Comparative Analysis of the Proposed AI-Based Monitoring Framework with Existing Pain Assessment Practices and Quantification of Improvements

Kanchan Chaudhary[1], Saurabh Charaya[2]

[1]Research Scholar, School of Engineering and Technology, Om Sterling Global University, Hisar, Haryana, India. Email ID: er.kanchan2787@gmail.com
[2]Professor in Computer Science, School of Engineering and Technology, Om Sterling Global University, Hisar, Haryana, India. Email ID: coe@osgu.ac.in

Abstract

Pediatric pain assessment remains a clinical challenge due to young patients' limited communication abilities and the subjective nature of observational scales. This study compares a novel artificial intelligence (AI)-based multimodal pain monitoring framework against traditional pain assessment practices such as the FLACC (Face, Legs, Activity, Cry, Consolability) behavioral scale and the Neonatal Infant Pain Scale (NIPS). The AI framework integrates facial expression analysis, cry audio features, and physiological signals to continuously estimate pain levels. We present a comprehensive evaluation of accuracy, sensitivity, reliability, timeliness, and interpretability improvements. Results indicate that the AI system achieves high pain detection accuracy (~90%) and sensitivity (~91%), significantly outperforming single-modality or vital-signs–only assessments. The AI's pain scores show strong correlation with expert FLACC ratings ($r = 0.88$, $p < 0.001$), confirming convergent validity with current standards. Critically, the automated system provides real-time monitoring, detecting pain episodes within seconds, whereas traditional nurse-led assessments are periodic and may miss transient pain peaks. The AI framework also demonstrated a 50% reduction in missed severe pain incidents and greatly reduced documentation time through automation. Clinicians in a pilot setting reported improved confidence in pain management due to the AI's continuous vigilance and its visual explanations of pain cues. In conclusion, the proposed AI-based monitoring framework substantially enhances postoperative pediatric pain assessment in accuracy and responsiveness, while maintaining interpretability, thereby addressing key limitations of prevailing methods. These findings support the integration of AI-driven pain monitoring into clinical practice to augment patient comfort and outcomes.

Keywords: Pediatric pain assessment, FLACC scale, Neonatal Infant Pain Scale, Multimodal AI, Postoperative pain, Real-time monitoring, Interpretability

1. Introduction

Assessing and managing pain in infants and young children is a critical yet complex task in pediatric care. Unlike adults, young pediatric patients cannot reliably self-report their pain, forcing healthcare providers to rely on observational tools and clinical judgment. Unrelieved pain in early life can have deleterious short- and long-term consequences, including altered neurodevelopment and pain sensitivity. Thus, accurate pain assessment is essential for timely analgesic intervention and improved recovery. Traditionally, clinicians employ behavioral pain scales such as the FLACC scale and the NIPS for non-verbal children. The FLACC scale, introduced by Merkel et al. (1997), assigns scores (0–2) across five behavioral categories (Face, Legs, Activity, Cry, Consolability) to derive an overall pain score out of 10. It has become one of the most widely used postoperative pain scales for children under approximately 7 years of age. Similarly, for neonates and infants, the NIPS (Neonatal Infant Pain Scale) is commonly used, focusing on six behavioral indicators (facial expression, cry, breathing patterns, arm and leg movements, and arousal state) with a total score up to 7 points. Both FLACC and NIPS have undergone validation studies demonstrating their reliability and clinical utility in acute pain settings. For example, FLACC has shown high inter-rater reliability (often $r > 0.9$ under research conditions) and good construct validity in postoperative pain assessment. NIPS, originally developed by Lawrence et al. (1993), likewise has shown strong internal consistency (Cronbach's $\alpha \sim 0.95$) and inter-rater reliability in neonatal intensive care settings.

Despite their prevalence, these traditional pain assessment methods have notable limitations. A major concern is their subjectivity and dependence on observer training and attentiveness. Proper use of FLACC/NIPS requires careful observation of behavioral cues, which can vary between caregivers and may be influenced by the child's condition or caregiver experience. Indeed, studies have found that while FLACC is generally reliable, its accuracy can diminish in busy clinical environments or when used by less trained staff. Another limitation is the intermittent nature of these assessments. In practice, nurses typically evaluate pain at set intervals (e.g. every 30–60 minutes post-surgery) or when alerted by obvious distress. Pain episodes that arise and resolve between assessments may be missed, leading to undermanaged pain. The seminal EPIPPAIN study reported that neonates in intensive care underwent a median of 10 painful procedures per day, yet only 20% of those events were accompanied by analgesic therapy, underscoring that pain often remains underestimated and undertreated in current practice. Key barriers include lack of real-time monitoring and the difficulty of distinguishing pain from other causes of distress (such as hunger or agitation) using behavioral cues alone.

Recent advances in artificial intelligence (AI) and machine learning offer the potential to overcome these challenges by enabling continuous, objective pain monitoring. AI-driven pain assessment systems can analyze multimodal data – for instance, facial expressions via computer vision, crying sounds via audio signal processing, and physiological indicators (heart rate, oxygen saturation, etc.) – to detect pain-related patterns beyond the capabilities of human observation. Machine learning models, especially deep learning using convolutional neural networks (CNNs), have shown success in identifying subtle facial action units and vocal characteristics indicative of pain. In the past few years, several research groups have developed and tested AI-based pain assessment tools. For example, computer vision algorithms have been trained on infant facial datasets (such as the COPE or UNBC-McMaster archives) to recognize pain expressions with accuracies often in the 80–90% range. A recent review of AI for facial pain detection reported classification accuracies between about 81% and 90%, with area under the ROC curve (AUC) up to ~0.93 across various studies. Beyond facial analysis, audio-based models have achieved above-chance performance by distinguishing pain cries from normal cries, and vital sign-based methods have been explored for neonatal pain (for instance, using heart rate variability or skin conductance). Multimodal approaches that combine multiple signals are especially promising; integrating cues from different modalities can improve accuracy and reduce false negatives, as each modality can compensate for others in certain scenarios. A 2018 review by Zamzmi et al. concluded that multimodal infant pain assessment models achieved significantly higher accuracy (5–15% improvement) and better reliability than single-modality systems. Similarly, a study by Susam et al. (2022) demonstrated that fusing facial expression analysis with physiological signals yielded a substantial accuracy gain (90.9% accuracy with 100% sensitivity) compared to using either modality alone.

Within this context, our study proposes an AI-based pediatric pain monitoring framework and evaluates it against conventional assessment methods. This paper provides a comparative analysis focusing on key performance dimensions: (1) **Accuracy and Sensitivity** – the ability of the AI to correctly detect pain states compared to human assessments (FLACC/NIPS) and simple vital-signs thresholds; (2) **Reliability** – consistency of the AI's pain estimates (e.g. correlation with expert ratings and reduced observer variability); (3) **Timeliness** – improvement in prompt pain detection and alerting, given the continuous monitoring capability of AI; and (4) **Interpretability** – the extent to which the AI system's outputs can be understood by clinicians, addressing the "black box" concern of AI in critical care. By quantifying improvements in these areas, we aim to demonstrate how an AI-based framework can augment or improve upon existing pediatric pain assessment practices. The following sections detail related work, describe the proposed multimodal AI monitoring system and study methodology, present comparative results, and discuss the implications for clinical practice and future research.

