

Review

A Review of Deep Learning Model Approach for Pain Assessment in Infant Cry Sounds

Anthony McCofie ^{1,*}, Dmitry Goldgof ^{1,*}, Jacqueline Hausmann ¹, Peter R. Mouton ², Yu Sun ¹
and Md Imran Hossain ¹

¹ Bellini College of Artificial Intelligence, Cybersecurity, and Computing, University of South Florida, 4202 E Fowler Ave, Tampa, FL 33620, USA

² SRC Biosciences, 11810 W Kennedy Blvd UNIT 310, Tampa, FL 33606, USA

* Correspondence: anthonymccofie@usf.edu (A.M.); goldgof@usf.edu (D.G.)

Abstract

Infant cries serve as a primary indicator of distress and pain; however, distinguishing pain-related cries from those triggered by other needs remains a challenging task, even for trained professionals. Timely and accurate pain assessment is essential for appropriate medical intervention, particularly in preverbal infants who cannot express their needs verbally. Recently, Deep Learning (DL) models have demonstrated significant potential in addressing this challenge by enabling automated and efficient pain assessment through audio signal processing. In this survey, we review methods for pain assessment from infant cry sounds, covering deep learning architectures, modern Transformer-based models, and emerging Vision-Language Model (VLM) pipelines. The review includes approaches that integrate Mel-spectrogram representations of cry audio with multimodal model frameworks to improve robustness, interpretability, and cross-modal reasoning in pain detection. By summarizing recent advancements and identifying limitations and open challenges in current methodologies, this review aims to provide insights into future research directions that may enhance the robustness, generalizability, and clinical applicability of automated infant pain assessment tools.

Keywords: infant cry analysis; neonatal pain assessment; deep learning; audio signal processing; cry classification; pain indicators

1. Introduction

1.1. Background and Clinical Importance

Pain assessment in infants remains a significant clinical challenge because infants cannot verbally communicate essential aspects of pain, including its location, duration, and intensity. Accurate and timely recognition of pain is especially important in neonatal and pediatric care, where untreated or poorly managed pain may negatively affect physiological stability, neurodevelopment, and overall clinical outcomes. As a result, reliable pain assessment is fundamental to guiding appropriate interventions and improving the quality of care delivered to this vulnerable population [1].

1.2. Limitations of Traditional Pain Assessment Methods

In current clinical practice, infant pain is commonly evaluated using observational and physiological methods, including behavioral pain scales, facial expression analysis, body movement assessment, and vital-sign monitoring. Although these approaches are



Academic Editor: Krzysztof J. Cios

Received: 30 January 2026

Revised: 11 March 2026

Accepted: 12 March 2026

Published: 19 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

widely used and clinically valuable, they are inherently limited by subjectivity and inconsistency. Their effectiveness often depends on the experience, training, and judgment of healthcare professionals, and their interpretation may be influenced by fatigue, workload, and time constraints [1–3]. In addition, physiological indicators are often nonspecific and may reflect stress, illness, or environmental disturbance rather than pain alone. These limitations highlight the need for more objective, consistent, and scalable approaches to infant pain assessment.

1.3. Rationale for Deep Learning-Based Cry Analysis

Among the available indicators of infant distress, crying is one of the earliest and most prominent acoustic expressions of discomfort. Infant cry signals contain rich temporal, spectral, and prosodic information that may reflect underlying physiological and emotional states, making them a promising target for automated analysis. Advances in digital signal processing and artificial intelligence have created new opportunities to extract clinically meaningful patterns from cry audio. In particular, deep learning has emerged as a powerful framework for modeling complex, high-dimensional acoustic data and for learning discriminative representations directly from raw signals or transformed audio features such as Mel-Frequency Cepstral Coefficient (MFCC), spectrograms, and other time–frequency representations [4–6]. Models such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), convolution-based architectures, Transformer-based, and more recently VLM approaches have been increasingly explored for infant cry analysis and pain-related classification tasks [7,8].

1.4. Challenges in Cry-Based Pain Detection

Cry-based pain detection also presents important clinical challenges. Infant crying is not specific to pain and may also be triggered by hunger, discomfort, fear, fatigue, or medical conditions, making it difficult to distinguish pain-related crying from other forms of distress. Cry signals are also affected by age, gestational maturity, recording conditions, background noise, inter-subject variability, and differences in annotation protocols and dataset quality [9,10]. Furthermore, many existing studies rely on relatively small or imbalanced datasets, use different preprocessing pipelines and evaluation metrics, and often lack external validation, which limits reproducibility and generalizability. These issues make it necessary to critically examine not only the strengths of deep learning methods, but also the practical limitations and translational barriers associated with cry-based pain assessment systems [11].

1.5. Objectives, Scope, and Novelty of This Review

Given the rapid growth in this field, a focused review is needed to synthesize the emerging body of work on deep learning approaches for pain assessment in infant cry sounds. The objective of this review is to examine current research on deep learning-based infant pain detection from cry audio, with attention to commonly used datasets, signal representations, model architectures, and reported challenges. The scope of this review is specifically centered on automated pain assessment using infant cry sounds, while also acknowledging related developments in multimodal neonatal monitoring where relevant. By concentrating on cry-based pain assessment rather than general infant cry classification, this review aims to clarify the state of the field from both technical and clinical perspectives. The contribution and novelty of this review lie in its focus on deep learning model approaches for pain assessment in infant cry sounds, rather than broader infant cry classification or general biomedical audio processing tasks. In contrast to prior reviews that often address neonatal monitoring at a broad level or combine multiple infant conditions without specific emphasis on pain, this review specifically evaluates the methodological

landscape of cry-based pain detection, highlights the unique challenges associated with using cry as a pain biomarker, and identifies key gaps that must be addressed for clinical translation. By bringing together evidence across model design, feature representation, dataset limitations, and application constraints, this review provides a structured foundation for researchers and clinicians seeking to advance more reliable and objective tools for infant pain assessment.

1.6. Overview of the Paper

The structure of the survey paper is as follows: Section 2 presents the methodology adopted for the review, including the paper selection and organization process. Section 3 discusses infant pain and the traditional pain assessment tools developed to evaluate pain and sedation levels in both preterm and full-term neonates, which are still used in some NICUs. Section 4 reviews commonly used cry audio feature extraction techniques for infant pain assessment, including traditional time–frequency representations, deep learning-based learned features, and emerging multimodal and vision–language model approaches. Section 5 reviews datasets commonly used for infant pain assessment and examines the challenges associated with data collection. Section 6 integrates the major challenges, limitations, and future directions in infant cry-based pain assessment. The paper concludes with Section 7.

2. Methodology

The review was conducted to examine existing research on deep learning model approaches for pain assessment in infant cry sounds. The primary objective was to identify the major computational methods that have been applied to this problem, and acoustic representations used, and compare the evaluation strategies and performance trends reported in the literature. Because the paper is intended as a narrative review rather than a systematic review, the methodology also emphasizes a structured approach to literature identification, selection, and synthesis. A PRISMA-style flow diagram is included to provide a transparent summary of the literature identification, screening, and final study count used in this review. Its inclusion is intended solely to improve reporting clarity and does not indicate that this work is presented as a formal systematic review.

2.1. Review Scope and Research Focus

The review focuses on studies that investigate automated pain assessment in infants using cry audio signals, with particular attention to deep learning-based techniques. The scope includes research involving neural network architectures such as Convolutional Neural Network (CNN), RNN, LSTM networks, hybrid deep learning frameworks, transfer learning methods, and more recent transformer- or multimodal-based approaches when relevant. In addition to model architectures, the review considers commonly used input representations, including spectrograms, Mel-spectrograms, and MFCC, as well as the datasets and evaluation metrics used to assess performance.

2.2. Literature Search Method

Relevant studies were identified through a targeted search of major scholarly sources. The search was guided by combinations of keywords related to infant crying, pain assessment, deep learning, audio classification, and neonatal pain detection. Representative search terms included infant cry, pain assessment, pain detection, deep learning, neonatal cry analysis, audio classification, and related variations. Additional papers were also identified through backward reference checking of selected articles to capture influential and closely related studies. The search process was designed to collect representative and technically relevant literature rather than to provide an exhaustive catalog of all published work.

Priority was given to studies that directly addressed automated pain recognition from infant cry signals and that presented clear methodological or experimental contributions.

2.3. Study Selection Basis

Studies were selected based on their relevance to the central theme of the review. Papers were considered for inclusion if they examined infant cry sounds in the context of pain assessment and employed deep learning methods as a primary analytical approach. Preference was given to studies that provided sufficient detail on model design, input features, datasets, and evaluation outcomes. Works focused solely on traditional machine learning methods, non-audio pain indicators, or unrelated infant vocalization tasks were considered only when useful for contextual comparison and background discussion. Because this is a narrative review, formal inclusion and exclusion criteria were not applied in the same rigid manner as in systematic reviews. Instead, selection was guided by relevance, methodological clarity, and contribution to understanding current trends and challenges in the field. The keywords selected for this review paper were used as search criteria in the Google Scholar search engine. Although this manuscript is presented as a narrative review rather than a formal systematic review, we retained a PRISMA-style flow diagram to improve transparency in reporting the literature identification and screening process. The purpose of the diagram is not to imply full PRISMA compliance or to classify the manuscript as a systematic review. Instead, it is included as a clear visual summary of the number of records identified, screened, excluded, and ultimately discussed in the review. Because the paper synthesizes prior studies in a structured manner, the flow diagram helps readers and reviewers understand how the literature set was formed and provides a concise account of the study count underlying the discussion. We believe its inclusion strengthens methodological clarity and reproducibility of the review process, even within the context of a non-systematic review. The references listed at the end of the paper correspond to the 95 studies included in the review process shown in Figure 1. We have also included the PRISMA 2020 Checklist as Supplementary Materials.

2.4. Data Extraction

For each selected study for the review, key information was examined and organized according to core dimensions such as model architecture, type of input representation, dataset characteristics, classification objective, and performance metrics. The reviewed studies were then grouped to support comparative analysis. The literature was group according to the types of deep learning models used, the acoustic features or signal representations adopted, and the evaluation practices used to measure effectiveness.

This organization enabled the review to highlight recurring methodological patterns, identify common challenges such as limited dataset size and class imbalance, and discuss emerging directions for improving robustness and clinical applicability.

2.5. Limitations of the Review Approach

As a non-systematic review, this paper does not claim exhaustive coverage of all available studies on infant cry-based pain assessment. Although care was taken to identify relevant and representative literature, some publications may not have been included due to differences in terminology, indexing, or publication source. In addition, variations in datasets, experimental setups, and evaluation metrics across studies may limit direct comparability. Nevertheless, the review provides a focused and structured summary of the most relevant deep learning approaches and offers insight into current progress, research gaps, and future opportunities in this area.

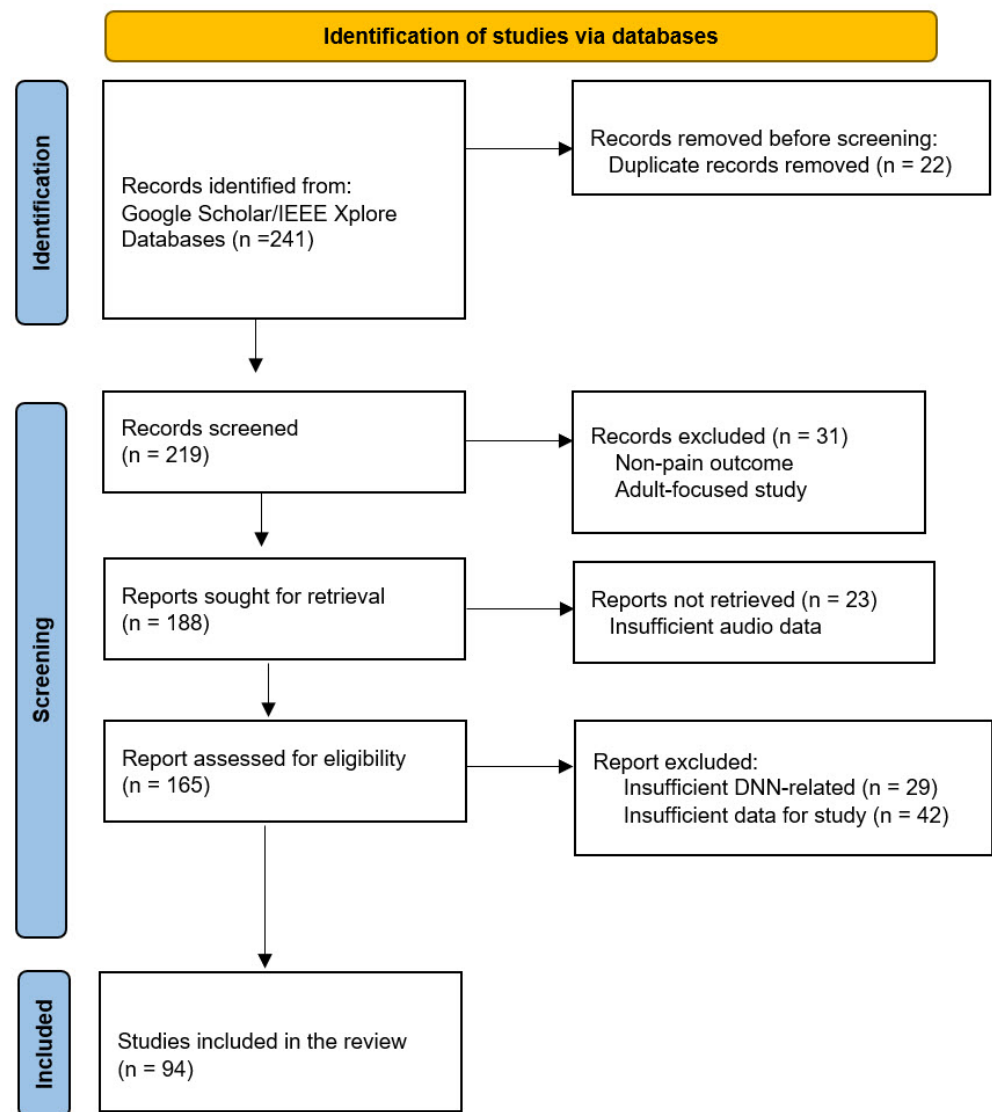


Figure 1. PRISMA-style flow diagram of literature identification and screening used in this review.

