any single-modality used in isolation. Facial expression was a common modality across all studies and often demonstrated the highest accuracy among individual modalities in five studies.[34,36,37,40,45] For instance, Zamzmi et al (2016)[36] reported an accuracy of 88% for facial expression alone, while their multimodal system achieved 95%. In contrast, limb movement, when assessed as a single modality, showed lower accuracies in studies by Salekin et al (2021)[9] (66.67%) and Zhu et al[40] (68.89%). Cry as a single indicator achieved the highest unimodal accuracy (76.61%) and AUC (0.82) in the study by Salekin et al (2021),[9] potentially due to factors such as obscured facial views or reduced limb movement in some neonates. Egede et al[32] reported the Root Mean Square Error (RMSE) of 1.89, the Mean Absolute Error (MAE) of 1.71, and the Pearson Correlation Coefficient (PCC) of 0.46 as the core outcome measures.

Zamzmi et al[34,37] explored different data fusion strategies. Their findings indicated that decision-level fusion (accuracy 95.56%) outperformed feature-level fusion (accuracy 92.28%), although feature-level fusion yielded a higher AUC and a lower false negative rate. Two studies by Salekin et al[9,39] using the USF-MNPAD-I dataset (facial

Table 4 A Comparative Analysis of Single-Modal and Multimodal Pain Assessment Across Various Studies

Reference	Facial Expression		Body Move		Crying sound		Vital Signs		EEG		Multimodal	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Zamzmi G et.al (2016)[36]	88%	NA	86%	NA	NA	NA	82%	NA	NA	NA	95%	NA
Zamzmi G et.al (2017)[37]	93.2%	NA	87.5%	NA	87.8%	NA	73.6%	NA	NA	NA	97%	NA
van der Vaart M et.al (2019)[38]	70% (0.58–0.80)	0.76 (0.64–0.85)	NA	NA	NA	NA	73% (0.62–0.83)	0.77 (0.63–0.87)	64% (0.52–0.75)	0.75 (0.60–0.85)	81% (0.70–0.89)	0.9 (0.78–0.95)
Salekin MS et.al (2019)[45]	GS: 88.32%	GS:0.82	84.64%	0.77	NA	NA	NA	NA	NA	NA	92.48%	0.90
	LS: 88.87%	LS:0.89										
Salekin MS et.al (2021)[9]	70.76%	0.81	66.67%	0.78	76.61%	0.82	NA	NA	NA	NA	78.95%	0.88
Zamzmi G et.al (2022)[34]	88.87%	0.89	84.64%	0.77	83.35%	0.69	81.73%	0.72	NA	NA	92.28%F	0.91F
											95.56%D	0.87D
Salekin MS et.al (2022)[39]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	82.02%	0.91
Zhu H et.al (2024)[40]	84.44%	NA	68.89%	NA	NA	NA	NA	NA	NA	NA	91.04%	NA

Notes: D:Decision Fusion; F: Feature Fusion; the values in parentheses represent 95% credible intervals.

Abbreviations: GS, Geometric+SVM; LS, LBP-TOP+SVM; LBP-TOP, Local Binary Patterns from Three Orthogonal Planes; SVM, Support Vector Machine; AUC, Area Under the Curve; EEG, Electroencephalogram.

https://doi.org/10.2147/JPR.S578721

expression, limb movement, cry) showed that a spatio-temporal deep learning approach achieved an accuracy of 78.95% and AUC of 0.88, while a subsequent attentional cross-modal feature fusion approach achieved higher accuracy (82.02%) and AUC (0.91). Van der Vaart, M. et al[38] using five modalities to discriminate between a clinically required heel lance and a control non-noxious stimulus, reported a multimodal accuracy of 81% (95% CI: 0.70–0.89) and AUC of 0.90 (95% CI: 0.78–0.95), superior to any single-modality approach.

Discussion

This systematic review synthesized evidence from nine studies on intelligent neonatal pain assessment utilizing multimodal data fusion. The primary finding is that multimodal approaches consistently outperform single-modality methods in terms of assessment accuracy. These results highlight the potential of integrating diverse pain indicators through advanced computational techniques to enable more robust and reliable neonatal pain evaluation, thereby overcoming the limitations associated with traditional observational scales and single-modality AI systems.