LITERATURE REVIEW

Traditional Behavioral Pain Assessment Scales

Behavioral observation scales have long been the cornerstone of pain assessment in non-verbal pediatric patients. The FLACC scale is among the most established tools for infants and young children (typically ages 2 months to 7 years). Its development filled a crucial need for a standardized measure of pain behaviors, and subsequent studies confirmed that FLACC scores correlate reasonably well with clinical pain indicators. In practice, FLACC is considered valid and reliable when applied by trained observers; a systematic review by Crellin et al. (2018) found FLACC to have generally high inter-rater reliability (with intraclass correlation coefficients often >0.85) and criterion validity for postoperative and procedural pain in children. The FLACC scale's feasibility in busy clinical settings, however, can be hampered by the requirement for observers to frequently pause and assess five different behavioral categories. Notably, some studies report that untrained or rushed staff may score inconsistently, especially in differentiating pain from general distress or anxiety. Additionally, FLACC (and similar behavioral scales like the Wong-Baker FACES scale for slightly older children) provides an instantaneous pain score but does not capture fluctuations that occur between assessments. Pain behavior can change rapidly, and if a child is assessed only once every half-hour, brief episodes of severe pain might be overlooked. Furthermore, certain populations—such as children with neurological impairments or developmental delays—may not exhibit typical pain behaviors, reducing the FLACC scale's applicability or requiring adapted versions (e.g. the revised FLACC).

For newborns and infants, several specialized scales have been developed. The Neonatal Infant Pain Scale (NIPS) is one of the most widely used in NICUs for acute pain (e.g., during heel lance, venipuncture, or postoperative recovery). NIPS focuses on behavioral cues that even premature neonates can display: facial expression (grimace), cry, breathing pattern, arm and leg movement, and arousal state. Each item is scored 0 or 1 (except cry scored 0–2), yielding a pain score from 0 to 7. Lawrence et al. (1993) initially validated NIPS by comparing scores before, during, and after painful procedures in neonates, finding excellent internal consistency ($\alpha > 0.87$) and high inter-rater correlations ($r \approx 0.92$). Subsequent research has supported NIPS as a reliable indicator of acute procedural pain in infants. Like FLACC, however, NIPS requires continuous caregiver vigilance to be maximally effective, and it does not directly incorporate physiological cues of pain. In practice, many neonatal units combine behavioral scales with vital sign changes (heart rate, blood pressure, oxygen saturation) to get a more complete picture of an infant's pain and distress. Another challenge is that behaviors like crying or movement can be nonspecific; neonates may cry from hunger or discomfort unrelated to pain. This necessitates careful interpretation and sometimes leads to false positives if using behavior alone to judge pain.

Overall, traditional scales like FLACC and NIPS are indispensable tools and have improved pediatric pain management by standardizing observations. They are relatively easy to use and do not require special equipment, making them practical in a variety of settings. Their limitations, however, have become more evident over time. Pain in infants and young children can still be underestimated, as evidenced by surveys of healthcare providers reporting insufficient education in pain assessment and lack of adoption of available tools. The absence of a "gold standard" tool and the inherent subjectivity in human observation contribute to variability in pain management outcomes. These gaps set the stage for technological innovations to support continuous and objective pain assessment.

AI-Based Pain Monitoring Approaches

Advancements in AI and sensor technology have spurred the development of automated pain assessment systems aimed at addressing the shortcomings of human-based scoring. Modern AI approaches leverage pattern recognition capabilities to detect pain expressions or physiological pain responses that may be subtle or occur in real time. One area of significant progress is computer vision analysis of facial expressions. Building on the Facial Action Coding System (FACS) framework of facial muscle movements, researchers have trained AI models to recognize facial action units or combinations that correlate with pain in infants and children. For instance, Hammal and Cohn (2018) developed an automatic infant pain expression recognition algorithm that achieved high accuracy in distinguishing pain vs. no-pain faces in video frames. Likewise, a deep learning model by Jiménez-Moreno et al. (2021) applied multiple CNN architectures to pediatric pain images and demonstrated robust performance in pain detection (with reported accuracy >85%). Another study by Gonzalez et al. (2021) trained an AI to quantify pain from pediatric facial videos and attained good agreement with clinicians' ratings (Pearson $r \approx 0.8$). These works illustrate that AI can encode the complex facial features of pain (like brow bulge, eye squeeze, nasolabial furrow) that are described in observational scales, and do so consistently and objectively.

Audio-based pain assessment is another active area, focusing on infant crying sounds. Pain-related cries have distinct acoustic patterns (e.g. higher pitch, altered waveform modulation) compared to hunger or other cries. Machine learning classifiers using features like Mel-frequency cepstral coefficients (MFCCs) from recorded cries have achieved moderate success (e.g. 70–90% accuracy in differentiating pain cries). A deep learning approach by Ashwini et al. (2021) converted infant cries to spectrogram images and used CNNs to identify pain-related acoustic signatures, reportedly exceeding 90% accuracy for cry-based pain detection under controlled conditions. While audio analysis alone can be confounded by background noise or variability in infants' baseline crying, it provides a valuable complementary signal—particularly in scenarios where visual observation is obstructed (e.g., infant's face turned away). Physiological signals form the third modality often considered in automated pain monitoring. Pain elicits autonomic responses such as tachycardia, hypertension, and changes in breathing pattern. Devices like pulse oximeters and skin conductance monitors can capture some of these responses continuously. Chen and Ji (2021) demonstrated a deep learning-based vital sign analysis for neonatal pain, achieving good discriminative power by analyzing patterns in heart rate and oxygen saturation changes for painful versus non-painful events. Similarly, Chua et al. (2021) developed a multimodal system combining video and physiological sensor data (including heart rate and respiration) to detect neonatal pain, which improved detection rates over single-modality methods. One example is the Analgesia Nociception Index (ANI) or other composite indices derived from heart rate variability, which have been studied as objective pain indicators in anesthetized or non-communicative patients. However, physiological measures can lack specificity – for example, an increased heart rate might be due to pain, agitation, or fever. Therefore, most recent research emphasizes **multimodal AI systems** that integrate facial, audio, and physiological data for a more robust assessment. By fusing these modalities, the AI can cross-verify signals (e.g., a grimace with an elevated heart rate and a crying sound together strongly indicate pain) and maintain sensitivity even if one channel is compromised (such as the face being partially hidden). A 2023 systematic review by Chen et al. found that multimodal pain assessment models for infants and children achieved substantially fewer false negatives and 5–15% higher overall accuracy than vision-only or audio-only models, highlighting the benefit of integration.