3. Infant Pain Analysis

The assessment of pain in crying infants is a vital research field focused on detecting and understanding the pain experienced by newborns, who are unable to communicate their discomfort verbally. Conventional approaches are often subjective and differ between caregivers. In this section, we review some of the commonly used methods.

3.1. Clinical and Behavioral Pain Assessment Methods (Ground Truth Formation)

Conventional approaches to assessing pain in infants mainly depend on non-verbal cues and clinical observations. These techniques emphasize physiological, behavioral, and observational signs. Commonly used traditional methods include the following:

a. **Behavioral Observation Methods:** This approach involves observing various behaviors such as facial expressions, crying, breathing patterns, and movements of the arms and legs, as well as the overall arousal state. Each behavior is assigned a score, with higher scores indicating more intense pain. Common scales used for this purpose include the Neonatal Infant Pain Scale (NIPS), Premature Infant Pain Profile-Revised (PIPP-R), and the FLACC Scale [2,3,12–14].

b. **Physiological Measurements:** Physiological measurements provide an objective complement to behavioral pain scales, capturing respiratory responses that often accompany distress [15]. In neonates and infants, pain can trigger sympathetic activation and elevated blood pressure, along with increased respiratory rate or altered breathing regularity (e.g., rapid, shallow, or periodic breathing) [16]. Clinical studies have demonstrated that all vital signs—including heart rate, respiratory rate, SpO₂ (oxygen saturation), and both diastolic and systolic blood pressure—can be significantly altered during painful procedures [16–18]. Continuous bedside monitoring enables detection of short-lived fluctuations and supports tracking whether physiological parameters normalize after comfort measures or analgesic interventions [19]. Common physiological measures used alongside cry audio includes:

- **Heart Rate Fluctuations (HR):** Pain activates the sympathetic nervous system, causing a measurable increase in heart rate. A concurrent rise in HR with high-intensity crying or elevated spectral frequencies provides strong confirmation of pain-related stress. Research demonstrates that heart rate changes are among the most sensitive physiological indicators of pain in neonates. During painful procedures, infants experience significant increases in heart rate that can be precisely measured and documented [19].
- **Oxygen Saturation (SpO₂):** Painful episodes may induce transient drops in oxygen saturation due to altered breathing patterns and excessive expiratory effort during crying. Monitoring SpO₂ helps confirm whether extreme cry features align with physiological compromise. Oxygen saturation levels, measured through pulse oximetry (SpO₂), represent a critical physiological parameter that reflects an infant's response to painful stimuli. Research has documented that minimum oxygen saturation was significantly lower during painful procedures compared to less invasive alternatives ($83 \pm 12\%$ vs. $86 \pm 10\%$, $p = 0.035$) [19].
- **Blood Pressure Changes (BP):** Both systolic and diastolic blood pressure tend to increase during nociceptive stimulation. When BP elevations occur simultaneously with high-pitched, strained cries, the confidence in pain detection increases. Clinical investigations have demonstrated that painful procedures cause statistically significant alterations in blood pressure parameters [16].
- **Respiratory Rate (RR):** Crying related to pain is often associated with rapid, shallow, or irregular respiration. RR serves as a key correlational indicator between vocal distress and physiological instability. Alterations in respiratory patterns provide another key physiological marker of neonatal distress and pain. Clinical observations reveal that painful stimuli trigger changes in respiratory rate, which can be systematically monitored and recorded [16].

c. **Multi-Modal Assessments:** The Neonatal Pain, Agitation, and Sedation Scale (N-PASS) combines multiple dimensions to assess pain and sedation. This includes crying, facial expressions, and physiological parameters such as heart rate and oxygen saturation. N-PASS is a clinical tool developed to assess pain and sedation levels in preterm and full-term neonates. It incorporates behavioral and physiological indicators, providing a comprehensive evaluation of neonatal distress. Since its inception, N-PASS has been widely adopted in Neonatal Intensive Care Unit (NICU) around the world, helping clinicians better manage pain and sedation in infants, particularly those undergoing medical procedures or in prolonged care. Its integration into clinical practice has contributed to improving neonatal care and ensuring some improved pain assessment and management. The N-PASS tool is widely used and has several advantages; however, the use of this approach is not without challenges and limitations [3,20].

Although these observer-based scales are not audio-only pain assessment systems, they provide the reference labels required for supervised machine learning models trained on infant cry audio. Most automatic pain detection systems, including deep learning and multimodal models, depend on these clinical scores to define pain versus no-pain ground truth [21].

3.2. Physiological Signal-Based Pain Validation

While acoustic analysis of infant cry signals provides important behavioral cues for detecting pain, many clinical studies strengthen their findings by incorporating physiological measures that objectively reflect the infant's internal stress response. These physiological indicators help verify whether specific acoustic markers, such as increased pitch, higher energy concentration, or irregular phonation, truly correspond to underlying nociceptive events rather than non-pain-related agitation [22].

Physiological signal-based validation is relevant since it plays a critical role as a ground-truth confirmation layer in automated infant pain detection systems [22]. While cry audio provides fast behavioral evidence, integrating physiological signals reduces false-positive classifications (e.g., distinguishing hunger cries from pain cries), Improves clinical acceptance and trust in AI-based pain assessment tools, and Enables the development of multimodal deep learning frameworks, where acoustic and physiological signals jointly inform prediction models. Establishing consistent correlations between cry acoustics and physiological stress responses is essential for translating automated pain detection technologies into real neonatal clinical settings [14].

In addition to acoustic and physiological analysis, infant pain is commonly validated using clinician-administered pain scoring systems such as the Neonatal Facial Coding System (NFCS), NIPS, the PIPP-R, and the N-PASS shown in Table 1. These scoring tools evaluate facial expression, limb movement, breathing patterns, and arousal states, and are frequently used as ground truth labels for training and validating machine learning models [3,22–24].

Table 1. Standard N-PASS scoring rubric for neonatal pain/agitation and sedation assessment. The table lists each behavioral/physiological category and the corresponding score anchors across the full range from sedation (−2: deep sedation; −1: mild sedation) to baseline (0: normal) and pain/agitation (+1: mild pain; +2: severe pain), enabling consistent bedside scoring.

Category	−2 (Sedation)	−1 (Mild Sedation)	0 (Normal)	+1 (Mild Pain)	+2 (Severe Pain)
Crying/Irritability	No cry, not arousable	Moans, minimal cry	No cry when awake	Irritable, intermittent cry	High-pitched, continuous cry
Behavior State	No response to stimulation	Minimal response	Appropriate for gestational age	Restless, difficult to calm	Constant agitation
Facial Expression	None, flat	Minimal, weak grimace	Relaxed, neutral	Intermittent grimace	Continuous grimace or clenched jaw
Extremities/Muscle Tone	Flaccid	Decreased tone	Normal tone	Increased tone, flexion	Rigid, clenched fists/toes
Vital Signs (HR, RR, BP, O ₂ sat)	Decrease >20% from baseline	Decrease 10–20%	Stable, baseline	Increase 10–20%	Increase >20% or frequent desaturations

3.3. Acoustic Characteristics of Cry Signals

Acoustic characteristics of cry signals provide measurable sound properties of an infant's cry that describe its structure and how it is perceived in time and frequency domains. Research demonstrates that acoustic features such as fundamental frequency, intensity, duration, and spectral properties show significant associations with infant distress levels and pain states [22]. The cries of infants experiencing pain exhibit distinctive acoustic patterns that differ measurably from hunger cries and fussiness, with machine learning algorithms achieving high accuracy in distinguishing these states [25]. Notably, pain-related

acoustic features outrank intensity measures alone, highlighting the importance of spectral and temporal characteristics as reliable biomarkers [25]. Pain alters the acoustic structure of infant cries in a consistent, quantifiable way that machine learning models can learn from. In newborn pain assessment, these acoustic characteristics form the core “biomarkers” that can be objectively quantified and used to discriminate painful from non-painful cries. These biomarkers play a vital role in training deep learning models for infant pain detection from cry audio [4]. The physiological changes that manifest as distinct and quantifiable acoustic alterations in the cry signal include the following:

- **Higher Pitch (Fundamental Frequency, F_0):** The fundamental frequency (F_0) of an infant’s cry represents the rate at which the vocal folds vibrate during vocalization. Research demonstrates that cry acoustics are strongly associated with neurophysiological and behavioral signals, with fundamental frequency being one of the most extracted and analyzed features depending on infants’ arousal state [22]. When infants experience heightened distress or pain, increased muscle tension in the laryngeal region causes the vocal folds to vibrate more rapidly, directly resulting in higher-pitched vocalizations. This physiological response reflects the infant’s level of arousal and emotional state, making F_0 a critical biomarker for assessing infant distress and health status [22]. Pain cries consistently exhibit significantly higher F_0 compared to hunger or discomfort cries.
- **Increased Loudness (Amplitude/Intensity):** Stronger respiratory drive and subglottal pressure generate higher sound pressure levels, making pain cries louder and more forceful [22,26]. Investigations examining different distress levels in newborns have revealed that heightened arousal states produce significantly elevated acoustic responses across multiple parameters, including fundamental frequency, along with physiological indicators such as increased brain activity, cerebral oxygenation, heart rate, body tension, and rigidity [22].
- **More Irregular Vibration:** Instability in vocal fold control produces nonlinear acoustic phenomena such as jitter, shimmer, subharmonics, and chaotic spectral components, which are commonly observed in pain-induced cries [27]. Research investigating infant cry acoustic features as developmental markers has distinguished between smoothly-modulated cry patterns and highly unstable patterns characterized by irregular fundamental frequency trajectories [28].
- **Abrupt Temporal Changes:** Pain cries often display sudden onsets, abrupt interruptions, and short, sharp cry bursts due to irregular respiratory control and involuntary muscle contractions [22,27,29]. Research characterizing neonatal infant cry signals has documented that cries exhibit sudden variations in fundamental frequency and resonance, with these abrupt acoustic changes occurring within tiny duration solitary utterances [27].

These consistent acoustic deviations provide critical discriminative features for automated pain assessment systems [30]. Features such as fundamental frequency, energy, spectral entropy, and temporal modulation patterns are widely used in deep learning models for infant pain classification. The physiological basis of these changes strengthens the reliability of acoustic cry analysis as a non-invasive biomarker for pain detection [22,31].

Understanding how pain physiologically alters cry production helps justify the acoustic features selected for computational analysis. In feature extraction, these acoustically driven changes guide the choice of time-domain, frequency-domain, and time–frequency representations such as Mel-frequency cepstral coefficients (MFCCs), log-Mel spectrograms, and nonlinear vocal metrics. These representations capture the distinctive patterns introduced by pain, enabling machine learning and deep learning models to detect them effectively [4,25,32].

3.4. Categories of Infant Cries

In clinical and developmental research, infant cries are typically classified into distinct categories based on their underlying cause, acoustic characteristics, and behavioral context [7,27]. Understanding these categories is essential for pain-related cry detection because the acoustic structure of pain cries differs significantly from that of other cry types [25]. We provide an overview of the commonly recognized categories of infant cries and highlight their acoustic and physiological attributes relevant to automated pain-assessment systems. Studies have shown that machine learning models can achieve high accuracy (91–95.5%) in classifying these distinct cry types when trained on appropriate acoustic features [7,27].