Superior Performance of Multimodal Approaches

All nine included studies demonstrated that multimodal pain assessment achieved higher accuracy than single-modality approaches. Among the seven studies that directly compared performance,[9,34,36–38,40,45] facial expression consistently emerged as the highest-performing individual modality in five studies,[34,36,37,40,45] with accuracies ranging from 84.44% to 93.20%. However, multimodal integration substantially improved performance. For instance, Zamzmi et al (2016)[36] reported that facial expression alone achieved 88% accuracy, while their multimodal system (facial expression, body movement, vital signs) reached 95%. These findings confirm that different modalities provide complementary information, and their integration captures the multifaceted nature of neonatal pain more comprehensively than any single indicator.

Optimal Modality Combinations

Despite the promising performance of multimodal pain assessment, substantial heterogeneity exists across the included studies regarding the specific modalities combined, the computational algorithms employed for feature extraction and fusion, and the metrics used for reporting outcomes. This variability, while reflective of an emerging field, makes direct comparisons between different multimodal systems challenging and precludes a quantitative meta-analysis. Consequently, it remains difficult to definitively conclude which specific combination of modalities or fusion strategy yields universally optimal performance. Zamzmi et al[34] compared fusion strategies, finding that decision-level fusion achieved slightly higher accuracy (95.56%) than feature-level fusion (92.28%), though feature-level fusion yielded superior AUC and lower false negative rates. Although research[50,51] indicates that integrating multimodal approaches based on cortical, behavioral, and physiological response assessments significantly enhances the accuracy of pain evaluation, the incremental benefits and optimal weighting of modalities such as crying, body movements, vital signs (heart rate, blood pressure, oxygen saturation, respiratory rate), and neurophysiological indicators (eg., EEG, near-infrared spectroscopy) still require further systematic investigation. Based on current evidence, and practical considerations, a framework integrating facial expressions, body movements, crying sounds, and vital signs appears to be the most clinically feasible approach for neonatal pain assessment across diverse contexts, including acute, chronic, and post-operative pain. Although EEG-based systems have demonstrated scientific validity and reliability,[52] their broader clinical implementation must balance time efficiency, economic cost, and practicality, particularly in acute procedural pain settings.

Limitations of Single-Modality AI–Based Pain Assessment in Neonates

As outlined in the Introduction, the limitations of manual pain assessment scales in clinical practice have already been presented; therefore, they are not reiterated here. Nevertheless, it remains necessary to articulate the inherent limitations of single-modality AI-based pain assessment methods in order to elucidate the rationale for prioritizing multimodal pain assessment in neonates. The majority of single-modality studies incorporated facial expression, which aligns with its established role as a primary indicator of pain. Facial expressions encompass various characteristics, including brow

protrusion, eye constriction, nasolabial groove, and mouth opening.[53] It is worth noting that changes in facial appearance during non-painful situations may resemble those associated with pain, thereby diminishing the reliability of these features as specific indicators of pain. In clinical practice, single-modality assessment based solely on facial expressions may be limited due to reduced visibility of the neonatal face, which can result from interventions such as endotracheal intubation, nasal cannula placement during non-invasive ventilation, or the use of eye masks in phototherapy. Crying is a manifestation of pain in neonates; however, it may also be elicited by various non-pain-related stimuli, including hunger, physical discomfort, or unmet emotional needs. Body movements, while indicative of discomfort or pain, can also occur spontaneously or in response to non-noxious stimuli. They may be restricted or absent in sedated, extremely preterm, or critically ill neonates, thereby limiting their discriminative value as a standalone pain indicator. Variations in heart rate, blood pressure, oxygen saturation, and respiratory patterns are among the most commonly utilized physiological indicators of pain. While research has established a significant association between changes in vital signs and pain,[30,54] these alterations may also reflect the child's underlying medical condition, medication effects, or emotional states such as hunger and fear,[55] thereby limiting their specificity as pain indicators. Roué JM et al[56] employed electroencephalography (EEG) monitoring combined with machine learning algorithms to investigate the correlation between pain and cerebral electrical activity, aiming to distinguish between painful and non-painful events in neonates undergoing acute pain procedures. However, EEG signals are susceptible to artifacts from movement, electrical interference, and environmental noise, and require specialized equipment and technical expertise for data acquisition and interpretation. Furthermore, the relationship between EEG patterns and pain intensity remains incompletely characterized, particularly in preterm infants with immature neural development. This limitation may restrict the generalizability and real-time applicability of EEG-based assessments in routine clinical practice. These collective limitations underscore that no single indicator is sufficient, necessitating the integration of diverse data streams for robust neonatal pain evaluation.