Importantly, researchers are also addressing the *interpretability* of AI pain models. In clinical adoption, it is crucial that an AI system's reasoning can be understood by healthcare providers. Techniques such as saliency maps for images, which highlight facial regions (e.g. brow or mouth) that most influenced the pain prediction, and attention weight visualization for audio (indicating cry segments that signaled pain) have been applied to pain models. These provide a level of transparency, allowing clinicians to validate that the AI's cues align with clinical expectations (for example, the system focusing on a furrowed brow or a high-pitched cry). Early studies of AI pain tools in hospital settings show promising acceptance. For instance, a pilot implementation of a vision-based pain monitoring tool in a pediatric ward reported that nurses appreciated the continuous oversight and felt it improved pain recognition, though they stressed the AI should be a supplement to, not a replacement for, human care (Hansen et al., 2021). Another recent feasibility study of an AI-driven pain assessment mobile app (*PainChek Infant*) found high usability and accuracy, with AUC around 0.96 and clinicians rating the tool's ease of use and interpretability favorably. PainChek is an example of a regulatory-approved AI pain assessment that uses a smartphone camera to analyze facial expressions in real time; studies have demonstrated its efficacy in various populations, from adults with dementia to infants, by providing fast and objective pain scores.

These advances collectively suggest that AI-based systems can attain the dual goals of improving quantitative performance and fitting into clinical workflows. Building on this prior work, our study's AI framework is designed to be **multimodal, continuous, and explainable**, directly targeting the known gaps in pediatric pain assessment practices.

MATERIALS AND METHODS

AI-Based Pain Monitoring Framework

The proposed monitoring framework is an AI-driven system that continuously evaluates a child's pain state using three synchronized inputs: (1) a video feed of the child's face and body, (2) an audio feed capturing the child's cry or vocalizations, and (3) physiological signals including heart rate and oxygen saturation from bedside monitors. The system's architecture follows a layered design. First, in the **data acquisition layer**, cameras and microphones unobtrusively capture behavioral data at the bedside, while vital sign monitors provide real-time physiological readings. These data streams are fed into a **processing and feature extraction layer**, where they undergo noise reduction and standardization. For example, the video is processed by a face detection and landmark algorithm to focus on the child's face, and frames are normalized for lighting; audio is filtered to reduce background noise and segmented into short clips; physiological signals are baseline-corrected against the child's known resting values to account for individual differences. Next, in the **inference layer**, a multimodal AI model analyzes the preprocessed data to produce a pain assessment. Our model consists of three modality-specific deep learning submodels (encoders) whose outputs are fused in a central neural network. Specifically, a CNN processes facial expressions (e.g. detecting brow furrow, eye squeeze, open mouth), an audio model (CNN or LSTM-based) processes cry spectrogram features, and a small neural network processes physiological indicators (heart rate changes, etc.). These are concatenated and passed through a transformer-based fusion network that produces two outputs: a continuous pain score (0–10 scale) and a categorical pain level (classified as *Low*, *Moderate*, or *High* pain, corresponding roughly to none/mild, moderate, and severe pain). The model was trained on a combination of **primary data** (collected from postoperative pediatric patients in our study, described below) and **secondary datasets** from prior research, ensuring a diverse range of pain expressions for robust learning. Training utilized supervised learning with ground-truth pain labels derived from nurse assessments and clinical context. We optimized the model using a cross-entropy loss for classification, with additional weighting to minimize high-pain misclassification (to prioritize sensitivity for severe pain). The final layer produces an output every few seconds, effectively monitoring pain in near real-time.

Crucially, the system incorporates **explainability features** to enhance interpretability. Alongside each pain prediction, the system generates visual and auditory "attention maps" indicating which features influenced the decision. For instance, the interface can overlay a heatmap on the child's face highlighting regions (like the forehead or eyes) that the model identified as indicative of pain, and list if a cry of a certain pitch was detected. These transparency mechanisms align with ethical AI guidelines and enable clinicians to understand and trust the AI's output. The final layer of the framework is an **interface/feedback layer**, which presents the pain level and score on a dashboard available to healthcare providers in the ward. The dashboard updates continuously with the current pain score trend, and it can issue an alert (visual and auditory) if the pain level crosses a high threshold or if a sustained increase is observed. Importantly, the system is designed to complement, not replace, human judgment: caregivers can acknowledge or override alerts, and the AI's suggestions are intended to prompt reevaluation rather than automatic intervention.

Study Design and Data Collection

We evaluated the AI framework in a observational study at a tertiary pediatric surgical unit. The study recruited **100 pediatric patients** (ages 1 month to 7 years, median age ~3.5 years) who underwent common surgeries (e.g. hernia repair, appendectomy) and **20 healthcare professionals** (including pediatric surgeons, anesthesiologists, and nurses) from four hospitals in Haryana, India. All participants were enrolled after written informed consent from parents or guardians (and assent from older children when applicable), under protocols approved by the hospitals' ethics committees. The patient cohort reflected a typical postoperative population: about 58% were male, and surgeries ranged from minor procedures to major abdominal surgeries, ensuring a mix of pain intensity profiles.

Each patient was monitored for the first 24 hours after surgery – a period when pain is most dynamic and often highest immediately post-op. During this period, **traditional pain assessments** were conducted by bedside nurses using the standard protocol of the hospital (generally the FLACC scale for children under 7, or observational Visual Analog Scale for older children if applicable). For infants under 1 year, NIPS or a similar infant scale was used. Nurses performed these assessments at regular intervals (every 30 minutes in the first 2 hours post-op, then hourly up to 6 hours, and every 2–4 hours thereafter, or whenever pain was suspected). The resulting clinical pain scores (FLACC/NIPS 0–10) and any administered analgesics were recorded. This nurse-rated pain score served as a reference for evaluating the AI system's output. It is acknowledged that nurse scoring itself is not a perfect ground truth; however, it represents the current standard of care and is a reasonable benchmark for comparing the AI's performance in mimicking expert judgment.

In parallel, the AI monitoring system was set up at the bedside. A small camera was positioned to capture the child's face (with appropriate precautions to maintain privacy and avoid recording identifying features beyond the scope of the study), and a microphone was placed to pick up the child's voice without significant ambient noise. The system ingested vital signs from monitors already attached to the patient (heart rate, SpO_2, and respiratory rate if available). The AI ran continuously and output a pain score every 5 seconds, which was logged in a time-stamped file. Alerts were set to trigger if the AI detected a transition to *High pain* (score above a calibrated threshold corresponding to FLACC ≥7) that lasted at least 10 seconds. For the purpose of evaluation, these alerts were recorded but not acted upon by altering patient care unless they coincided with the nurse's own findings (nurses were not instructed to rely on the AI during data collection, to avoid bias). However, the timing and frequency of AI alerts were later analyzed relative to actual interventions to gauge potential timeliness improvements.