- **Basic Cry (Hunger/Discomfort Cry):** Basic cries are the most common type observed in infants and occur in response to general needs such as hunger, thirst, mild discomfort, or the need for attention [27,33]. These cries exhibit a relatively low fundamental frequency (F_0), fairly regular phonation patterns, and rhythmic temporal structure. The cry units typically follow a consistent inspiratory/expiratory cycle, with moderate intensity and short-duration expiratory phases [29]. Because basic cries are associated with non-urgent states, they lack the abrupt onset, pitch escalation, and chaotic energy patterns characteristic of pain cries. In acoustic analysis, basic cries often show stable spectral harmonics, lower energy concentration, and minimal frequency modulation [4].
- **Pain Cry (Nociceptive Cry):** Pain cries arise from acute nociceptive stimuli such as heel lances, venipunctures, vaccinations, or clinical procedures [34]. They are the most acoustically distinct category, enabling highly accurate automated classification [4,9]. Pain cries typically begin with a sudden onset, characterized by a short inspiratory phase, followed by a long, loud expiratory signal [29]. Studies investigating newborn cry sounds have emphasized that expiratory segments represent the most informative parts of cry signals for diagnostic purposes, as these phases contain the richest acoustic information for distinguishing between different infant states [9]. Pain cries may also include hyperphonation, where F_0 rapidly escalates due to heightened subglottal pressure and increased vocal fold tension [22,32]. Temporal patterns become chaotic, and pauses between expiratory bursts can shorten dramatically [27]. From a signal-processing standpoint, pain cries frequently display broad spectral peaks, nonlinear phenomena (e.g., subharmonics, biphonation), and higher energy in high-frequency bands, making them suitable for automated detection through acoustic or deep learning techniques [4,32]. Deep learning-based cry diagnostic systems utilizing Self-Supervised Learning models (wav2vec 2.0, WavLM, HuBERT) can extract feature representations directly from raw audio signals, effectively capturing complex patterns associated with pathological conditions and achieving accuracy rates approximately 90% [9].
- **Fear Cry (Alarm or Startle Cry):** Fear cries emerge when infants experience sudden threatening stimuli, such as loud noises, startling movements, or separation anxiety. These cries share some similarities with pain cries but are distinguished by a gradual buildup of intensity rather than an abrupt onset [35]. Research investigating different distress levels in newborns has documented that infants exhibit varying arousal states—including resting, cry, and distress conditions—each associated with distinct neurophysiological and behavioral patterns [22]. Fear cries contain higher-than-normal F_0 , but with more controlled phonation compared to pain cries. They often exhibit tremulous modulation due to heightened sympathetic arousal. Mel-spectrogram representations typically show alternating regions of stable and unstable harmonic structure. Although fear cries may sound distressed, they generally lack the extreme hyperphonation and disorganized temporal patterns found in pain cries.

This distinction can be useful in automated systems to reduce false positives when classifying infant states [4,32].

- **Anger/Frustration Cry:** Anger cries occur when an infant is physically restrained, overstimulated, or prevented from achieving a goal. These cries are characterized by increased intensity, short inspiratory phases, and harsh phonation [22,28,29]. In contrast to pain cries, anger cries usually exhibit lower fundamental frequency but greater amplitude variability. Their temporal structure includes longer expiratory phases with intermittent breaks but remains more rhythmic than nociceptive cries [28]. Spectral analysis typically reveals moderate harmonic distortion, reflecting high subglottal pressure without significant nonlinearities. The acoustic signature of anger cries may overlap with pain cries in terms of loudness, which presents challenges for classifiers that rely solely on amplitude-based features [4,36].
- **Fatigue or Tired Cry:** Fatigue-related cries tend to be softer, lower-pitched, and less structured than other categories. They feature intermittent crying bouts, longer pauses, and decreasing intensity over time as the infant becomes more exhausted [27,33]. Acoustically, these cries exhibit lower energy distribution, reduced harmonic richness, and minimal frequency modulation. Their relaxed, drawn-out temporal pattern contrasts sharply with the urgency and tension embedded in pain cries [31,37]. In machine learning systems, fatigue cries often form a distinct cluster due to their low spectral complexity and smooth temporal contours [4,32].
- **Distress Cry:** Distress cries occur in infants experiencing gastrointestinal discomfort or prolonged distress. They are typically loud, inconsolable, and can last for extended periods. Research on infant cry classification has identified discomfort as a distinct category alongside pain, hunger, sleep, and other emotional states, with gastrointestinal issues such as lower wind pain representing specific triggers for characteristic vocalization patterns [27,33]. Studies employing automated classification systems have documented that discomfort-related cries exhibit unique acoustic signatures that can be distinguished through machine learning approaches utilizing pitch, spectral, and temporal features [28,33]. Studies analyzing cry acoustics across different distress states have confirmed that fundamental frequency increases significantly during heightened arousal, with both pain and distress conditions producing elevated pitch values [22]. Distress cries share some acoustic similarities with pain cries, such as elevated pitch and high energy, but are more cyclic, with alternating periods of intense crying and recovery [4,37]. Spectrograms often show repeated high-energy bursts, but these patterns are more periodic compared to the chaotic nonlinearity of authentic nociceptive pain cries. Colic cries can challenge automated systems because they mimic certain pain-related features while lacking the clear nociceptive trigger [32,36].

The categorization of infant cries provides critical context for developing reliable pain detection algorithms [38]. Pain cries are acoustically distinct, but they can overlap with fear, anger, and distress cries in pitch, intensity, or energy distribution [25]. Research using the ChatterBaby algorithm has demonstrated that while pain cries have distinctive acoustic characteristics, they share neuronal pathways with other distress states, making differentiation challenging [25]. Advanced machine learning and deep learning models, especially those using spectrogram-based visual features, must therefore learn to differentiate between these categories by identifying diagnostic acoustic markers unique to nociceptive crying [4,8]. Incorporating knowledge of cry categories enhances feature engineering, improves model robustness, and reduces misclassification in real-world clinical settings [6].

3.5. Challenges in Infant Cry Signal Interpretation

Interpreting infant cry signals for pain assessment remains a complex task due to the multifaceted nature of neonatal vocal behavior and the influence of environmental, physiological, and developmental factors. Although significant progress has been made in computational analysis of cry acoustics, several challenges hinder the development of reliable and generalizable pain-classification models [22,39]. Real-world deployment remains particularly challenging, with nearly 90% of developed models remaining undeployed to practical clinical applications [38].

- **High Inter-Infant Variability:** Cry acoustics vary significantly across infants due to differences in gestational age, lung capacity, neuromuscular development, and medical conditions [40]. Preterm infants, for example, often exhibit weaker phonation, reduced frequency range, and irregular cry patterns compared to full-term infants, which complicates standardized interpretation. This inter-infant variability makes it difficult to develop universal acoustic descriptors that reliably indicate pain across heterogeneous populations [40].
- **Overlapping Acoustic Characteristics Across Cry Types:** Pain cries often share overlapping features with cries induced by hunger, discomfort, or stress [25]. Although pain cries tend to show higher fundamental frequency (F_0), greater jitter/shimmer, and more turbulent energy, these characteristics are not always exclusive to pain states. Hence, algorithms may misclassify cries when acoustic cues are ambiguous or when multiple stressors co-occur (e.g., handling, cold stress, separation) [38].
- **Sensitivity to Environmental Noise and Recording Conditions:** Neonatal intensive care units (NICUs) and clinical wards are typically high-noise environments with sounds from equipment, conversations, and other infants [41]. Background noise and reverberation degrade the acoustic quality of cry recordings, reducing feature extraction accuracy, particularly for high-frequency energy components used in pain detection. Variable microphone distance, placement, and recording hardware introduce additional inconsistencies that can bias model performance, especially in small datasets [42].
- **Scarcity and Imbalance of Pain-Labeled Data:** Ethical considerations and clinical constraints limit the collection of high-quality, well-annotated pain-cry datasets. Painful procedures such as heel lances or immunizations cannot be performed solely for research, leading to small sample sizes and imbalanced class distributions [37]. This scarcity contributes to model overfitting and reduces generalizability across different populations, institutions, or recording conditions [43]. Additionally, differences in annotation standards (e.g., N-PASS, NIPS, PIPP-R, DAN scores) make cross-dataset comparisons challenging [38].
- **Generalizability Across Clinical and Cultural Settings:** Cry patterns may vary due to cultural caregiving differences, hospital practices, language exposure, or NICU protocols. A deep learning model trained in one clinical environment may struggle to generalize across hospitals with different noise profiles, equipment, or caregiving routines [38,41]. This highlights the need for larger, multi-site datasets and robust domain adaptation techniques.
- **Annotation Subjectivity and Inconsistent Ground Truth:** Human-labeled pain indicators often suffer from subjectivity, even when validated scales such as N-PASS, PIPP, or NIPS are used [1]. Clinical staff may differ in their interpretation of infant behavior, leading to inconsistencies in dataset labels [44]. Inter-rater variability reduces the reliability of the ground truth used to train supervised models, affecting both accuracy and reproducibility [38].

Figure 2 shows an audio classification workflow that begins with data acquisition, where audio signals are collected for analysis. The raw audio data undergoes preprocessing to enhance its quality and prepare it for further steps. Spectrograms are generated from the audio signals to provide a visual representation of their frequency components over time. These spectrograms serve as inputs to deep learning models designed for classification tasks. The models process the spectrograms and predict probabilities for pain versus non-pain categories. Figure 2 illustrates an audio classification workflow, a modified version of Alzubaidi et al., 2021 [45].

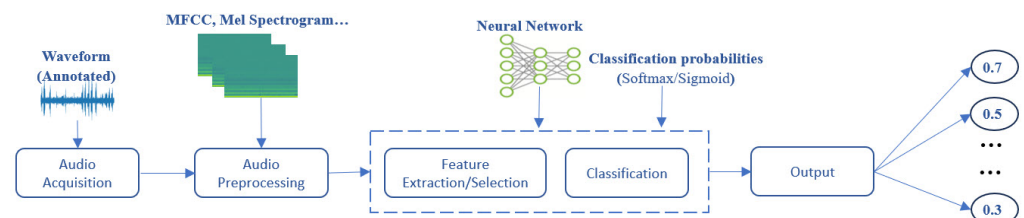


Figure 2. End-to-end infant cry audio classification pipeline for pain assessment. Raw cry recordings are acquired and labeled, then preprocessed (e.g., denoising/segmentation and time–frequency conversion) to generate spectrogram representations. Next, feature extraction/embedding is performed using a deep learning backbone, followed by a classifier that outputs pain vs. non-pain class probabilities, which are finally converted into the predicted label (pain/non-pain) for each recording.

4. Cay Audio Feature Extraction

To facilitate deep learning-based analysis of infant cry signals, raw audio recordings are first converted into spectrogram representations [46,47]. Spectrograms provide a time–frequency view of the signal, capturing dynamic variations in spectral energy and frequency that are not readily observable in the raw waveform [48]. Deep learning architectures, particularly CNN and VLM, are highly effective at learning discriminative patterns from these visual representations, making spectrograms a suitable input for cry classification tasks [47]. Additionally, transformer-based approaches utilizing audio spectrograms have demonstrated remarkable performance in pathology detection tasks, with some models achieving classification accuracies exceeding 98% [8].

Different feature representation domains capture complementary acoustic characteristics of infant cry signals [4]. Since pain-related cues are expressed across temporal, spectral, and perceptual dimensions, incorporating features from multiple domains enhances model robustness and improves classification performance. The following are four of the most widely used feature representation domains:

- **Time-Domain Features:** Time-domain features describe the amplitude and temporal structure of the cry waveform [49]. They are particularly useful for capturing cry intensity, duration, and temporal instability, which are commonly associated with pain-evoked vocalizations [22]. Pain cries typically exhibit higher energy, longer expiratory phases, and irregular amplitude patterns compared to non-pain cries [25]. Research using the ChatterBaby algorithm found that pain cries showed significantly higher energy and distinct amplitude characteristics compared to fussy or hungry cries [25]. These features are associated with raw waveform, frame-level and amplitude envelopes [49]. Time-domain features do not require spectrogram conversion, making them computationally efficient but less expressive for complex pattern learning [49]. The autocorrelation function (ACF) of sound signals has been successfully used to extract time-domain features that achieve high sensitivity in cry detection while maintaining lower computational loads [49]. They are best suited for traditional machine-learning models (SVM, Random Forest, k-NN) [27,50]. Studies have

demonstrated that support vector machines, random forests, and k-nearest neighbor classifiers achieve strong performance when trained on carefully selected time-domain and frequency-domain features [27]. However, they are less effective for CNN-based models due to the lack of spatial structure, which is why deep learning approaches typically rely on spectrogram representations that provide the two-dimensional structure necessary for convolutional operations [47].