Data Quality and Representativeness

A critical aspect for the clinical translation of these intelligent systems is the quality and representativeness of the datasets used for their development and validation. While three studies in this review constructed multimodal neonatal pain datasets, demographic information on study participants is often limited. This particularly applies to gestational age at birth, corrected gestational age for preterm infants, age, weight, and relevant clinical conditions apart from painful stimuli. These details are of critical importance, as such factors exert substantial influences on both pain expression and physiological responses. Furthermore, the included study populations encompass a broad spectrum of gestational ages, ranging from 26 to 42 weeks. It is noteworthy that preterm infants, particularly those of lower gestational ages, exhibit immature central nervous system development and diminished regulatory capacities. This is especially true for extremely preterm infants, who demonstrate limited behavioral capabilities in expressing pain compared to their full-term counterparts.

Moreover, a majority of the datasets (over 66.67% based on the included studies that developed new systems or datasets and reported on validation) lacked external validation using independent cohorts from different clinical settings or populations. Only 33.33% of studies (3/9)[9,32,38] employed external validation sets. This is a significant gap, as models validated only on internal data or subsets of the same dataset may not generalize well to real-world clinical practice. The development of large-scale, diverse, and publicly accessible benchmark datasets, ideally with standardized annotation protocols, would greatly accelerate progress and facilitate more rigorous comparative evaluations in this field.

Ground Truth Standardization Issue

Standardization in outcome reporting and ground truth labeling is another key area for improvement. Four different pain scales (NIPS, PIPP/PIPP-R, N-PASS, NFCS) were used for video labeling across the reviewed studies. While all are validated scales, their differing constructs and scoring can introduce variability. The N-PASS, recommended by some guidelines for its comprehensive assessment of pain and sedation in both term and preterm neonates across various pain types (acute, chronic, postoperative),[7,13] could serve as a candidate for a more standardized reference scale in future multimodal research. This would improve consistency in data annotation and facilitate comparisons across studies.

Moreover, the ability of N-PASS to assess sedation levels opens an avenue for developing intelligent systems that not only assess pain but also monitor sedation, helping to avoid excessive analgesia and its associated risks—a critical unmet need in neonatal pain management.

Postoperative Pain Assessment and Missing Modality Handling

The assessment of postoperative pain using AI, particularly multimodal systems, is an area that warrants more attention. While acute procedural pain is transient, postoperative pain can be prolonged and fluctuating, posing a greater challenge for continuous and accurate assessment. Only three studies (33.33%) included in this review specifically addressed postoperative pain.[9,34,39] Salekin et al[9,39] pioneered work in this domain by recording neonates from preoperative baseline through to 3 hours post-surgery. However, as noted in their original work, this 3-hour window might coincide with emergence from anesthesia, and neonates often receive analgesia, potentially masking true pain expression. Future studies should consider extending recording durations and developing algorithms robust to the confounding effects of sedation and analgesia to capture a more complete picture of postoperative pain dynamics.

An additional challenge is handling missing modality data—a common occurrence in clinical practice. Four studies[9,32,36,45] did not explicitly address this issue, one[37] excluded incomplete cases, and four[34,38–40] described specific solutions, including attention-based reconstruction of missing modalities,[39] training on data with interference annotations,[40] or using algorithms inherently robust to missing data.[38] These findings underscore that model robustness to incomplete data is a critical requirement for real-world deployment.