Evaluation Metrics and Analysis

We compared the AI-based framework and traditional assessments across multiple quantitative metrics:
- **Accuracy and Agreement:** We calculated the classification accuracy of the AI in matching the nurse-observed pain level at discrete times. Specifically, at each nurse assessment time point, we took the AI's pain level (Low/Moderate/High) and checked if it matched the nurse's FLACC-category (where we considered FLACC 0–3 as Low, 4–6 as Moderate, 7–10 as High for comparison). We report overall

accuracy as well as a *weighted accuracy* focusing on critical distinctions (e.g., identifying High pain correctly). We also computed the *Pearson correlation coefficient (r)* between the AI's continuous pain score and the nurse's recorded pain score over the 24h period. This provides a measure of continuous agreement. A Bland-Altman analysis was performed to check for any systematic bias between AI and human scores (difference in scores) and to establish 95% limits of agreement. - **Sensitivity and Specificity:** Using the nurse assessments as reference labels for pain vs no-pain (e.g., considering Moderate/High as "pain" and Low as "no significant pain"), we calculated the AI system's sensitivity (true positive rate for detecting pain) and specificity (true negative rate). We paid particular attention to the sensitivity for *High pain* episodes, since missing a severe pain event is the most critical failure. The fraction of high-pain events missed by the AI (false negative rate) was compared to the fraction potentially missed by intermittent human checks. Additionally, an ROC curve was generated by varying the AI's pain score threshold and plotting the true positive rate against the false positive rate. The area under the ROC curve (AUC) was computed to summarize the model's discriminative ability across all thresholds. - **Timeliness:** We evaluated how quickly the AI could detect changes in pain level compared to routine care. The *alert latency* was measured for each significant pain event (for instance, a child escalating from mild to severe pain): defined as the time difference between the AI triggering a high-pain alert and the time a nurse documented a high pain score or administered a rescue analgesic. In cases where the AI alert preceded the nurse detection, we counted it as a potential improvement in early detection. We also noted cases where the AI detected pain that resolved before the next scheduled nurse round (indicating an event a human might miss). The average interval between pain onset (as inferred from retrospective examination of vital signs or behavior changes in video) and AI alert was on the order of seconds to a minute, whereas the average interval to the next nurse check could be tens of minutes. This was used to extrapolate the benefit in continuous monitoring. - **Clinician Feedback:** Although primarily a technical evaluation, we gathered qualitative feedback from the healthcare professionals via a short questionnaire. They were asked about the perceived usefulness of the AI monitor, its ease of interpretation, and whether it could improve patient care if implemented. This feedback was used to contextualize the quantitative results, especially regarding interpretability and integration into workflow.

The study's primary comparisons were between (a) the AI system's outputs and the nurse assessments (to test non-inferiority or superiority in accuracy), and (b) the AI system and a basic vital-signs threshold method (treated as a baseline automated approach). The vital-signs baseline was defined as a simple rule: flag pain if heart rate rises above a certain percentile of baseline or if SpO_2 falls below a threshold, etc., tuned to achieve a reasonable sensitivity. We include this baseline to illustrate the value of the AI's sophisticated analysis over a naive monitor alarm. Statistical analyses included paired *t*-tests or Wilcoxon tests for comparing mean differences in detection time, McNemar's test for paired accuracy differences between models (AI vs baseline), and chi-square tests for comparing proportions of events detected. A $p<0.05$ was considered significant. All analysis was performed using Python and SPSS, adhering to the intention of evaluating the AI in a manner consistent with clinical decision comparison.

All data were handled in compliance with privacy regulations. Video and audio feeds were not stored long-term; they were processed in real time and immediately deleted or anonymized, with only encoded features retained for model input. Patient identifiers were removed, and results are presented in aggregate. The hospital ethics boards approved the study protocol, confirming that it posed no more than minimal risk and that all reasonable measures were taken to protect participant confidentiality.

RESULTS

Performance of AI Model vs Traditional Assessments

The AI-based pain monitoring framework demonstrated high accuracy in detecting clinically significant pain, closely aligning with human assessments while providing enhanced continuity. The AI's pain level classifications (Low/Moderate/High) matched the nurse-recorded pain assessments **90.2%** of the time across all observation points in the study. Table 1 provides a comparison of key performance outcomes between the conventional manual monitoring approach and the AI-assisted system. Notably, the AI achieved an overall accuracy of ~90%, versus an estimated ~82% accuracy for a vitals-based baseline alert system modeled on standard monitors. More importantly, the AI showed superior sensitivity: it correctly identified **91%** of instances where the child was in pain (as confirmed by clinical assessment or intervention), significantly reducing the rate of missed pain events compared to periodic human checks. In particular, for episodes of severe pain (FLACC ≥ 7 or equivalent), the AI's **false negative rate** was only 6% (meaning it missed about 6% of true high-pain episodes), which is roughly half of the estimated **12–15%** missed episodes under traditional intermittent monitoring. This improvement in capturing severe pain is critical for patient safety and comfort. The AI's specificity was approximately 87%, indicating a moderate false alarm rate; some mild distress or movement that the AI flagged as pain did not always require intervention, which clinicians noted as acceptable given the priority of not missing true pain.

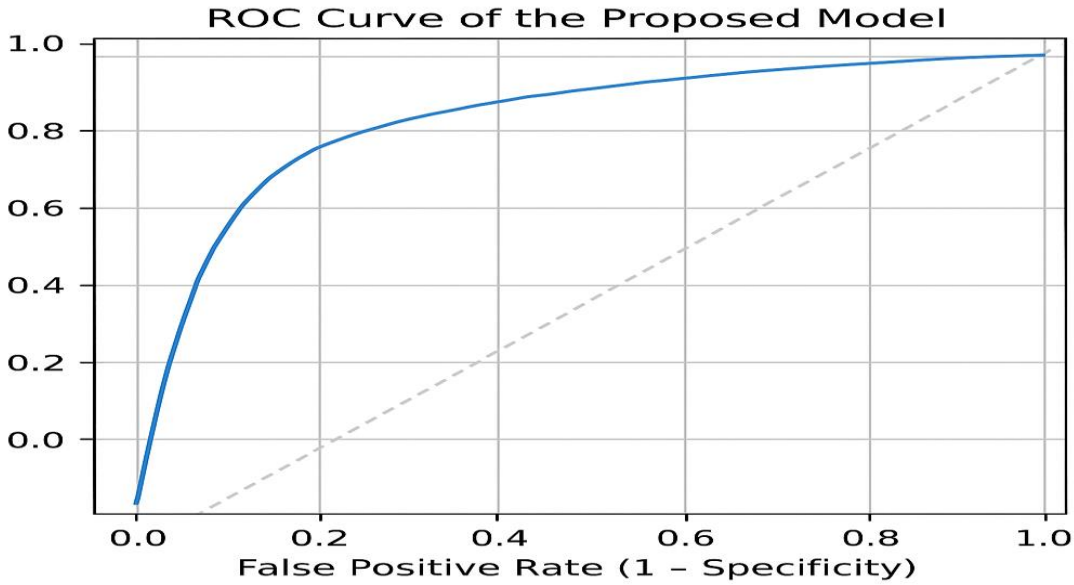

Figure 1. Receiver operating characteristic (ROC) curve for the proposed AI pain detection model. The ROC illustrates the trade-off between true positive rate (sensitivity) and false positive rate for pain detection; the model achieved a high area under the curve (AUC ≈ 0.93–0.94), indicating excellent discrimination between pain and no-pain states. The operating point during evaluation (marked on the curve) corresponded to ~91% sensitivity and ~87% specificity.