- **Frequency-Domain Features:** This feature captures how acoustic energy is distributed across frequencies, which reflects vocal fold tension, harmonic structure, and noise components [30]. Spectral analysis provides insights into the harmonic ratio and frequency distribution patterns that characterize different types of infant cries [31]. These representations summarize frequency information but discard temporal dynamics. Frequency-domain features are less suitable for CNNs unless converted into 2D representations [6,47]. Research has demonstrated that support vector machines, Gaussian mixture models, and multilayer perceptron networks effectively process frequency-domain features for infant cry classification tasks [49]. Studies using extracted frequency features with machine learning classifiers have achieved strong performance, with accuracies exceeding 88% when combined with appropriate feature selection methods [27].
- **Time-Frequency Features:** Time-frequency features jointly represent how frequency content evolves, making them highly effective for analyzing non-stationary signals such as infant cries [47]. The short-time Fourier transform (STFT) technique is employed to transform cry signals into spectrogram images, capturing the dynamic spectral characteristics essential for accurate classification [47]. Time-frequency features are best associated with STFT Spectrogram, Mel Spectrogram, and Log-Mel Spectrogram [46]. Among these, log-Mel spectrograms are the most widely used due to their perceptual relevance and compact representation [51]. Research has demonstrated that Mel spectrograms effectively capture the nuanced characteristics of infant speech and provide robust input for deep learning models [5]. Making them highly suitable for models such as CNN, Vision-Language Models (CLIP, BLIP-2, GPT-4V-style), and CNN-RNN or CNN-Transformer hybrids [8]. Transformer-based approaches utilizing audio spectrograms have achieved remarkable accuracies exceeding 98% in pathology classification tasks [8]. Additionally, hybrid architectures combining CNNs with LSTM networks have demonstrated outstanding performance, with studies reporting accuracies up to 99% when processing spectrogram representations [52]. Time-frequency spectrograms provide the spatial structure required for deep visual feature learning, making them the dominant choice in modern infant pain classification systems [46,47]. The integration of spectrogram-based representations with deep convolutional architectures enables automatic feature extraction and superior classification performance compared to traditional hand-crafted feature approaches [50].
- **Cepstral Features:** Cepstral features provide a compact representation of the spectral envelope, closely modeling human auditory perception [32]. They are highly effective in capturing vocal tract characteristics and phonatory variations associated with pain-induced crying. Research has demonstrated that cepstral coefficients successfully capture the unique qualities of infant cry sounds across different emotional and pain states [53]. MFCC is an example of the associated representation [5]. MFCCs are derived from Mel spectrograms followed by a cepstral transform, effectively compressing spectral information while preserving discriminative cues [7]. Although MFCCs perform well with traditional models [54], they are increasingly being complemented by deep learned embeddings [55]. Studies have shown that combining MFCCs with

deep learning architectures such as convolutional neural networks achieves superior classification accuracy, with some systems reporting accuracies exceeding 95% for infant cry classification tasks [7]. Additionally, self-supervised learning approaches like wav2vec 2.0 have emerged as powerful alternatives that can extract meaningful representations directly from raw audio, offering performance gains beyond traditional MFCC-based features [55].

In deep learning-based infant pain classification, feature extraction is typically grounded in time–frequency representations, enabling models to capture detailed acoustic and temporal characteristics [50]. VLM-based approaches treat cry audio as visual spectrogram image representations and learn more abstract, semantic features through alignment with textual descriptions. Transformer-based algorithms applied to audio spectrograms have shown remarkable capability in classifying infant crying into distinct pathological categories, achieving over 98% accuracy by exploiting self-attention mechanisms that preserve the integrity of diagnostic features [8].

Table 2 summarizes various deep learning models and feature extraction techniques used for audio classification across different infant cry databases, along with their reported accuracies. The study highlights that MFCC is the most commonly used feature across models, often paired with techniques like CNN, and LSTM. Models combining CNN with other architectures, such as RNN, show strong performance, particularly with spectrogram features. Hybrid models, like CNN+SVM and VGG-16+LSTM, demonstrate diverse application of spectral and temporal features. Some methods, like Ji et al.'s [56] CNN with Baby2020, achieve near-perfect accuracy, showcasing the potential of advanced architectures. Spectrogram-based methods consistently perform well, reinforcing the importance of visual feature representation in cry analysis. Overall, the table highlights the diverse methodological approaches and their effectiveness in infant cry sound classification. For additional details, refer to the reference section, which includes the literature review papers for this study.

Table 2. Summary of deep learning-based infant cry audio classification studies, highlighting model architecture, input features, datasets, and reported performance. For each study (authors), the table lists the AI network/architecture, the feature extraction or input representation used (e.g., time–frequency features such as MFCCs or spectrograms), the dataset(s) evaluated, and the reported classification accuracy (as stated in the original paper). The database field specifies either the dataset name or the location where the study's data were collected.

Ref. and Year	Model	Features	Database	Accuracy
[56] 2022	CNN	MFCC	Baby Chillanto/Baby2020	99.28%
[57] 2021	VGG-16+LSTM	Spectrogram	USF-MNPAD-I	79.00%
[58] 2019	VGG16	MFCC	USF-MNPAD-I	96.77%
[59] 2025	CNN/BiLSTM	MFCC	Baby Cry Sounds	89.65%
[60] 2025	CNN/LSTM	MFCC	Baby Cry Sense	89.80%
[61] 2025	CRNN	MFCC	Donate-a-Cry/Baby Chillanto	97.00%
[62] 2024	CNN	Mel-spectrograms	Kyoto University Hospital	92.40%
[63] 2025	CNN	MFCC	Kaggle+ICSD+Donate-a-Cry	95.20%
[47] 2021	CNN + SVM	MFCC	NTUH	93.33%
[7] 2025	CNN	MFCC	Baby Cry	95.5%

NTUH: National Taiwan University Hospital Yunlin Branch; ICSD: Infant Cry and Snoring Detection dataset.

Table 3 presents an expanded summary of the studies introduced in Table 2 by reporting additional methodological and performance-related details. In particular, it includes information on dataset size, task formulation, split strategy, class distribution, leakage-prevention reporting, and the reported precision, F1-score, and accuracy values for each study. For studies in which some of these details were not explicitly provided by the original authors, the corresponding entries are indicated as NR.

Table 3. Supplementary summary of the studies listed in Table 2, providing additional reporting details on data size, task definition, unit of data split, train/test/validation split, class balance, data-leakage control, precision, F1-score, and accuracy. For each referenced study, the table extends the comparative overview in Table 2 by including dataset sample size and selected performance metrics. Where task definition, split unit, class balance, or leakage-control details were unavailable, they are marked as NR.

Ref. and Year	Data Size	Task Definition	Unit of Split	Data Split			Class Balance	Data-Leakage Control	Precision	F1-Score	Accuracy
				Train	Test	Validation					
[57] 2021	216	Pain vs. no-pain classification	Subject	LOSO	LOSO	LOSO	DA	Subject-wise via LOSO	79.64%	79.63%	79.63%
[64] 2019	3250	Infant cry classification	Recording	80%	20%	-	NR		NR	NR	99.28%
[58] 2019	4914	Pain assessment/classification	NR	CV	CV	CV	DA		NR	NR	96.77%
[59] 2025	7088	Infant cry classification	Recording	80%	20%	-	SMOTE	NR	89.00%	88.00%	89.70%
[60] 2025	NR	Infant cry classification	Recording	70%	15%	15%	NR	NR	90.20%	89.70%	89.80%
[61] 2025	(DC+BC) **	Multi-class infant cry classification	Recording	80%	10%	10%	NR	NR	98.00%	97.00%	97.00%
[62] 2024	6523	Infant cry classification	Subject	CV	CV	CV	NR	Subject-wise via LOSO	96.00%	93.50%	92.40%
[63] 2025	9000	Infant cry classification	Recording	70%	15%	15%	DA		92.00%	93.00%	95.20%
[47] 2021	300	Pain vs. no-pain classification	Recording	80%	20%	-	NR		93.33%	90.32%	93.33%
[7] 2025	889	Infant cry classification	Recording	NR	NR	NR	NR	NR	NR	94.60%	95.50%

NR: Not Reported. CV: Cross-Validation. LOSO: Leave-One-Subject-Out. Recording: Audio recording file. Subject: Multiple audio recording file per subject. DA: Data Augmentation. ** Donate-a-Cry merged with Baby Chillanto.

4.1. Cry Signal Augmentation Features

In the realm of machine learning and artificial intelligence, data is often hailed as the new oil, powering the algorithms and models that drive innovation across industries [43]. Data can be scarce or, at best, insufficient, particularly in the context of training robust and accurate models. The scarcity of labeled data has historically posed a significant challenge, leading to the development and evolution of techniques like data augmentation. Over time, this technique has become an integral part of machine learning workflows, especially in the context of classifiers, where it plays a critical role in enhancing model performance, generalization, and robustness [6].

In their study, Salekin et al. [58] applied data augmentation to each audio segment in the audio modality of the USF-MNPAD dataset. Each segment was augmented using three frequency modifications ($f/2$, $f/3$, $2f/3$), six noise levels (0.01, 0.05, 0.001, 0.005, 0.003, and 0.03), and combinations of both frequency and noise perturbations. This augmentation strategy produced 27 variants per audio event, consisting of 3 frequency-only transformations, 6 noise-only transformations, and 18 combined frequency–noise transformations, resulting in a total of 4914 augmented signals. Using this augmented dataset, the authors achieved an accuracy of 96.77%.

Data augmentation is used to increase the diversity of infant cry samples and mitigate the limited availability of labeled pain and no-pain data. Preprocessing steps with data augmentation techniques effectively address dataset imbalances, enabling more balanced training sets for classification models [58]. By introducing controlled variations such as noise perturbations, models become more robust to real-world recording conditions and environmental artifacts commonly present in clinical settings. Research has demonstrated that acoustic scene simulation and properly designed synthetic datasets can help detectors handle challenging circumstances including different environmental noises, improving performance in real NICU environments [65]. Studies show that denoising techniques enhance model robustness when applied to cry signals captured under noisy conditions [42]. This improves generalization performance and reduces overfitting, leading to more reliable infant pain classification from cry sounds. Findings confirm that models trained on augmented data with real-world variations demonstrate significantly improved external validity and generalization compared to models trained without such augmentation [41].

Figure 3 presents examples of infant cry representations, including both original and augmented data. The non-augmented samples, (a) a no-pain cry waveform, (b) its corresponding no-pain log-Mel spectrogram, (c) a pain cry waveform, and (d) its corresponding pain log-Mel spectrogram serve as baseline representations of the two classes. The augmented samples (e)–(h), generated by adding white noise to the no-pain and pain cry waveforms and their corresponding log-Mel spectrograms, introduce realistic variability while preserving the underlying cry characteristics. This form of data augmentation increases the diversity of training data, improves robustness to noise, and enhances the model's ability to generalize when distinguishing between pain and no-pain infant cries.

4.2. Deep Learning-Based Audio Feature Extraction

Early research on infant cry analysis predominantly relied on handcrafted acoustic features derived from classical digital signal processing techniques [50]. These features are typically computed over short-time frames of the cry signal and are designed to capture perceptually and physiologically meaningful characteristics of infant vocalizations. Commonly employed descriptors include time-domain features (e.g., zero-crossing rate and short-time energy), frequency-domain features (e.g., spectral centroid, spectral roll-off, and formant frequencies), and time–frequency representations such as spectrograms [38]. Among these, MFCC are the most widely used due to their ability to approximate human auditory

perception while compactly encoding vocal tract information [7]. The MFCC technique efficiently captures audio characteristics from crying sounds, making it a dominant choice for infant cry classification systems. Studies have demonstrated that MFCCs combined with deep learning models achieve superior accuracy, with some systems reporting over 95% classification performance [7]. CNN are frequently applied to spectrograms treated as two-dimensional images, enabling the automatic learning of localized time–frequency patterns associated with pain-related cry characteristics, such as elevated pitch, abrupt energy variations, and prolonged vocalizations [47]. In addition, RNN and temporal convolutional models have been employed to capture temporal dependencies within cry sequences. Research has shown that CNN-RNN models outperform traditional methods in accuracy when analyzing cry patterns over time [66].

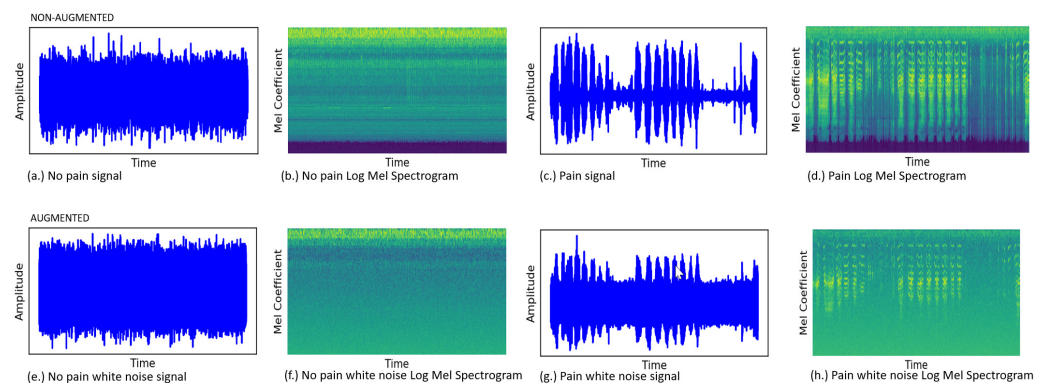


Figure 3. Examples of infant cry representations used in this study. The figure includes eight images: four non-augmented samples consisting of (a) a no-pain cry waveform, (b) the corresponding no-pain log-Mel spectrogram, (c) a pain cry waveform, and (d) the corresponding pain log-Mel spectrogram; and four augmented samples generated using white-noise augmentation, consisting of (e) a no-pain cry waveform with added white noise, (f) the corresponding no-pain log-Mel spectrogram with white noise, (g) a pain cry waveform with added white noise, and (h) the corresponding pain log-Mel spectrogram with white noise. The audio waveform is displayed in blue for visualization only, and the color itself has no analytical significance. On the other hand, in the log mel spectrogram representations, brighter colors denote higher feature magnitudes and darker colors denote lower magnitudes, thereby highlighting variations in acoustic energy and structure that may help distinguish pain-related crying patterns from non-pain vocalizations.