Development and Clinical Translation of Multimodal Pain Assessment Systems

The development of next-generation multimodal neonatal pain assessment systems requires rigorous, multidimensional consideration of both technical and clinical implementation factors. Key considerations include: (1) minimizing invasiveness in data acquisition—such as reducing sensor burden and optimizing camera placement; (2) integrating validated clinical instruments like the Neonatal Pain, Agitation, and Sedation Scale (N-PASS) to enable simultaneous, differentiated assessment of pain and sedation effects, thereby supporting nuanced clinical decision-making; (3) establishing robust, standardized protocols for video annotation. These protocols should include mandatory independent dual annotation by trained raters, formal calculation of inter-rater reliability (eg., Cohen's kappa or intraclass correlation), and resolution of discrepancies by a certified neonatal pain assessment specialist to ensure annotation fidelity and modeling validity. Furthermore, at a frame rate of 30 fps, one hour of video yields 108,000 frames—a scale that imposes considerable demands on annotation time, personnel resources, and quality control. Equally critical are system-level practicalities: computational efficiency, real-time inference capability, hardware compatibility, and usability for frontline clinical staff—all of which directly influence clinical integration and sustainability. Although this review emphasizes algorithmic performance metrics, translational viability hinges equally on these operational dimensions. We therefore urge researchers to transparently report training and inference costs, hardware specifications, latency benchmarks, and deployment requirements to enable rigorous cost-effectiveness analyses, reproducibility, and scalable implementation across diverse healthcare settings.

Limitations

This systematic review has several limitations. Firstly, by focusing on automated multimodal assessment methods that reported specific performance metrics like accuracy or AUC, we may have excluded studies that developed multimodal datasets without implementing or reporting on computational algorithms, as well as studies using qualitative or unimodal outcome measures. Consequently, our review does not provide a comprehensive synthesis of unimodal approaches or standalone datasets. Secondly, while we summarized the types of feature extraction and classification methods, an in-depth technical analysis of the preprocessing steps for each modality was beyond the scope of this review, though these steps are crucial for algorithm performance. Thirdly, the exclusion of non-English language publications might have led to the omission of relevant studies from other regions. Finally, the inherent heterogeneity in the primary studies limited our ability to perform a meta-analysis and draw stronger quantitative conclusions about the optimal combination of modalities. Moreover, while this review focused on accuracy and AUC—the most prevalent metrics in current literature

—these do not fully address clinical reliability. Most included studies lacked advanced evaluations such as calibration and uncertainty quantification, which are essential for ensuring that model predictions are trustworthy in high-stakes neonatal care. Thus, future research should transition from solely optimizing discriminative performance to enhancing the reliability and interpretability of model outputs through these robust statistical frameworks.

Conclusion

The integration of multimodal data through artificial intelligence offers a scientifically sound and effective pathway toward automated and objective neonatal pain assessment, demonstrating superior accuracy compared to single-modal approaches. This advancement holds significant promise for improving pain management in vulnerable neonates. While current evidence, predominantly generated by computer scientists, has established strong algorithmic performance, critical gaps remain in external validation, clinical integration, and real-world applicability. The optimal combination of modalities and computational strategies for diverse clinical contexts remains undefined. Based on available evidence, we recommend a multimodal framework integrating facial expressions, body movements, crying patterns, and vital signs (heart rate, blood pressure, oxygen saturation, and respiratory rate). We further advocate standardizing video annotation using the N-PASS scale and extending postoperative pain recording durations to capture authentic pain variability.

Future research must prioritize large-scale, externally validated datasets that address postoperative and chronic pain while accounting for clinical confounders such as sedation. The ultimate translational goal is a multimodal monitoring system that provides continuous, automated pain classification with proportional clinical alerts, enabling timely, evidence-based interventions. Interdisciplinary collaboration between neonatologists, neonatal nurses, computer scientists, and engineers will be paramount in advancing this field and ultimately improving the well-being of neonates.

Data Sharing Statement

All data generated or analyzed during this study are included in this published article and its Supplementary Materials. Additional data are available from the corresponding author (Queyun Zhou.) upon reasonable request.

Acknowledgments

The authors express their sincere gratitude to all individuals who contributed to this systematic review. We also acknowledge the authors of the primary studies included in this review and the funding agencies that provided support for this study. We also wish to express our deep appreciation to Professors Zamzmi G and Salekin MS for their pioneering contributions to the field of multimodal neonatal pain assessment.

Funding

This work was supported by the Scientific Research Project of Shenzhen Science and Technology Innovation Commission (JCYJ20240813112516022).

Disclosure

The authors declare that they have no competing interests for this work.