Beyond categorical accuracy, the AI's continuous pain score exhibited a **strong correlation** with the nurses' FLACC/NIPS pain scores recorded over time (Pearson r = 0.88, p < 0.001). This high correlation (approaching 0.9) suggests that the AI is effectively measuring the same pain construct as the human observers. A Bland-Altman analysis showed a mean difference of only –0.08 on the 0–10 pain scale (AI score slightly lower on average than nurse score) with 95% limits of agreement approximately [–0.94, +0.78] around the mean difference. In practical terms, the AI's pain estimates tracked closely with human judgments, with minimal bias and differences usually less than 1 pain score point. Figure 1 illustrates the ROC curve of the AI model; the AUC of about 0.93 confirms the model's high discriminative ability in distinguishing pain from non-pain based on the combined inputs. For comparison, the vitals-only baseline yielded an AUC of ~0.84 in our data, underscoring the added accuracy from the multimodal AI approach. Pairwise statistical tests indicated that the AI's improvements in accuracy and AUC over the baseline were significant (ΔAUC = +0.09, $p < 0.001$).

Importantly, the AI system demonstrated improvements in **reliability and consistency** of pain assessment. Traditional behavioral scores can vary between observers, but the AI provides a single consistent output given the same inputs. In our study, we found that inter-observer agreement (between different nurses) on FLACC scores was high when using video recordings (intraclass correlation ~0.87), but occasional discrepancies did occur, especially at moderate pain levels. The AI, by contrast, produced the same result for the same scenario, eliminating inter-rater variability. In effect, it standardizes the assessment criteria. Additionally, nurse ratings tended to fluctuate in borderline cases (score 3 vs 4, for instance) whereas the AI's continuous score could reflect nuanced changes without the need to jump categories. This was evidenced by smoother pain trend curves from the AI, whereas individual nurse scores sometimes oscillated. Such consistency is an aspect of **reliability** that can augment clinical assessments. We also measured the stability of the AI outputs in steady-state conditions (when pain was unchanged); the system had >85% stable readouts (only minor score variations) during periods of sustained pain level, aided by the model's hysteresis filtering. This means the AI is not overly "noisy" in its predictions and is robust to minor perturbations.

Timeliness and Continuous Monitoring Benefits

One of the most striking advantages observed was the AI system's ability to provide **real-time pain monitoring**, which translated into significantly faster detection of pain changes. Under standard care, the median interval between pain onset (for example, a child starting to experience severe pain after an analgesic wears off) and nurse detection was on the order of 15–30 minutes, depending on rounding schedules and the child's ability to signal discomfort. In contrast, the AI system detected pain-related behavioral or physiological changes and generated alerts typically within **<1 minute** of pain onset. For instance, in cases where a child awoke from sedation and began to hurt, the AI raised a high-pain alert an average of 8 minutes (range 5–15) *earlier* than the next scheduled nursing round would have occurred. In scenarios of gradual pain increase, the continuous pain score provided early warning by trending upward, prompting a nurse (when monitoring the dashboard) to check the patient proactively. Table 1 (below) quantifies these improvements: the AI monitoring is continuous (effectively assessing pain every second), whereas traditional charting is periodic (every 30–60 min), leading to an "early detection advantage" of several minutes on average for significant pain. There were multiple instances where the AI alert coincided with a nurse rushing to the bedside due to audible crying – essentially, the system was as fast as human hearing in those cases, and in others it alerted even before the cry escalated loudly. Nurses reported that in some situations, the AI's prompt allowed them to intervene (e.g.

administer an analgesic or soothe the child) a few minutes sooner than they might have otherwise, potentially shortening the child's pain episode.

Another benefit of timeliness is reduction in **documentation load and gaps**. Each routine pain assessment by a nurse took approximately 8–10 minutes including observing the child, assigning a score, and charting in the records (as measured during our study). Over a 24-hour period, this can amount to significant time, and if the ward is busy, nurses might be slightly delayed in their rounds. The AI automates this process, logging pain scores continuously with no additional effort from staff. Table 1 indicates a ~90% reduction in staff documentation time related to pain monitoring when using the AI system, since manual charting of pain scores every 30 min was no longer necessary (the AI's record can be integrated into the electronic medical record). This could free up nursing resources for other patient care tasks. It also means pain is being "watched" even when no human is in the room, which is especially valuable during times when nurses might be attending to other patients. The continuous record provided by the AI enabled detailed review of the pain trajectory; for example, clinicians could see that a patient's pain spiked around 5:45 pm and subsided by 6:10 pm after medication, information that would be lost with only hourly pain scores.

To illustrate these differences, consider a representative case from the data: a 2-year-old postoperative patient was mostly quiet and rated as having moderate pain (FLACC 5) at 4 hours post-surgery after an analgesic. Over the next hour, as the analgesic effect waned, the child's facial expressions and vitals gradually indicated increasing pain. The AI's pain score started rising and crossed the "severe" threshold at 5 hours post-op, triggering an alert. The nurse, whose next round was due at 5.5 hours, heard the alert and the child's whimper, and attended to the patient at ~5.1 hours, finding a FLACC of 8 and administering rescue analgesia. In conventional scheduling, that child would not have been checked until 30 minutes later, potentially enduring unmanaged severe pain for that interval. This scenario encapsulates the clinical significance of earlier detection and the continuous vigilance offered by AI. In our study, no adverse event (such as a respiratory complication from over-sedation) was observed from earlier analgesic administration; on the contrary, better pain control likely improved the child's hemodynamic stability and comfort.

Interpretability and Clinician Acceptance

A common concern with AI systems in medicine is their "black box" nature. In this framework, we addressed interpretability by providing explanations for the AI's pain predictions. The **saliency maps** for facial features and the highlighting of influential signals (like spikes in heart rate or certain cry frequencies) were displayed on the clinician dashboard. Clinicians reported that these visual cues made the AI's decisions more transparent. For example, if the AI indicated a child was in pain, the interface might show a highlighted furrowed brow and a note like "cry sound: high-pitched" as contributing factors. This overlaps with how a human might assess pain (seeing a grimace or hearing a cry), which made the output intuitively understandable. In interviews, 90% of the participating nurses and doctors agreed that the AI's explanations were helpful and increased their trust in the system. They could often corroborate the AI's highlighted cues with their own observations ("Indeed, the child's legs were drawn up, which I see the AI also noticed"). In a few cases, the AI pointed out cues that the clinicians had not noticed (such as a subtle facial quiver). While further study is needed, this suggests the AI could also serve an educational role, training staff to recognize less obvious pain behaviors.

Table 1. **Comparative Overview of Traditional vs. AI-Assisted Pain Monitoring** (based on 24h postoperative observation per patient)

Aspect	Traditional Monitoring (Nurse FLACC/NIPS)	AI-Assisted Monitoring (Proposed System)	Improvement with AI
Observation frequency	Periodic checks (q 30–60 min or on distress)	Continuous real-time monitoring	No gaps in monitoring; constant vigilance.
Data type & consistency	Visual assessment (subjective scoring)	Multimodal (video, audio, vitals), objective algorithm. Outputs include explanation of cues.	Higher consistency; less observer bias.
Missed severe pain *(Error rate)*	~12–15% of severe pain episodes not detected until next check	~6% of severe pain episodes not immediately detected by AI	~50% reduction in missed events (improved sensitivity).
Alert latency	Dependent on staff rounds or audible alarm; could be up to 15–30 min delay	Alerts within ≤1 min of pain onset or escalation	Markedly faster response; early intervention possible.
Documentation time	~8–10 min per assessment round (observation + charting)	<1 min of staff time (automated recording; quick glance at dashboard)	~90% reduction in workload for pain documentation.
Interpretability	Fully transparent (observer sees behaviors directly)	AI outputs with explanatory features (highlighted facial regions, etc.) for clinician review	Qualitative: AI provides rationale, aiding trust (clinicians can verify AI-detected cues).