Although recent infant-cry pain assessment studies increasingly rely on Deep Neural Network (DNN) and vision-inspired pipelines on time–frequency representations, Support Vector Machine (SVM) even though not a DNN remain an important and still competitive classical method that is often used either (i) as a standalone classifier on engineered acoustic features such as MFCCs, or (ii) as the final classifier on top of deep features (i.e., a hybrid deep-feature + SVM approach) [47]. This persistence of SVMs in contemporary infant pain classification research reflects both their theoretical robustness and practical effectiveness, particularly in medical settings where interpretability, computational efficiency, and performance on limited training data remain critical requirements. The following studies illustrate representative feature-based approaches for extracting meaningful information for infant pain classification.

Vincent et al. (2021) [47] specifically discriminated neonatal cries into three categories: pain, hunger, and sleepiness. Their research addressed the challenge that conventional deep learning requires enormous amounts of data for effective classification by proposing a hybrid approach combining deep learning feature extraction with machine learning classification. The study employed a Deep Convolutional Neural Network (DCNN) for automatic feature extraction from infant cry signals. The DCNN was combined with a

Support Vector Machine (SVM) classifier to create a hybrid deep learning-machine learning architecture. Neonatal cry auditory signals were first transformed into spectrogram images using the Short-Time Fourier Transform (STFT) technique. The DCNN then took these spectrogram images as input and automatically extracted relevant features. These features were subsequently passed to the SVM classifier for final classification into pain, hunger, or sleepiness categories. Various SVM kernel techniques were compared, including radial basis function (RBF), linear, and polynomial kernels. The study utilized neonatal cry audio signals categorized into pain, hunger, and sleepiness classes. The approach was specifically designed to work effectively with moderate numbers of data samples. The CNN-based feature extraction combined with SVM-RBF classifier achieved an accuracy of 88.89% in classifying infant cries into the three categories (pain, hunger, and sleepiness). The study demonstrated that combining deep learning feature extraction with machine learning classification could achieve promising results. The hybrid CNN-SVM approach proved effective for discriminating pain cries from other cry types, with the SVM-RBF kernel providing superior performance compared to other kernel methods.

In their research, Mala et al. (2025) [7] developed a deep learning-based system to identify and classify infant crying into four distinct classes: pain, hunger, sleepiness, and wet diaper. The study aimed to address the challenges faced by first-time parents in understanding the meaning behind infant cries, particularly given that approximately 130 million babies are born annually worldwide. The study employed CNN as the primary deep learning technique for classifying infant cry sounds. The CNN architecture was specifically designed to leverage spatial hierarchies and patterns in acoustic features extracted from crying sounds. MFCC were used as the feature extraction technique to efficiently capture audio characteristics from infant crying sounds. These MFCC features were then fed into the CNN model for classification. They compared their CNN approach with various existing machine learning models to validate its effectiveness in automatically interpreting infant cries and determining the underlying reasons, including pain. Among all models evaluated, the MFCC combined with Convolutional Neural Networks achieved the highest accuracy of 95.50% for classifying infant cry sounds and determining the reasoning behind the crying, including pain classification. The study utilized a dataset of infant cry audio samples categorized into four classes: pain, hunger, sleepiness, and wet diaper. The MFCC features extracted from each audio sample served as input to the neural network. They demonstrated that deep learning techniques, specifically CNNs combined with MFCC feature extraction, could significantly outperform existing machine learning approaches in infant cry classification. The high accuracy achieved makes this technology-assisted approach promising for early childhood development applications and for providing valuable assistance to parents in identifying pain and other needs in infants.

Mohanty et al. (2025) [59] presented a hybrid machine learning and deep learning framework for classifying infant cries to identify critical needs including hunger, pain, and sleepiness. The study recognized that accurately interpreting infant cries is essential for timely caregiving and that traditional methods often lack the precision needed for reliable real-time infant monitoring. They used hybrid architecture combining Convolutional Neural Network-Bidirectional Long Short-Term Memory Network (CNN-BiLSTM) enhanced with an attention mechanism. The CNN component extracted spatial features from time–frequency representations, while the BiLSTM captured temporal dependencies in cry signals. The attention mechanism specifically improved focus on important time–frequency features relevant for distinguishing pain from other cry types. The approach included comprehensive feature extraction using MFCCs (Mel-Frequency Cepstral Coefficients) along with Delta and Delta2 derivatives to capture both spectral and temporal patterns in infant cry signals. These features were fed into the CNN-BiLSTM architecture where CNN

layers processed spatial characteristics and BiLSTM layers modeled temporal evolution. The attention layer then weighted the importance of different features for final classification. Data augmentation and SMOTE (Synthetic Minority Oversampling Technique) were applied to improve class balance and model generalization. The dataset included infant cry samples categorized into multiple classes including hunger, pain, and sleepiness. SMOTE was applied to address class imbalance and ensure effective learning of patterns for all cry types, including pain. The deep CNN-BiLSTM model with attention mechanism achieved superior performance with 89.65% validation accuracy and a weighted F1-score of 88%. The study demonstrated that incorporating attention mechanisms into a hybrid CNN-BiLSTM architecture significantly enhances infant cry classification performance, including pain detection. The proposed system proved to be sensor-free, computationally efficient, and suitable for real-time infant monitoring applications. The attention mechanism's ability to focus on important temporal and spectral features was particularly valuable for distinguishing pain cries from other cry types.

Saraswathy et al. (2018) [67] investigated binary classification of infant cry signals across three groups: healthy and deaf infants, infants with jaundice, and infants separated from the deaf group. They combined features from the time and frequency domains into a unified time–frequency ($t - f$) domain, generating $t - f$ representations and extracting $t - f$ -based features. Key features such as MFCC and Linear Prediction Coefficient (LPC) were utilized to evaluate the effectiveness of this approach. The extracted feature vectors were validated using Probabilistic Neural Networks (PNN) and General Regression Neural Network (GRNN). Their method demonstrated a 90.00% classification accuracy, highlighting the potential of $t - f$ techniques for analyzing infant cry signals.

4.3. VLM-Based Infant Cry Pain Assessment Approach

As of 2025, only one published paper is known to specifically address infant pain assessment from cry sounds using VLM with an explicit prompting-based approach [68]. VLMs enable joint learning from visual and textual modalities. Multimodal learning with transformers has become a hot topic in AI research, with transformer-based models achieving great success in various machine learning tasks [69]. In this approach, infant cry audio is first transformed into visual representations, typically log-Mel spectrogram images (or MFCC), which preserve temporal and spectral structure [8]. VLMs process these spectrogram images alongside natural-language prompts describing pain-related acoustic patterns (e.g., “high-pitched, intense, and irregular cry”). This allows the model to associate visual acoustic patterns with semantic concepts, enabling few-shot or zero-shot classification without extensive supervised training [55]. Compared to traditional deep learning pipelines, VLM-based feature extraction emphasizes semantic alignment and interpretability, making it particularly appealing in clinical contexts. The ability to incorporate expert knowledge through textual prompts also provides flexibility when labeled cry datasets are limited. Self-supervised learning approaches demonstrate that leveraging unlabeled data can significantly lower the cost of building robust models, particularly when annotated medical datasets are expensive and time-consuming to create [55,70].

In their innovative study, McCofie et al. (2025) [68] introduced a novel approach specifically for infant pain detection that addressed the limitations of conventional deep neural networks, which typically demand large labeled datasets, substantial computational power, and often lack interpretability. The research represented the first application of few-shot prompting with vision–language models for infant pain classification, focusing specifically on pain detection rather than general cry classification. The study leveraged OpenAI's vision–language model, GPT-4(V), combined with mel spectrogram-based representations of infant cries. Unlike traditional deep learning approaches requiring extensive training

on large datasets, this method employed a few-shot prompting strategy that utilized the pre-trained capabilities of GPT-4(V) to interpret visual representations of cry sounds. Infant cry sounds were first converted into mel spectrograms. These mel spectrogram images were then presented to the GPT-4(V) model along with carefully designed prompts that guided the model to classify whether the cry indicated pain. The prompting strategy significantly reduced dependence on large training datasets while enhancing transparency and interpretability compared to traditional black-box deep learning models. The research utilized the USF-MNPAD-II dataset, which is specifically designed for infant pain assessment. This dataset contains cry recordings from infants under various conditions, enabling the model to learn pain-specific acoustic patterns. Using the USF-MNPAD-II dataset, the method achieved an accuracy of 83.33% with only 16 training samples, which stood in stark contrast to the 4914 samples required by the baseline model. This remarkable efficiency in learning from minimal data represented a significant advancement in the field. The study successfully demonstrated that vision–language models like GPT-4(V) could be effectively applied to infant pain classification through few-shot learning, dramatically reducing data requirements while maintaining high accuracy. This approach offered significant advantages in terms of interpretability, as the prompting method made the model’s decision-making process more transparent compared to traditional deep learning architectures. The research opened new avenues for developing practical pain assessment systems that could be deployed in resource-limited settings where large labeled datasets are unavailable. The workflow for the study is as shown in Figure 4, and the model comparison with baseline can be seen in Table 4.

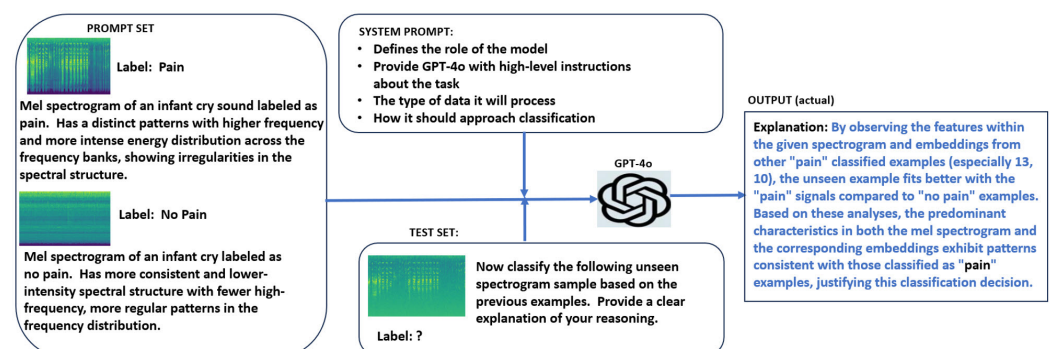


Figure 4. Few-shot VLM prompting pipeline for infant cry pain classification from log-mel spectrograms. The diagram illustrates the complete prompt package provided to a GPT-4o-style vision–language model: (i) a system prompt that sets the model’s role (infant cry pain-assessment assistant) and gives high-level instructions on how to interpret spectrogram images and produce a binary decision; (ii) a balanced few-shot exemplar set of labeled log-mel spectrograms for Pain and No Pain, where each exemplar includes the spectrogram image, its class label, and a short class-defining description (e.g., pain typically exhibits higher pitch, stronger intensity, longer duration, and more irregular energy distribution, whereas no-pain tends to be lower-pitched, more rhythmic, shorter, and more stable); and (iii) an unseen test spectrogram (unknown label) supplied as the query. The model uses the system instructions and the exemplar patterns to infer the label of the test spectrogram and returns the predicted class (Pain/No Pain), optionally with a brief rationale consistent with the exemplar definitions [68]. The symbol ‘?’ in ‘Label: ?’ indicates an unlabeled, previously unseen sample that is provided to the model for inference, after which the model assigns the predicted class label.

Table 4 presents a comparative evaluation of [68] using VLM and few-shot prompting approach against baseline methods [58] using tradition methods. As shown, the pre-trained GPT-4o-based prompting strategy attains an accuracy of 83.33%, surpassing the conventional LPCC/MFCC + SVM model (82.35%), which highlights the benefit of employing

structured prompt-driven inference for infant pain assessment. While its performance is marginally below that of the Spectrogram + N-CNN approach (91.20%), ref. [68] relies on a few-shot prompting. Nevertheless, it delivers competitive accuracy, underscoring its ability to learn effectively from limited data.

Table 4. Comparison of model performance by showing that pre-trained GPT-4o prompting achieves competitive accuracy using only 16 few-shot examples, outperforming LPCC/MFCC+SVM methods, although it remains below the performance of spectrogram-based CNNs trained on 4914 samples.

Method	Training Data Siz	Few-Shot Size	Accuracy
LPCC/MFCC + SVM [58]	4914	N/A	82.35%
Spectrogram + N-CNN [58]	4914	N/A	91.20%
GPT-4o (Prompting) [68]	Pre-Trained	16	83.33%

N/A: Not Applicable.

5. Datasets for Infant Pain Assessment

The quality and structure of datasets are foundational to the validity of deep learning models for infant pain assessment, since model performance is inherently constrained by the reliability, representativeness, and clinical relevance of the underlying data. In this domain, variations in dataset size, recording environment, annotation procedures, pain scales, and class distribution can substantially influence the outcomes and the interpretation of reported results. Accordingly, this section examines the principal dataset and data-quality considerations that shape current progress in automated pain assessment from infant cry sounds [41,71].