References

1. Raja SN, Carr DB, Cohen M. et al. The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises. *Pain*. 2020;161(9):1976–1982. doi:10.1097/j.pain.0000000000001939
2. Groisman B, Bermejo-Sánchez E, Romitti PA, et al. Join World Birth Defects Day. *Pediatr Res*. 2019;86(1):3–4. doi:10.1038/s41390-019-0392-x
3. McPherson C, Miller SP, El-Dib M, et al. The influence of pain, agitation, and their management on the immature brain. *Pediatr Res*. 2020;88(2):168–175. doi:10.1038/s41390-019-0744-6
4. Walker SM. Long-term effects of neonatal pain. *Semin Fetal Neonatal Med*. 2019;24(4):101005. doi:10.1016/j.siny.2019.04.005
5. Hummel P, van Dijk M. Pain assessment: current status and challenges. *Semin Fetal Neonatal Med*. 2006;11(4):237–245. doi:10.1016/j.siny.2006.02.004
6. Carbajal R, Eriksson M, Courtois E, et al. Sedation and analgesia practices in neonatal intensive care units (EUROPAIN): results from a prospective cohort study. *Lancet Respir Med*. 2015;3(10):796–812. doi:10.1016/S2213-2600(15)00331-8
7. Zheng XL. Evidence-based guideline for neonatal pain management in China (2023). *Chin J Contemp Pediatr*. 2023;25:109–127.