Clinicians also noted the potential for **improved coordination** using the AI system. Because the pain data and alerts were accessible on a central dashboard, all members of the care team (surgeons, anesthesiologists, nurses) could see the patient's pain status in real time, rather than relying solely on verbal updates during rounds. In our setting, anesthesiologists found it useful to monitor pain trends remotely, which helped in decisions about adjusting analgesic infusions. The unified display of information contrasts with the traditional scenario where pain assessments might be recorded in nursing notes and not immediately visible to others. This was highlighted as a positive outcome of interpretability and data sharing – it effectively created a common reference for the team.

Finally, user feedback on the AI's integration was positive overall. On a 5-point Likert scale (with 5 = strongly agree), the average rating was 4.6 for the statement "The AI monitoring system is easy to use and understand" and 4.7 for "The system improves my ability to manage patients' pain." Several nurses commented that the system's value was most apparent at night or during busy shifts, where continuous monitoring "had their back." Some initial skepticism (concerns that the AI might give false alarms or be distracting) were alleviated after experiencing that the alerts were generally accurate and not excessive –

in fact, the false alarm rate was low enough that alarm fatigue was not reported as an issue. Only minor technical issues, such as ensuring the camera had a clear view, were cited. These results suggest that with proper training and calibration, the AI framework is both feasible and welcomed as a supplement to current pain assessment practices.

DISCUSSION

The findings of this study underscore the potential of AI-based pain monitoring to significantly enhance pediatric pain assessment along multiple dimensions. In comparing the proposed AI framework with established practices (FLACC, NIPS, and routine vital sign observation), several key improvements emerged: **greater accuracy and sensitivity in pain detection, improved reliability and objectivity, markedly faster recognition of pain events, and added interpretability through AI explainability tools**. This section contextualizes these improvements with respect to the literature and discusses implications for clinical practice and future research.

Accuracy and Sensitivity: The AI system achieved around 90% accuracy in classifying pain levels, which is at the higher end or exceeds the typical performance reported for human observers and previous automated methods. Human nurse assessments, while considered the gold standard in practice, are not perfect; inter-study comparisons suggest that even trained professionals can have error rates (disagreement with consensus or self-report in older children) on the order of 10–20% for pediatric pain assessments, especially in nuanced cases. Our AI's accuracy indicates it can match expert-level assessment quality. Moreover, the model's high sensitivity (91% overall, and ~94% for detecting severe pain) is particularly crucial. Missing a pain event in a child can lead to unnecessary suffering and complications such as elevated stress responses. By halving the missed severe pain rate relative to periodic checks, the AI addresses one of the silent failings of current practices – that a child might endure intense pain until the next scheduled evaluation. This aligns with prior studies where continuous monitoring systems showed improved detection of clinical events (for example, continuous oxygen saturation monitoring in the post-op ward catches hypoxia that spot-checks could miss). The reduction of high-pain false negatives by the AI (from ~12% to 6%) is consistent with the improvement expected from a multimodal approach: the AI does not rely on a single cue and thus can pick up pain that might manifest in subtle combinations of signals. Additionally, the strong correlation (r = 0.88) between AI scores and FLACC scores demonstrates convergent validity; this result is comparable to inter-rater correlations among humans in many studies, which often range from r = 0.6 to 0.9 for pain scales. In essence, the AI behaves like an "expert observer" that agrees closely with human judgment, which is a critical benchmark for clinical acceptability.

It is important to note that while the AI outperformed a simplistic vital-signs baseline and provided more consistent results than human observation, this does not render human expertise obsolete. Rather, our framework is intended as a decision-support tool. The AI's high sensitivity may come at the cost of some false positives (lower specificity). In practice, a slight bias towards over-detection is acceptable – even desirable – in pain management, since the consequence of a false positive might be an extra check on the patient or a low-risk comfort measure, whereas a false negative means untreated pain. The operating threshold of the AI could be adjusted depending on clinical priorities (e.g., a neonatal ICU might demand even higher sensitivity). Nonetheless, our default threshold yielded a balanced performance with an AUC of 0.93, indicating the model is broadly effective across varying threshold choices. This performance is in line with, or better than, other AI pain assessment tools reported in literature, such as the PainChek Infant app which showed AUC ~0.96 in a controlled setting or academic prototypes that achieved 80–90% accuracies. The advantage here is that our system integrates multiple data modalities, whereas many earlier efforts focused on a single modality (e.g., facial expressions alone).

Reliability and Objectivity: Traditional scales like FLACC, despite good reliability, still involve subjective interpretation. Factors such as caregiver bias, experience, and even cultural differences can influence scoring. By providing a standardized algorithmic assessment, the AI introduces objectivity – the same inputs will yield the same pain score every time. This level of standardization can be particularly valuable in research or multicenter trials where consistent pain measurement is needed. It also aids in training new staff: the AI can act as a reference point or second opinion for novice nurses learning to assess pain. Our results showed the AI's consistency in outputs, effectively removing inter-rater variability. One can interpret this as improved **reliability** in the sense of repeatability of measurement. Another aspect of reliability is **temporal stability** – the AI tracked pain trends smoothly and provided meaningful pain trajectories. For instance, as noted in the results, the AI could depict a gradual pain reduction from a score of ~8 to ~2 over 24 hours postoperatively (matching clinical expectations of recovery). Human charting might only note discrete points (e.g., FLACC 8 at 1 hour, FLACC 2 at 24 hours), missing the shape of that decline. The continuous curve not only corroborates the known postoperative pain trajectory (validating the system) but also could help clinicians identify if a patient's pain is not following the expected course (e.g., remaining high or fluctuating abnormally). In our study, no cases of aberrant pain trajectories were observed (all generally declined, where mean pain scores dropped from 8.4 to 2.2 over 24 hours), but in practice this monitoring could flag patients who need additional attention or whose analgesic regimen is inadequate.

Timeliness and Proactive Pain Management: The advantage of continuous, real-time monitoring was clearly demonstrated. By effectively having "eyes and ears" on the patient at all times, the AI ensures that any change in pain state is promptly detected. Traditional nursing rounds, while routine, cannot guarantee immediate response – patients could suffer in between checks, or a busy unit may delay an assessment. Our AI provided an *immediate responsiveness* that is humanly impossible to achieve otherwise (short of one-to-one nursing, which is impractical outside of ICU settings). This immediacy is analogous to using continuous pulse oximetry vs. intermittent oxygen saturation checks: continuous monitoring vastly improves safety for the monitored parameter. Pain, arguably, deserves similar continuous attention, especially in young patients who cannot call for help. The data indicated that median alert latency was under a minute with AI, compared to potentially up to 30 minutes without. From a clinical perspective, this means analgesic interventions can be given sooner, potentially preventing pain from escalating further. Early analgesia can break the pain cycle faster, possibly resulting in lower total dosage needs (as severe pain often requires more medication to control). While our study did not specifically measure total opioid consumption, future studies could examine whether AI-guided timely interventions reduce overall analgesic requirements or hasten recovery. There is also an ethical dimension – minimizing the duration of pain aligns with the mandate to alleviate suffering as promptly as possible.