5.1. Dataset Size and the Role of Data Augmentation

A central limitation in infant pain assessment research is the small size and limited diversity of available cry datasets. Many studies rely on recordings from a relatively small number of infants, often collected in specific neonatal units or controlled environments, which can restrict variability in age, clinical condition, recording quality, and background noise [4,41]. In addition, recording context differs across studies. Some datasets are captured in real clinical settings during routine painful procedures, while others are obtained under more controlled or semi-structured conditions. These differences directly influence ecological validity, since recordings from clinical environments better reflect real-world deployment conditions but may also contain more noise, overlapping speech, and equipment interference [72].

Because such datasets are often limited in both scale and heterogeneity, data augmentation has become an important strategy for improving model robustness and reducing overfitting [73]. Common augmentation methods include time shifting, pitch variation, additive noise, time stretching, and spectrogram masking, all of which can increase training diversity without requiring new data collection. However, augmentation must be applied carefully, as excessive or unrealistic transformations may distort clinically meaningful cry characteristics and compromise label fidelity. For this reason, augmentation is most useful when it preserves the acoustic cues associated with pain expression and is restricted to training data, while validation and test sets remain untouched to ensure reliable performance assessment [73–75].

5.2. Labeling Methods and Ground-Truth Definition

Accurate labeling is fundamental because model performance is directly constrained by the quality of the ground truth. In infant pain assessment, labels may be assigned using several approaches, including expert clinical judgment, nurse assessments, standardized

pain scoring systems, or indirect physiological validation. The level of confidence in a dataset depends heavily on who performed the annotation, what criteria were followed, and whether inter-rater agreement was established. Labels based solely on coarse observational judgment may introduce subjectivity, while annotations supported by trained clinicians or validated pain scales are generally more reliable [2,76].

5.3. Data Quality Challenges

A common limitation in infant cry datasets is class imbalance, where one class, often no pain or a dominant distress category substantially outnumbers the other. This imbalance can bias models toward the majority class and produce misleadingly high accuracy if more informative metrics such as sensitivity, recall, or F1-score are not reported [4]. Beyond class imbalance, dataset quality may also be affected by sampling bias, repeated recordings from the same infant, inconsistent recording durations, background noise, missing metadata, and variability in microphone placement. These factors can reduce the robustness of learned representations and complicate fair comparison across studies. Consequently, a dataset-focused discussion should examine whether studies used balancing strategies, subject-wise partitioning, quality control procedures, and preprocessing steps designed to reduce confounding influences. Attention to these issues is essential for determining whether reported performance reflects genuine pain discrimination or dataset-specific artifacts [73,77].

5.4. Hospital-Based Infant Pain Datasets

Hospital-based datasets focus on pain cries elicited during routine medical procedures such as vaccinations, heel lances, and ear piercing, often accompanied by standardized pain scales and clinical observations. Studies have recorded neonates undergoing painful procedures, including heel blood sampling, arterial blood sampling, intravenous injection, subcutaneous injection, peripheral intravenous cannulation, and other routine medical interventions [78]. In these corpora, audio is recorded in neonatal or pediatric units with dedicated microphones or bedside recorders before, during, and after a painful procedure, and segments are later extracted to represent pre-procedure baseline, peak pain during the stimulus, and recovery. Pain labels are commonly derived from validated behavioral scales such as N-PASS and NIPS or from the procedure type and timing, with scoring performed by trained nurses or clinicians observing the infant in real time or from synchronized video [20]. These datasets have been used in recent work that converts cry waveforms into time–frequency images or cepstral features and applies deep learning architectures to classify pain versus non-pain states, reporting high discrimination accuracy in NICU settings [57,68,79].

- USF-MNPAD-II: One of the widely cited resource is the Neonatal Pain Assessment Database (often referred to as NPAD or USF-MNPAD), collected by the University of South Florida from 75+ (ongoing) neonates undergoing routine painful procedures in a neonatal intensive care setting. Audio was recorded in parallel with video and vital signs around procedures such as heel lancing, with segments labeled as baseline (no pain), procedural pain, and post-procedure, using the Neonatal Infant Pain Scale (NIPS) and N-PASS scored by trained nurses who observed the infant during each phase. The labels in this corpus therefore reflect standardized behavioral pain scores rather than only clinical event tags, and multiple studies have used it to train and evaluate both handcrafted-feature and deep learning models, including convolutional networks operating on spectrograms for binary pain vs. no-pain classification [57,68]. The primary aim is to facilitate classification tasks and other related pain assessments for neonatal care. The dataset was developed as part of a funded research initiative

by the University of South Florida's Computer Science and Engineering Department and was conducted in collaboration with Tampa General Hospital [80,81]. The USF-MNPAD dataset is publicly accessible for research purposes on a case-by-case basis, subject to a written request.

- **Baby Chillanto:** The Baby Chillanto Database is a publicly available corpus compiled in Mexico for infant cry pathology and condition classification, and it has been repeatedly cited in cry analysis research. Recordings consisting of 2268 cry samples, collected from infants under controlled conditions and then segmented into one-second audio samples that are grouped into five classes: asphyxia, deaf, hunger, normal, and pain, with labeling performed by clinical experts who knew the infant's diagnosis and recording context. Because it includes both pathological and normal cries along with a distinct pain class, this database has been used to train multiclass classifiers that distinguish pain from non-pain states within a broader clinical screening framework, often using standard speech-processing features and deep learning models [82].

Several infant pain cry databases are available for research, either through institutional online repositories or by written request. Donate A Cry, is an audio dataset of infant cries designed to support the classification of infant needs and discomfort. The dataset includes 457 audio recordings, each approximately seven seconds long, collected from infants aged 0–2 years. The recordings were processed and cleaned prior to labeling, resulting in a curated set of cry samples organized by need category [83]. Another resource is the ChatterBaby database developed at UCLA, a large-scale audio collection of infant cries used to train the ChatterBaby algorithm and apps. The database was initially constructed from approximately 2000 pain-related cry recordings, including a clearly identified subset of about 300 samples captured during painful procedures such as vaccinations and ear piercings [25,83].

Table 5 summarizes available infant/neonate cry sound datasets commonly used in pain assessment research. Most of these databases do not report key details such as the pain scale employed, the type of pain recorded, or the number of subjects. In contrast, the USF-MNPAD datasets provide more comprehensive documentation.

Table 5. Available infant/neonatal cry datasets used in pain assessment research, with key cohort and annotation details. For each dataset, the table summarizes the source/publication (authors), dataset name, infant age range, pain context/type (e.g., procedural/clinical), recorded modalities (face, body, audio, vital signs), number of subjects, number of data used (size), and the pain assessment scale/ground-truth labeling method used. The database field specifies either the dataset name or the location where the study's data were collected.

Database	Age Range	Pain Type	Modalities				Subjects	Size	Pain Scale	Ref.
			Face	Body	Audio	Vital Sign				
USF-MNPAD-I	GA 27–41 weeks	Proc + Post Op.	✓	✓	✓	✓	45	4914	NIPS/N-PASS	[58,80]
USF-MNPAD-II	GA 27–41 weeks	Proc.	✓	✓	✓	✓	75+	7958	N-PASS	[68,80]
Donate-A-Cry	0–2 yrs	belly pain	✗	✗	✓	✗	NR	457	NR	[84]
Baby Chillanto	0–6 mths	NR	✗	✗	✓	✗	NR	2268	NR	[82]
DA-IICT	0–2 yrs	NR	✗	✗	✓	✗	NR	251	NR	[85]
NTUH	1–10 days	NR	✗	✗	✓	✗	29	300	NIPS	[47]

NR: Not Reported, GA: Gestational Age, Proc + Post Op.: Procedural + Post Operative pain, Proc.: Procedural pain, NTUH: National Taiwan University Hospital Yunlin Branch.

5.5. Dataset Availability

Regarding accessibility, Baby Chillanto is described as a publicly available database, and it has been used in multiple studies classification where pain is treated as one of several outcome categories. Using the Baby Chillanto dataset, research has explored the use of typical features used in audio signal analysis—chromagram, Log-Mel spectrogram, and Mel frequency cepstral coefficients (MFCC) to determine whether deep learning models

can perform classification tasks [86]. The USF neonatal pain datasets (NPAD/USF-MNPAD series) are collected under institutional review protocols, and while not as openly distributed as some public corpora, they are extensively used in publications on neonatal pain assessment, including recent work that applies CNN-based models and few-shot vision–language prompting for pain cry detection [68,80,81].

6. Challenges, Limitations, and Future Directions

Deep Learning models for pain assessment face challenges like limited and imbalanced datasets, which can lead to overfitting and poor generalization across different populations. Ethical concerns regarding data privacy and the interpretation of model predictions can hinder widespread clinical adoption.

Data quality and quantity have a significant impact on the effectiveness of infant cry classification for pain due to the inherent challenges in collecting and labeling infant cry data. Data scarcity is a major challenge when training deep learning models, as DL demands a large amount of data to achieve exceptional performance [43]. High-quality data is essential because variations in recording conditions, such as background noise or recording devices, can introduce inconsistencies, making it difficult to accurately classify infant cry signal. A cry detection model trained on in-lab data underperforms when presented with real-world data, highlighting the challenges posed by environmental variations [41]. The limited quantity of available data exacerbates this issue, as infant cry recordings for pain, in particular, are less common and harder to obtain, leading to an imbalance in the dataset [43]. This imbalance can cause the classification model to be biased toward the more prevalent class, often resulting in inaccurate predictions [6,48]. The variability in cry sounds between different infants, influenced by factors like age and health, adds another layer of complexity, necessitating a large and diverse dataset to train a robust model [41]. Without sufficient data, the model may struggle to generalize well to unseen examples, especially for edge cases with atypical cries. Research has shown that infant cry classification models typically offer low amounts of training data, which limits their effectiveness [87]. The subjective nature of labeling cries as pain or no pain also introduces potential inaccuracies, as human annotators may disagree, leading to noisy labels [88]. These limitations together make it challenging to develop a reliable infant cry classification system that can differentiate between cries of pain and no pain across diverse scenarios

6.1. Interpretability and Transparency

The lack of transparency and explainability of the deep learning models currently used in healthcare, especially in classifying infant cries for pain, is a major drawback [89]. While deep learning has demonstrated impressive accuracy in various audio tasks, the decision-making process of these models remains opaque. This “black-box” nature makes it difficult to understand how model arrive at a prediction, undermining their trustworthiness in healthcare applications. In the case of audio data, such as infant cries, these models rely on complex, high-dimensional patterns that are difficult to interpret, making the explanations for model outputs less comprehensible to humans. Without clear transparency, medical professionals may find it challenging to validate these systems in clinical settings.

Additionally, existing explainable AI (XAI) methods, developed for visual data, often struggle to generalize effectively to audio models. Attribution-based methods, which focus on highlighting salient pixels in an image, face several challenges. These include issues such as inconsistent heatmap generation, reproducibility problems, sensitivity to perturbations, and a lack of insight into the actual decision-making process [90]. Moreover, these methods provide little to no information about the underlying audio patterns contributing to the model’s predictions.

6.2. Interpretability and Explainability

The field of Audio Explainability will benefit from future research that concentrates on designing explainability techniques optimized for audio data and resolving the deficiencies of current XAI methods:

Robust and Reproducible Attribution Methods: Future efforts should focus on developing robust attribution techniques that are reproducible across different datasets and models. These methods should not only highlight relevant regions of the audio signal but also ensure that the attributions are consistent across similar examples. Reproducibility is crucial for establishing trust, particularly in high-stakes applications like healthcare.

Non-attribution Explanations in Audio: Future research in non-attribution explanations for audio models should prioritize automated concept extraction, counterfactual reasoning, and optimized prototype-based methods. Concept-based explanations can be adapted for audio by automatically extracting relevant features (e.g., cry intensity, duration) using techniques like ACE [91], minimizing annotation costs and human error. To enhance counterfactual explanations, research should focus on improving perturbation quality, ensuring meaningfulness and plausibility within medical contexts. Additionally, prototype-based methods require optimization to efficiently manage large audio datasets and ensure selected prototypes are unbiased by irrelevant noise. By addressing completeness, scalability, and robustness, future research can develop effective non-attribution explanations for audio models, enhancing interpretability and trustworthiness in healthcare applications such as infant pain detection. Finally, future audio explainability research should integrate clinical expertise, providing actionable insights into specific cry features indicative of pain. The goal is to develop novel, domain-specific XAI techniques that deliver robust, scalable, and clinically relevant explanations, ensuring AI's safe and trustworthy adoption in healthcare.

6.3. Integration with Multimodal Data

Model integration with multimodal data can enhance the accuracy and reliability of pain classification by combining auditory signals with complementary data sources like physiological, environmental, and visual inputs. This approach leverages the strengths of different data modalities, providing a more holistic understanding of the infant's condition, which can help differentiate between subtle variations in cries that might indicate pain or other needs. Advanced deep learning models that can handle multimodal inputs offer the potential to process and integrate these diverse data streams in a simultaneous manner. Future research could explore the synergy between audio signals and other sensor data, refining classification models and improving real-time responsiveness. This direction promises to advance both clinical and at-home applications, offering better tools for caregivers and healthcare professionals to assess and respond to infants' needs in a more effective and safe manner [57,92,93].