8. Cong X, Wu J, Vittner D, et al. The impact of cumulative pain/stress on neurobehavioral development of preterm infants in the NICU. *Early Hum Dev*. 2017;108:9–16. doi:10.1016/j.earlhumdev.2017.03.003
9. Salekin MS, Zamzmi G, Goldgof D, et al. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Comput Biol Med*. 2021;129:1.
10. Peng T, Qu S, Du Z, et al. A Systematic Review of the Measurement Properties of Face, Legs, Activity, Cry and Consolability Scale for Pediatric Pain Assessment. *J Pain Res*. 2023;16:1185–1196. doi:10.2147/JPR.S397064
11. Stevens B, Johnston C, Petryshen P, et al. Premature Infant Pain Profile: development and initial validation. *Clin J Pain*. 1996;12(1):13–22. doi:10.1097/00002508-199603000-00004
12. Lawrence J, Alcock D, McGrath P, et al. The development of a tool to assess neonatal pain. *Neonatal Netw*. 1993;12(6):59–66.
13. Hummel P, Puchalski M, Creech SD, et al. Clinical reliability and validity of the N-PASS: neonatal pain, agitation and sedation scale with prolonged pain. *J Perinatol*. 2008;28(1):55–60. doi:10.1038/sj.jp.7211861
14. Krechel SW, Bildner J. CRIES: a new neonatal postoperative pain measurement score. Initial testing of validity and reliability. *Paediatr Anaesth*. 1995;5(1):53–61. doi:10.1111/j.1460-9592.1995.tb00242.x
15. Grunau RE, Oberlander T, Holsti L, et al. Bedside application of the Neonatal Facial Coding System in pain assessment of premature neonates. *Pain*. 1998;76(3):277–286. doi:10.1016/S0304-3959(98)00046-3
16. Pölkki T, Korhonen A, Axelin A, et al. Development and preliminary validation of the Neonatal Infant Acute Pain Assessment Scale (NIAPAS). *Int J Nurs Stud*. 2014;51(12):1585–1594. doi:10.1016/j.ijnurstu.2014.04.001
17. Debillon T, Zupan V, Ravault N, et al. Development and initial validation of the EDIN scale, a new tool for assessing prolonged pain in preterm infants. *Arch Dis Child Fetal Neonatal Ed*. 2001;85(1):F36–41. doi:10.1136/fn.85.1.F36
18. Andersen RD, Munsters JMA, Vederhus BJ, et al. Pain assessment practices in Swedish and Norwegian neonatal care units. *Scand J Caring Sci*. 2018;32(3):1074–1082. doi:10.1111/scs.12553
19. Brahnam S, Chuang CF, Shih FY, et al. Machine recognition and representation of neonatal facial displays of acute pain. *Artif Intell Med*. 2006;36(3):211–222. doi:10.1016/j.artmed.2004.12.003
20. Chen S, Luo F, Chen X, et al. A Video Database of Neonatal Facial Expression based on Painful Clinical Procedures. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference. 2019; 2019: 6565–6568.
21. Schiavenato M, von Baeyer CL. A Quantitative Examination of Extreme Facial Pain Expression in Neonates: the Primal Face of Pain across Time. *Pain Res Treat*. 2012;2012:251625. doi:10.1155/2012/251625
22. Lu GM, Yang C, Chen MY, et al. Sparse Representation Based Facial Expression Classification for Pain Assessment in Neonates. Paper presented at the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, Peoples R China. 2016.
23. Heiderich TM, Leslie ATFS, Guinsburg R. Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements. *Acta Paediatr Int J Paediatr*. 2015;104:e63–e69. doi:10.1111/apa.12861
24. Parodi E, Melis D, Boulard L, et al. Automated Newborn Pain Assessment Framework Using Computer Vision Techniques. Paper presented at the International Conference on Bioinformatics Research and Applications (ICBRA), Barcelona, SPAIN. 2017.
25. Cheng X, Zhu H, Mei L, et al. Artificial Intelligence Based Pain Assessment Technology in Clinical Application of Real-World Neonatal Blood Sampling. *Diagnostics*. 2022;12(8):1831. doi:10.3390/diagnostics12081831
26. Vincent KA, Srinivasan PMDR, Srinivasan K, et al. Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. *Front Public Health*. 2021;9:670352. doi:10.3389/fpubh.2021.670352
27. Aggarwal G, Jhajharia K, Izhar J, et al. A Machine Learning Approach to Classify Biomedical Acoustic Features for Baby Cries. *Journal of Voice*. 2025;39(6):1446–1455. doi:10.1016/j.jvoice.2023.06.014
28. Sun Y, de With PHN, Kommers D, et al. Automatic and Continuous Discomfort Detection for Premature Infants in a NICU Using Video-Based Motion Analysis. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. 2019;2019:5995–5999. doi:10.1109/EMBC.2019.8857597
29. Gruss S, Treister R, Werner P, et al. Pain Intensity Recognition Rates via Biopotential Feature Patterns with Support Vector Machines. *PLoS One*. 2015;10(10):e0140330. doi:10.1371/journal.pone.0140330
30. Faye PM, De Jonckheere J, Logier R, et al. Newborn infant pain assessment using heart rate variability analysis. *Clin J Pain*. 2010;26(9):777–782. doi:10.1097/AJP.0b013e3181ed1058
31. Ranger M, Gélinas C. Innovating in pain assessment of the critically ill: exploring cerebral near-infrared spectroscopy as a bedside approach. *Pain Manag Nurs*. 2014;15(2):519–529. doi:10.1016/j.pmn.2012.03.005
32. Egede J, Valstar M, Torres MT, et al. Automatic Neonatal Pain Estimation: an Acute Pain in Neonates Database. Paper presented at the 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, ENGLAND. 2019.
33. Salekin MS, Zamzmi G, Hausmann J, et al. Multimodal neonatal procedural and postoperative pain assessment dataset. *Data Brief*. 2021;35:106796. doi:10.1016/j.dib.2021.106796
34. Zamzmi G, Pai CY, Goldgof D, et al. A Comprehensive and Context-Sensitive Neonatal Pain Assessment Using Computer Vision. *IEEE Transactions on Affective Computing*. 2022;13(1):28–45. doi:10.1109/TAFFC.2019.2926710
35. Yang N, Zhuang Y, Jiang H, et al. Developing and Validating a Multimodal Dataset for Neonatal Pain Assessment to Improve AI Algorithms With Clinical Data. *Adv Neonatal Care*. 2024;24(6):578–585. doi:10.1097/ANC.0000000000001205
36. Zamzmi G, Pai CY, Goldgof D, et al. An approach for automated multimodal analysis of infants' pain. Paper presented at the 2016 23rd International Conference on Pattern Recognition (ICPR). 2016.
37. Zamzmi G, Pai CY, Goldgof D, et al. Automated Pain Assessment in Neonates. Paper presented at the 20th Scandinavian Conference on Image Analysis (SCIA), Tromso, NORWAY. 2017.
38. van der Vaart M, Duff E, Raafat N, et al. Multimodal pain assessment improves discrimination between noxious and non-noxious stimuli in infants. *Paediatr Neonatal Pain*. 2019;1(1):21–30. doi:10.1002/pne2.12007
39. Salekin MS, Zamzmi G, Goldgof D, et al. Attentional Generative Multimodal Network for Neonatal Postoperative Pain Estimation. *Med Image Comput Comput Assist Interv*. 2022;13433:749–759. doi:10.1007/978-3-031-16437-8_72