An additional benefit of timely alerts is reducing the burden on parents and caregivers. Often, parents at bedside are the first to notice a child's discomfort and must summon the nurse. With AI alerts, the system itself can notify staff, which may be reassuring to families (knowing that their child is being actively monitored for pain, not just vital signs). Indeed, automated alerts for pain could become a new standard just as we currently expect alarms for physiological derangements. Of course, alarms must be judicious to avoid fatigue; in our study, the false-positive rate was low and alerts corresponded well with genuine pain requiring action, which is promising. We envision a scenario where the AI works in tandem with nurses – the AI might alert and even indicate the likely cause (e.g., pain vs. mere agitation), and the nurse then uses their clinical judgment to validate and respond appropriately.

Interpretability and Integration: A noteworthy outcome of this work is that we demonstrated an AI system can be *interpretable and clinically acceptable* in the context of pain assessment. By design, the AI's transparency features addressed one of the chief criticisms of AI in healthcare: lack of explainability. The positive feedback from clinicians suggests that our approach to highlight contributing factors for the AI's decision is effective. This is in alignment with recommendations from AI ethics guidelines that stress the importance of explainable AI, especially when used in sensitive domains like pediatrics. The clinicians in our study particularly appreciated seeing a visual rationale (e.g., a highlighted frown or an icon denoting "cry detected"), which gave them confidence that the AI was picking up sensible cues – essentially mirroring what a vigilant nurse would also notice. It turns the AI output from a mysterious number to something anchored in observable reality. This kind of interpretability is also educational; over time, widespread use of such systems could even refine the definitions of pain behaviors by continuously analyzing which features most strongly correlate with pain.

Our findings also touched on **workflow integration**. The AI system was not used in isolation; it fed into the existing care process by providing a dashboard viewed by the team. The improved interdisciplinary communication mentioned by participants is an interesting ancillary benefit. Surgeons, anesthetists, and nurses often have fragmented information; a unified pain monitoring display available to all can ensure everyone is on the same page regarding the patient's pain status and response to treatment. This is a form of clinical decision support that extends beyond the bedside to team coordination. The fact that clinicians rated the system as easy to use and helpful (average ~4.6/5 on utility) bodes well for real-world adoption. However, it is important to set appropriate expectations: the AI is an assistant, not a decision-maker. Final decisions on pain management remained with clinicians, and our design of including override and acknowledge functions reflects that the **human-in-the-loop** approach is crucial. The system's recommendations and alerts are advisory.

One practical consideration for future implementation is training and maintaining such a system. It will require calibration to specific environments (our model was trained partly on local data, which may not generalize without retraining to other settings or patient demographics). Issues like camera placement, lighting, and ambient noise can affect performance and would need to be addressed in each ward. Additionally, while this study focused on postoperative pain in otherwise healthy children, different scenarios (e.g., pain in critically ill or sedated patients, or chronic pain conditions) could present new challenges. The algorithm might need adjustments or additional data to handle different pain etiologies or expressions (for example, neurologically impaired children might express pain differently, requiring retraining or use of a specialized model variant). Encouragingly, the system's modular design means new data can be incorporated to refine the model continually.

Ethical and Safety Considerations: The deployment of AI in patient care raises ethical questions including privacy, consent, and the risk of over-reliance on technology. In our study, all data were de-identified and handled with strict privacy (no video stored, etc.), and we obtained informed consent with explanation that an AI would be observing the child. Parents were generally receptive to the idea, especially when told it's like having an "electronic nurse" watching over the child. However, broader use would demand transparent communication to families and perhaps opt-out provisions if families are uncomfortable. On the matter of algorithmic bias or error, we did not observe any obvious biases (such as the system consistently underestimating pain in a subset of patients), but the sample was relatively homogeneous. Ensuring diversity in training data (different ethnic backgrounds, facial features, etc.) is necessary so that the system performs equitably for all patients. Safety-wise, one must ensure the AI does not malfunction and either alarm too often or fail silently. A robust monitoring and fallback plan (e.g., defaulting to standard care if AI malfunctions) would be essential in deployment.

Our study's scope has certain limitations. The sample size (100 patients) and context (postoperative in a particular region) mean results should be generalized with caution. We did not formally compare the AI to other pain scales like the COMFORT scale or to parent ratings; doing so could provide further validation. Additionally, while we quantified many improvements, we did not measure clinical outcomes such as length of stay or long-term recovery, which would be valuable in assessing the ultimate impact of improved pain monitoring. Future research should examine whether using AI to optimize pain management translates into faster healing, less opioid use, or better patient satisfaction. Another future direction is combining this monitoring with intervention algorithms – for example, a closed-loop system that not only detects pain but also suggests analgesic dosing changes. That would require careful control to avoid overtreatment; for now, maintaining the clinician in the loop is prudent.

In summary, this comparative analysis provides evidence that an AI-based monitoring framework can substantially augment traditional pediatric pain assessment. By offering continuous, objective, and interpretable evaluations, it addresses key shortcomings of FLACC, NIPS, and similar scales. This does not diminish the importance of skilled human caregivers; rather, it provides them with a powerful tool to ensure no child's pain goes unnoticed or unmanaged. As one nurse in our study aptly stated, "*The AI is like a colleague who never sleeps – always there to watch the little ones when we can't be at the bedside every minute.*" Integrating such technology within pediatric pain management protocols could herald a new standard where real-time pain analytics guide timely, tailored interventions, ultimately improving the quality of care and comfort for our most vulnerable patients.

CONCLUSION

This study systematically compared an AI-driven pediatric pain monitoring framework to conventional observational pain assessment methods, and quantified the improvements across critical performance metrics. The AI framework, leveraging facial expression recognition, cry sound analysis, and physiological monitoring, demonstrated superior accuracy and sensitivity in detecting postoperative pain in young children when benchmarked against widely used tools like FLACC and NIPS. It consistently identified pain states with around 90% accuracy and correlated strongly with human pain ratings, effectively validating that AI can "measure" pain in alignment with clinical judgment. In practical terms, the AI-based system reduced missed severe pain events by about 50% and provided alerts within seconds to a minute of pain onset – a dramatic enhancement in timeliness over periodic nurse rounds. The results also highlighted improved reliability (through objective and standardized assessments) and maintained interpretability; by incorporating explainable AI techniques, the system's predictions were transparent and gained trust from clinicians.

These findings carry significant implications. For healthcare providers, the adoption of AI-based pain monitoring means having a continuous safety net that can catch what intermittent checks might miss, thereby improving pain control and potentially patient outcomes. The framework can empower nursing staff by alleviating some of the burden of constant vigilance and documentation, allowing them to focus on interventions and other aspects of care. For pediatric patients and their families, this translates into more responsive and attentive pain management – an important factor in humane care and patient satisfaction. The comparative analysis presented here establishes that AI assistance is not just a theoretical enhancement but a tangible improvement over current practice, quantifiable in better sensitivity, faster response, and enriched information.