6.4. Data Collection and Dataset Expansion

Data collection and dataset expansion are crucial future directions for improving the accuracy of infant cry classification for distinguishing between cries of pain and no pain. By gathering a larger and more diverse set of cry recordings, the model can be trained on a broader range of acoustic variations, reducing bias and improving generalization across different infants. Expanding the dataset to include cries from various environments and conditions can also help the model become more robust in real-world applications. Moreover, adding more labeled data, especially from less represented groups, can enhance the model's ability to detect subtle differences between pain and no-pain cries. In the end,

these steps will contribute to creating a more reliable and sensitive tool for early diagnosis and intervention [94].

7. Conclusions

The review examined deep learning approaches for pain assessment in infant cry sounds, focusing on feature representations, model architectures, datasets, and evaluation practices. The literature shows a clear progression from handcrafted acoustic features, such as MFCCs and related time–frequency descriptors, to data-driven deep learning models, with CNN-based architectures remaining the most widely used. Recent studies have also begun exploring transformer-based methods and vision–language paradigms using spectrogram representations. A key trend is the increasing value of multimodal approaches, which combine cry audio with facial, physiological, or clinical information to better capture the multidimensional nature of neonatal pain and improve robustness beyond unimodal systems.

Despite these advances, major gaps remain. Existing datasets are often small, private, and heterogeneous, with inconsistent labeling methods, variable pain scales, and limited cross-dataset validation, all of which restrict reproducibility and generalizability. Future work should prioritize larger and more diverse clinically grounded datasets, standardized benchmarks, stronger external validation, and more interpretable models. Multimodal learning, along with few-shot, transfer, and prompt-based methods, offers a practical direction for addressing data scarcity and improving real-world performance. Overall, infant cry-based pain assessment remains a promising non-invasive strategy, but its clinical translation will depend on more rigorous validation, improved data quality, and systems designed for reliable use in practice.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/make8030076/s1>, PRISMA 2020 Checklist [95].

Author Contributions: Conceptualization, methodology, and manuscript preparation were performed by the authors. P.R.M. contributed to this work by reviewing and refining the conceptual and visual presentation of the methodology. Specifically, he provided critical feedback on the relevance and clarity of the diagrams, and assisted in improving domain-specific wording of the manuscript to ensure consistency with the stated research objectives. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Alfred P. Sloan Foundation University Center for Exemplary Mentoring (UCEM) award G-2020-12674 and by the National Institutes of Health Grants (NIH R21NR018756, R41HD109086).

Institutional Review Board Statement: Not applicable. This study is a review of previously published works and does not involve new studies with human participants or animals.

Informed Consent Statement: Not applicable. No human subjects were involved in this study.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This research was supported in part by a grant from the Alfred P. Sloan Foundation University Center for Exemplary Mentoring (UCEM) award G-2020-12674 and by the National Institutes of Health Grants (NIH R21NR018756, R41HD109086). We extend our gratitude to them for their generous support. The author would also like to express his sincere gratitude to the contributors whose invaluable insights and expertise made this work possible and greatly enriched the scope of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Llerena, A.; Tran, K.; Choudhary, D.; Hausmann, J.; Goldgof, D.; Sun, Y.; Prescott, S.M. Neonatal pain assessment: Do we have the right tools? *Front. Pediatr.* **2023**, *10*, 1022751. [[CrossRef](#)] [[PubMed](#)]
2. Sarkaria, E.; Gruszfeld, D. Assessing Neonatal Pain with NIPS and COMFORT-B: Evaluation of NICU's Staff Competences. *Pain Res. Manag.* **2022**, *2022*, 8545372. [[CrossRef](#)]
3. Huang, X.Z.; Li, L.; Zhou, J.; He, F.; Zhong, C.X.; Wang, B. Evaluation of three pain assessment scales used for ventilated neonates. *J. Clin. Nurs.* **2018**, *27*, 3522–3529. [[CrossRef](#)]
4. Zayed, Y.; Hasasneh, A.; Tadj, C. Infant cry signal diagnostic system using deep learning and fused features. *Diagnostics* **2023**, *13*, 2107. [[CrossRef](#)] [[PubMed](#)]
5. Muhammad, N.F.; Dewan, R.; Pusppanathan, J.; Suryanata, F.A. Baby crying sound classification using convolutional neural network. *J. Hum. Centered Technol.* **2024**, *3*, 67–74. [[CrossRef](#)]
6. Jamal, A.; Al-Azani, S. A machine-learning approach for children's pain assessments using prosodic and spectral acoustic features. In *Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
7. Mala, B.; Darandale, S.S. Deep Learning-based Classification of Infant Cries for Parental Assistance: A Study using MFCC and CNN Technique. In *Proceedings of the 2025 International Conference on Electronics and Renewable Systems (ICEARS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1812–1818.
8. Tami, M.; Masri, S.; Hasasneh, A.; Tadj, C. Transformer-based approach to pathology diagnosis using audio spectrogram. *Information* **2024**, *15*, 253. [[CrossRef](#)]
9. Shayegh, S.V.; Tadj, C. Deep audio features and self-supervised learning for early diagnosis of neonatal diseases: Sepsis and respiratory distress syndrome classification from infant cry signals. *Electronics* **2025**, *14*, 248. [[CrossRef](#)]
10. Tzeng, J.T.; Li, J.L.; Chen, H.Y.; Huang, C.H.; Chen, C.H.; Fan, C.Y.; Huang, E.P.C.; Lee, C.C. Improving the robustness and clinical applicability of automatic respiratory sound classification using deep learning-based audio enhancement: Algorithm development and validation. *JMIR AI* **2025**, *4*, e67239. [[CrossRef](#)] [[PubMed](#)]
11. Ji, C.; Mudiyansele, T.B.; Gao, Y.; Pan, Y. A review of infant cry analysis and classification. *Eurasip J. Audio, Speech, Music. Process.* **2021**, *2021*, 8. [[CrossRef](#)]
12. Morgan, M.E.; Kukora, S.; Nemshak, M.; Shuman, C.J. Neonatal Pain, Agitation, and Sedation Scale's use, reliability, and validity: A systematic review. *J. Perinatol.* **2020**, *40*, 1753–1763. [[CrossRef](#)]
13. Subramanian, P.; Chitgupikar, S.R.; Jalapu, S.; Ilyas, T. Comparative Analysis of Three Pain Scales for Evaluation of Procedural Pain in Neonates: A Cross-sectional Study. *J. Clin. Diagn. Res.* **2025**, *19*, 6. [[CrossRef](#)]
14. Chaudhary, K.; Charaya, S. Comparative Analysis of the Proposed AI-Based Monitoring Framework with Existing Pain Assessment Practices and Quantification of Improvements. *AIJFR-Adv. Int. J. Res.* **2025**, *6*, 1–19. [[CrossRef](#)]
15. Walas, W.; Malinowska, E.; Halaba, Z.P.; Szczapa, T.; Latka-Grot, J.; Rutkowska, M.; Kubiacyk, A.; Wrońska, M.; Piotrowski, A.; Skrzypek, M.; et al. Newborn infant parasympathetic evaluation for the assessment of analgesedation adequacy in infants treated by mechanical ventilation—A multicenter pilot study. *Arch. Med. Sci. AMS* **2021**, *17*, 1651. [[CrossRef](#)]
16. Dincer, E.; Özer, H.; Topçuoğlu, S.; Karatekin, G. Ultrasonography causes agitation and pain leading to hemodynamic disturbance in neonates: A prospective observational study. *Children* **2023**, *10*, 347. [[CrossRef](#)]
17. Relland, L.M.; Gehred, A.; Maitre, N.L. Behavioral and physiological signs for pain assessment in preterm and term neonates during a nociception-specific response: A systematic review. *Pediatr. Neurol.* **2019**, *90*, 13–23. [[CrossRef](#)]
18. Uberos Fernández, J.; Campos Martínez, A.M.; Ruiz-López, A.; Fernández-Marín, E.; García Serrano, J.L. Sensitivity and Specificity of the Newborn Infant Parasympathetic Evaluation Index in Pain Assessment of Very Low Birth Weight Infants. *Am. J. Perinatol.* **2024**, *41*, e430–e434.
19. Das, U.; Cho, E.; Balasubramanian, T.; Zhou, J.; Mansoor, S.; Kapoor, R.; Forbes, H.E.; Chase, S.; Kore, M.; Sidelnikov, D.; et al. Comparing infant pain and stress during retinopathy of prematurity screening using ophthalmoscopy and non-contact imaging. *Sci. Rep.* **2025**, *15*, 32771. [[CrossRef](#)] [[PubMed](#)]
20. Desai, S.A.; Nanavati, R.N.; Jasani, B.B.; Kabra, N. Comparison of neonatal pain, agitation, and sedation scale with premature infant pain profile for the assessment of acute prolonged pain in neonates on assisted ventilation: A prospective observational study. *Indian J. Palliat. Care* **2017**, *23*, 287. [[CrossRef](#)] [[PubMed](#)]
21. Manworren, R.C.; Horner, S.; Joseph, R.; Dadar, P.; Kaduwela, N. Performance evaluation of a supervised machine learning pain classification model developed by neonatal nurses. *Adv. Neonatal Care* **2024**, *24*, 301–310. [[CrossRef](#)] [[PubMed](#)]
22. Laguna, A.; Pusil, S.; Acero-Pousa, I.; Zegarra-Valdivia, J.A.; Paltrinieri, A.L.; Bazán, Á.; Piras, P.; Palomares i Perera, C.; Garcia-Algar, O.; Orlandi, S. How can cry acoustics associate newborns' distress levels with neurophysiological and behavioral signals? *Front. Neurosci.* **2023**, *17*, 1266873. [[CrossRef](#)]
23. Popowicz, H.; Kwiecień-Jaguś, K.; Olszewska, J.; Mędrzycka-Dąbrowska, W.A. Pain scales in neonates receiving mechanical ventilation in neonatal intensive care units—systematic review. *J. Pain Res.* **2020**, *13*, 1883–1897. [[CrossRef](#)]