40. Zhu H, Zhao Y, Chen X, et al. Video-Based Neonatal Pain Assessment in Uncontrolled Conditions. *IEEE J Biomed Health Inform*. 2024;28 (1):239–250. doi:10.1109/JBHI.2023.3324537
41. Cheng D, Liu D, Philpotts LL, et al. Current state of science in machine learning methods for automatic infant pain evaluation using facial expression information: study protocol of a systematic review and meta-analysis. *BMJ Open*. 2019;9(12):e030482. doi:10.1136/bmjopen-2019-030482
42. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372: n71. doi:10.1136/bmj.n71
43. Huang X, Lin J, Demner-Fushman D Evaluation of PICO as a knowledge representation for clinical questions. *AMIA Annu Symp Proc*. 2006; 2006: 359–363.
44. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536. doi:10.7326/0003-4819-155-8-201110180-00009
45. Salekin MS, Zamzmi G, Goldgof D, et al. Multi-Channel Neural Network for Assessing Neonatal Pain from Videos. Paper presented at the Proc. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy. 2019.
46. Schiavenato M, Byers JF, Scovanner P, et al. Neonatal pain facial expression: evaluating the primal face of pain. *Pain*. 2008;138(2):460–471. doi:10.1016/j.pain.2008.07.009
47. Stevens BJ, Gibbins S, Yamada J, et al. The premature infant pain profile-revised (PIPP-R): initial validation and feasibility. *Clin J Pain*. 2014;30 (3):238–243. doi:10.1097/AJP.0b013e3182906aed
48. Grunau RVE, Craig KD. Pain expression in neonates: facial action and cry. *Pain*. 1987;28(3):395–410. doi:10.1016/0304-3959(87)90073-X
49. Rui C, Yang Y, Shi Y, et al. Expert consensus on neonatal pain assessment and analgesia management (2020). *Chin J Contemp Pediatr*. 2020;22:923–930.
50. Gendras J, Lavenant P, Sicard-Cras I, et al. The newborn infant parasympathetic evaluation index for acute procedural pain assessment in preterm infants. *Pediatr Res*. 2021;89(7):1840–1847. doi:10.1038/s41390-020-01152-4
51. Roué JM, Rioualen S, Gendras J, et al. Multi-modal pain assessment: are near-infrared spectroscopy, skin conductance, salivary cortisol, physiologic parameters, and Neonatal Facial Coding System interrelated during venepuncture in healthy, term neonates? *J Pain Res*. 2018;11:2257–2267. doi:10.2147/JPR.S165810
52. Baxter L, van der Vaart M, Cobo MM, et al. Is noxious stimulus-evoked electroencephalography response a reliable, valid, and interpretable outcome measure to assess analgesic efficacy in neonates? A systematic review and individual participant data (IPD) meta-analysis protocol. *Syst Rev*. 2025;14(1):152. doi:10.1186/s13643-025-02890-4
53. Schiavenato M, Butler-O'Hara M, Scovanner P. Exploring the association between pain intensity and facial display in term newborns. *Pain Res Manag*. 2011;16(1):10–12. doi:10.1155/2011/873103
54. Slater R, Cantarella A, Gallella S, et al. Cortical pain responses in human infants. *J Neurosci*. 2006;26(14):3662–3666. doi:10.1523/JNEUROSCI.0348-06.2006
55. Bellieni CV. Pain assessment in human fetus and infants. *AAPS J*. 2012;14(3):456–461. doi:10.1208/s12248-012-9354-5
56. Roué JM, Avnit A, Gholami B, et al. Objective Detection of Newborn Infant Acute Procedural Pain Using EEG and Machine Learning Algorithms. *Paediatric and Neonatal Pain*. 2025;7(1):e70001. doi:10.1002/pne2.70001