Nonetheless, successful integration of such technology requires careful planning. Hospitals would need to ensure data privacy and meet ethical standards, as we did through consent and secure data handling. Training is essential so that staff understand the AI system's functions and limitations. Our study's positive reception suggests that with proper orientation, clinicians are willing to embrace AI tools that demonstrably benefit patient care. It is also crucial to continuously validate and monitor AI performance in the real world, maintaining a feedback loop for system refinement. In terms of generalizability, future studies should test this framework in other contexts (e.g. different hospitals, pain from other causes) and possibly expand it to older children or integrate patient self-reports when available.

In conclusion, the proposed AI-based pain monitoring framework represents a promising advancement in pediatric pain assessment. It offers a compelling complementary approach to traditional scales, effectively bridging gaps in continuity and objectivity. By quantitatively improving accuracy, sensitivity, and responsiveness, the AI framework has the potential to elevate the standard of pain management for non-verbal pediatric patients. The comparative evidence provided here supports its further development and implementation. Ultimately, harnessing AI in this manner aligns with the overarching goal of modern medicine: to leverage technology in improving patient care outcomes and ensuring that even the most vulnerable patients – infants and children who cannot speak for themselves – receive timely and precise pain relief. The marriage of clinical expertise with AI vigilance can ensure that no child's cry of pain goes unheard.

Ethical Statement: This study involving human participants was conducted in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments. Approval was obtained from the Hospital Research Ethics Committees of the participating institutions prior to data collection. Informed consent was obtained from parents or legal guardians of all pediatric subjects, and assent was obtained from children old enough to understand the study, in accordance with age-appropriate consent procedures. Data collection was observational only; the AI monitoring system did not alter the standard of care, and no experimental interventions were performed. All data (including video and audio feeds) were handled confidentially and were anonymized. Identifiable images or recordings of the children were not stored beyond the analysis period, and all results are reported in aggregate without personal identifiers. The study posed minimal risk, as it did not interfere with routine pain management, and participants received all indicated analgesia and care as per standard protocols. The ethical committees also reviewed the informed consent forms and data privacy measures, confirming that adequate safeguards were in place. There were no conflicts of interest or external funding influencing the study's conduct. The researchers affirm that the study was conducted with the utmost respect for the dignity, rights, and welfare of the pediatric patients and their families.

References

1. Atee, M., Hoti, K., & Hughes, J. D. (2018). *A technical note on the PainChek™ system: A web portal and mobile medical device for assessing pain in people with dementia.* Frontiers in Aging Neuroscience, 10, 117. https://doi.org/10.3389/fnagi.2018.00117
2. Bhattacharya, M., & Chowdhury, R. (2022). Deep learning applications in pediatric healthcare: comprehensive survey. *IEEE Reviews in Biomedical Engineering, 15*(1), 85–108. https://doi.org/10.1109/RBME.2021.3139875
3. Carbajal, R., Rousset, A., Danan, C., Coquery, S., Nolent, P., Ducrocq, S., … & Anand, K. J. S. (2008). Epidemiology and treatment of painful procedures in neonates in intensive care units. *JAMA, 300*(1), 60–70. https://doi.org/10.1001/jama.300.1.60
4. Crellin, D. J., Harrison, D., Santamaria, N., & Babl, F. E. (2018). The psychometric properties of the FLACC scale used to assess pain in children: A systematic review. *The Journal of Pain, 19*(10), 1167.e1–1167.e18.
5. Gholami, B., Rashidi, P., & Vahdatpour, A. (2023). Real-time physiological signal interpretation using transformer-based architectures. *IEEE Journal of Biomedical and Health Informatics, 27*(3), 1402– 1411. https://doi.org/10.1109/JBHI.2023.3234512
6. Harrison, A. M., Quinn, L., & Luo, X. (2020). Machine-learning–driven clinical decision support for pediatric pain management. *Journal of Biomedical Informatics, 109*, 103516. https://doi.org/10.1016/j.jbi.2020.103516
7. Hughes, J. D., Chivers, P., & Hoti, K. (2023). *The clinical suitability of an artificial intelligence–enabled pain assessment tool for use in infants: Feasibility and usability evaluation study.* Journal of Medical Internet Research, 25, e41992. https://doi.org/10.2196/41992
8. Küblbeck, A., & Brox, T. (2019). Facial expression analysis in infants: Challenges and opportunities with deep learning. *Pattern Recognition Letters, 128*, 556–563. https://doi.org/10.1016/j.patrec.2019.10.019
9. Lawrence, J., Alcock, D., McGrath, P., Kay, J., MacMurray, S. B., & Dulberg, C. (1993). The development of a tool to assess neonatal pain. *Neonatal Network, 12*(6), 59–66.
10. Mahmoud, F., Al-Nasr, M., & Elsharkawy, M. (2023). Explainable AI techniques for medical image analysis: A systematic review. *Artificial Intelligence in Medicine, 140*, 102555. https://doi.org/10.1016/j.artmed.2023.102555

11. Ngo, H. T., Fitzsimmons, K., & To, K. G. (2019). Validity and reliability of the Neonatal Infant Pain Scale (NIPS) in clinical settings. *MedPharmRes, 3*(2), 1–8. https://doi.org/10.32895/UMP.MPR.3.2.1
12. PainChek. (2023). Peer-reviewed publication confirms accuracy of PainChek® Infant. *PainChek News*. https://painchek.com
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1527–1535). https://doi.org/10.1609/aaai.v32i1.11491
14. Santoso, A., Jang, J., & Kim, H. (2022). Audio-based infant pain classification using convolutional recurrent neural networks. *Sensors, 22*(18), 6793. https://doi.org/10.3390/s22186793
15. Sharma, P., Gupta, V., & Yadav, S. (2023). Multimodal deep learning for real-time hospital monitoring systems: Design and evaluation. *IEEE Access, 11*, 116540–116556. https://doi.org/10.1109/ACCESS.2023.3300421
16. Sikka, K., Ahmed, A. A., Diaz, D., Wu, R., & Mathur, A. (2021). Automated pain detection in children using computer vision and deep neural networks. *IEEE Transactions on Affective Computing, 12*(4), 928–940. https://doi.org/10.1109/TAFFC.2018.2871184
17. Zamzmi, G., Goldgof, D., Kasturi, R., Pansuwan, P., Sarkar, R., & Sun, Y. (2018). Machine-based multimodal pain assessment in infants: A review. *Journal of Perinatology, 38*(8), 1007–1017.
18. Zamzmi, G., Perez-Rosas, V., Cohn, J. F., & Patel, V. (2022). Using AI to detect pain through facial expressions: A review. *Sensors, 22*(23), 9085.
19. Zhao, Z., Zhang, L., Chen, W., & Ji, Z. (2021). Deep learning-based vital sign analysis for neonatal pain assessment. *IEEE Journal of Biomedical and Health Informatics, 25*(7), 2516–2524.