24. Kappesser, J.; Kamper-Fuhrmann, E.; de Laffolie, J.; Faas, D.; Ehrhardt, H.; Franck, L.S.; Hermann, C. Pain-specific reactions or indicators of a general stress response?: Investigating the discriminant validity of 5 well-established neonatal pain assessment tools. *Clin. J. Pain* **2019**, *35*, 101–110. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Parga, J.J.; Lewin, S.; Lewis, J.; Montoya-Williams, D.; Alwan, A.; Shaul, B.; Han, C.; Bookheimer, S.Y.; Eyer, S.; Dapretto, M.; et al. Defining and distinguishing infant behavioral states using acoustic cry analysis: Is colic painful? *Pediatr. Res.* **2020**, *87*, 576–580. [\[CrossRef\]](#)
26. Myakala, P.R.; Nalumachu, R.; Sharma, S.; Mittal, V. A low cost intelligent smart system for real time infant monitoring and cry detection. In *Proceedings of the TENCON 2017—2017 IEEE Region 10 Conference*; IEEE: Piscataway, NJ, USA, 2017; pp. 2795–2800.
27. Chang, C.Y.; Bhattacharya, S.; Raj Vincent, P.D.; Lakshmana, K.; Srinivasan, K. An Efficient Classification of Neonates Cry Using Extreme Gradient Boosting-Assisted Grouped-Support-Vector Network. *J. Healthc. Eng.* **2021**, *2021*, 7517313. [\[CrossRef\]](#)
28. Kolacz, J. Investigating Infant Crying Persistence and Cry Acoustic Features as Early Risk Indicators for Social Adjustment: Developmental Associations with Infant Vagal Tone and Attachment Stress. Ph.D. Thesis, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 2016.
29. Abou-Abbas, L.; Tadj, C.; Fersaie, H.A. A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes. *J. Acoust. Soc. Am.* **2017**, *142*, 1318–1331. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Khalilzad, Z.; Hasasneh, A.; Tadj, C. Newborn cry-based diagnostic system to distinguish between sepsis and respiratory distress syndrome using combined acoustic features. *Diagnostics* **2022**, *12*, 2802. [\[CrossRef\]](#)
31. Khalilzad, Z.; Kheddache, Y.; Tadj, C. An entropy-based architecture for detection of sepsis in newborn cry diagnostic systems. *Entropy* **2022**, *24*, 1194. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Khalilzad, Z.; Tadj, C. Using CCA-fused cepstral features in a deep learning-based cry diagnostic system for detecting an ensemble of pathologies in newborns. *Diagnostics* **2023**, *13*, 879. [\[CrossRef\]](#)
33. Dharwadkar, N.V.; Dixit, A.A.; Kannur, A.K.; Kadampur, M.A.B.; Joshi, S. Identification of reasons behind infant crying using acoustic signal processing and deep neural network for neonatal intensive care unit. *Int. J. Inf. Retr. Res. (IJIRR)* **2022**, *12*, 17. [\[CrossRef\]](#)
34. Mendhakar, A.M.; Sreedevi, N.; Arunraj, K.; Shanbal, J.C. Infant screening system based on cry analysis. *Int. Ann. Sci.* **2019**, *6*, 1–7. [\[CrossRef\]](#)
35. Lindová, J.; Špinka, M.; Nováková, L. Decoding of baby calls: Can adult humans identify the eliciting situation from emotional vocalizations of preverbal infants? *PLoS ONE* **2015**, *10*, e0124317. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Keles, E.; Bagci, U. The past, current, and future of neonatal intensive care units with artificial intelligence: A systematic review. *NPJ Digit. Med.* **2023**, *6*, 220. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Coro, G.; Bardelli, S.; Cuttano, A.; Scaramuzzo, R.T.; Ciantelli, M. A self-training automatic infant-cry detector. *Neural Comput. Appl.* **2023**, *35*, 8543–8559. [\[CrossRef\]](#)
38. Owino, G.; Shibwabo, B. Advances in Infant Cry Paralinguistic Classification—Methods, Implementation, and Applications: Systematic Review. *JMIR Rehabil. Assist. Technol.* **2025**, *12*, e69457. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Aggarwal, G.; Jhajharia, K.; Izhar, J.; Kumar, M.; Abualigah, L. A machine learning approach to classify biomedical acoustic features for baby cries. *J. Voice* **2025**, *39*, 1446–1455. [\[CrossRef\]](#)
40. Ahila, A.; Princye, P.H.; Reshmi, T.S.; Vithyalakshmi, N.; Monisha, N.; Saravanakumar, R.; Lakshmi, S.J. Machine Learning Methods for Recognizing Infant Cries: An Extensive Examination of Audio Signal Processing. In *Proceedings of the 2025 6th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 10–12 September 2025; pp. 2242–2247. [\[CrossRef\]](#)
41. Yao, X.; Micheletti, M.; Johnson, M.; Thomaz, E.; de Barbaro, K. Infant crying detection in real-world environments. In *Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2022; pp. 131–135.
42. Owino, G.; Shibwabo, B. Noise-Resilient Bioacoustics Feature Extraction Methods and Their Implications on Audio Classification Performance: Systematic Review. *JMIR Biomed. Eng.* **2025**, *10*, e80089. [\[CrossRef\]](#)
43. Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Al-Dabbagh, B.S.N.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *J. Big Data* **2023**, *10*, 46. [\[CrossRef\]](#)
44. Ilhan, E.; Pacey, V.; Brown, L.; Spence, K.; Galea, C.; Adams, R.; Hush, J.M. Exploration and validation of behavioral pain measures and physiological pain measures factor structure extracted from the pain assessment tool item scores for infants admitted to neonatal intensive care. *Clin. J. Pain* **2021**, *37*, 397–403. [\[CrossRef\]](#)
45. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [\[CrossRef\]](#)

46. Herlea, D.M.; Iancu, B.; Ardelean, E.R. A Study of Deep Learning Models for Audio Classification of Infant Crying in a Baby Monitoring System. *Informatics* **2025**, *12*, 50. [\[CrossRef\]](#)
47. Ashwini, K.; Vincent, P.; Srinivasan, K.; Chang, C.Y. Deep learning assisted neonatal cry classification via support vector machine models. *Front. Public Health* **2021**, *9*, 670352.
48. Joshi, V.R.; Srinivasan, K.; Vincent, P.D.R.; Rajinikanth, V.; Chang, C.Y. A multistage heterogeneous stacking ensemble model for augmented infant cry classification. *Front. Public Health* **2022**, *10*, 819865. [\[CrossRef\]](#)
49. Narayanan, S.P.; Manikandan, M.S.; Cenkeramaddi, L.R. Fast Autocorrelation Feature-Based Infant Cry Detector for Resource-Efficient Affordable Edge Cry Sound Analysis Systems. In *Proceedings of the 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
50. Ashwini, K.; PM, D.R.V.; Srinivasan, K.; Chang, C.Y. Deep convolutional neural network based feature extraction with optimized machine learning classifier in infant cry classification. In *Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA)*; IEEE: Piscataway, NJ, USA, 2020; pp. 27–32.
51. Narayanan, S.P.; Manikandan, M.S.; Cenkeramaddi, L.R. Spectrogram and LSTM Based Infant Cry Detection Method for Infant Wellness Monitoring Systems. In *Proceedings of the 2024 16th International Conference on Human System Interaction (HSI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
52. Reddy, K.S.; Rithani, M.; Rangarajan, P.K.; Mohan, G.B. A Comparative Analysis: Enhancing Baby Cry Detection with Hybrid Deep Learning Techniques. In *Proceedings of the 2023 International Conference on Next Generation Electronics (NEleX)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
53. Maiti, A.; Dutta, C.; Banerjee, J.S.; Sarigiannidis, P. Ai for infant well-being: Advanced techniques in cry interpretation and monitoring. *J. Mech. Contin. Math. Sci.* **2024**, *19*, 1–27 [\[CrossRef\]](#)
54. Mekhfioui, M.; Fadel, W.; Hammouch, F.E.; Laayati, O.; Bouchouirbat, M.; El Bazi, N.; Satif, A.; Boujiha, T.; Chebak, A. Development of a Baby Cry Identification System Using a Raspberry Pi-Based Embedded System and Machine Learning. *Technologies* **2025**, *13*, 130. [\[CrossRef\]](#)
55. Gorin, A.; Subakan, C.; Abdoli, S.; Wang, J.; Latremouille, S.; Onu, C. Self-supervised learning for infant cry analysis. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
56. Ji, C.; Jiao, Y.; Chen, M.; Pan, Y. Infant cry classification based-on feature fusion and mel-spectrogram decomposition with cnns. In *Proceedings of the International Conference on AI and Mobile Services*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 126–134.
57. Salekin, M.S.; Zamzmi, G.; Goldgof, D.; Kasturi, R.; Ho, T.; Sun, Y. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Comput. Biol. Med.* **2021**, *129*, 104150. [\[CrossRef\]](#)
58. Salekin, M.S.; Zamzmi, G.; Paul, R.; Goldgof, D.; Kasturi, R.; Ho, T.; Sun, Y. Harnessing the power of deep learning methods in healthcare: Neonatal pain assessment from crying sound. In *Proceedings of the 2019 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*; IEEE: Piscataway, NJ, USA, 2019; pp. 127–130.
59. Mohanty, S.P.; Mohapatra, T.; Adhikari, S.; Sirisha, V.; Sahu, N.; Sahu, B.J.R. Infant Cry Classification Using CNN-BiLSTM With Attention Mechanism and MFCC Features. In *Proceedings of the 2025 International Conference on Artificial intelligence and Emerging Technologies (ICAET)*, Bhubaneswar, India, 28–30 August 2025; pp. 1–6. [\[CrossRef\]](#)
60. Sekar, R.R.; Sadhik, N.; Krishna, G.S.; Vardhan, T.H. Cryfusion: A Multimodal Infant Cry Classification and Health Monitoring System Using Deep Learning and Embedded Sensors. In *Proceedings of the 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–8.
61. Owino, G.; Kamanu, T.; Ndiritu, J.; Kikechi, C.B. Adaptive infant cry classification using multi-armed bandit modality selection in an attentive convolutional recurrent neural network model. *Complex Intell. Syst.* **2025**, *11*, 375. [\[CrossRef\]](#)
62. Shinya, Y.; Ueno, T.; Kawai, M.; Niwa, F.; Tomotaki, S.; Myowa, M. Listening deeper: Neural networks unravel acoustic features in preterm infant crying. *Sci. Rep.* **2025**, *15*, 23204. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Héritier, N.M.; Muyisa, M.K.; Grâce, K.L.; Pascal, K.U.; Dieu-Donné, K.K.; Olivier, M.C. Lightweight Deep Learning System for Infant-Cry Recognition with Real-Time Notification in Resource-Constrained Environments. *Mach. Learn. Appl.* **2025**, *22*, 100791.
64. Kheddache, Y.; Tadj, C. Identification of diseases in newborns using advanced acoustic features of cry signals. *Biomed. Signal Process. Control* **2019**, *50*, 35–44. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Severini, M.; Ferretti, D.; Principi, E.; Squartini, S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. *IEEE Access* **2019**, *7*, 51982–51993. [\[CrossRef\]](#)
66. Setyoningrum, N.G.; Utami, E.; Wibowo, F.W. A Comprehensive Survey of Infant Cry Classification Research Trends and Methods: A Systematic Review. In *Proceedings of the 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
67. Saraswathy, J.; Hariharan, M.; Khairunizam, W.; Sarojini, J.; Thiagar, N.; Sazali, Y.; Nisha, S. Time–frequency analysis in infant cry classification using quadratic time frequency distributions. *Biocybern. Biomed. Eng.* **2018**, *38*, 634–645. [\[CrossRef\]](#)

68. McCofie, A.; Kandiyana, A.; Mouton, P.R.; Sun, Y.; Goldgof, D. Few-Shot Prompting with Vision Language Model for Pain Classification in Infant Cry Sounds. In *Proceedings of the 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*; IEEE: Piscataway, NJ, USA, 2025; pp. 857–862.
69. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [\[CrossRef\]](#)
70. Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.Y.; Wang, Y.; Tian, Y.; Gao, W. Large-scale multi-modal pre-trained models: A comprehensive survey. *Mach. Intell. Res.* **2023**, *20*, 447–482. [\[CrossRef\]](#)
71. Onu, C.C. Making deep neural networks work for medical audio: Representation, compression and domain adaptation. *arXiv* **2025**, arXiv:2506.13970.
72. Mundri, D.H.; Solanki, S.S.; Mahto, K. An approach to finding the most suitable algorithm for noise reduction in a Neonatal cry signal. In *Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
73. Setyoningrum, N.G.; Utami, E.; Kusri, K.; Wibowo, F.W. Improving Infant Cry Recognition Using MFCC And CNN-Based Audio Augmentation. *J. Tek. Inform. (Jutif)* **2025**, *6*, 995–1016. [\[CrossRef\]](#)
74. Zhang, K.; Ting, H.N.; Choo, Y.M. Baby cry recognition by BCRNet using transfer learning and deep feature fusion. *IEEE Access* **2023**, *11*, 126251–126262. [\[CrossRef\]](#)
75. Zhang, X.; Shao, Y.; Xu, J.; Ma, Y.; Zhang, W.Q. Classification of short audio acoustic scenes based on data augmentation methods. In *Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1920–1926.
76. Guinsburg, R.; Almeida, M.F.B.d.; Peres, C.d.A.; Shinzato, A.R.; Kopelman, B.I. Reliability of two behavioral tools to assess pain in preterm neonates. *Sao Paulo Med. J.* **2003**, *121*, 72–76. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Xia, M.; Huang, D.; Wang, W. Multi-task learning for audio-based infant cry detection and reasoning. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 7434–7446. [\[CrossRef\]](#)
78. Chen, X.; Zhu, H.; Mei, L.; Shu, Q.; Cheng, X.; Luo, F.; Zhao, Y.; Chen, S.; Pan, Y. Video-based versus on-site neonatal pain assessment in neonatal intensive care units: The impact of video-based neonatal pain assessment in real-world scenario on pain diagnosis and its artificial intelligence application. *Diagnostics* **2023**, *13*, 2661. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Meesters, N.J.; Simons, S.H.; van Rosmalen, J.; Holsti, L.; Reiss, I.K.; Van Dijk, M. Acute pain assessment in prematurely born infants below 29 weeks: A long way to go. *Clin. J. Pain* **2019**, *35*, 975–982. [\[CrossRef\]](#)
80. Hausmann, J.; Salekin, M.S.; Zamzmi, G.; Mouton, P.R.; Prescott, S.; Ho, T.; Sun, Y.; Goldgof, D. Accurate neonatal face detection for improved pain classification in the challenging NICU setting. *IEEE Access* **2024**, *12*, 49122–49133. [\[CrossRef\]](#)
81. Salekin, M.S.; Zamzmi, G.; Goldgof, D.; Mouton, P.R.; Anand, K.J.; Ashmeade, T.; Prescott, S.; Huang, Y.; Sun, Y. Attentional generative multimodal network for neonatal postoperative pain estimation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 749–759.
82. Liu, Q.; Song, L.; Xu, D.; Long, Y. ICSD: An open-source dataset for infant cry and snoring detection. *arXiv* **2024**, arXiv:2408.10561. [\[CrossRef\]](#)
83. Hashemi, S.; Kolivand, H.; Khan, W.; Saba, T. Infant Cry Analysis: A Survey of Datasets, Features, and Machine Learning Techniques. *IEEE Trans. Affect. Comput.* **2025**, *14*, 21–40. [\[CrossRef\]](#)
84. Singh, S.K.; Anand, B.S.; Bhatia, A.; Singh, G. A Novel Approach for Infant Cry Classification Using Transformer Models. In *Proceedings of the 2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA)*; IEEE: Piscataway, NJ, USA, 2023; pp. 631–638.
85. Chittora, A.; Patil, H.A. Data collection of infant cries for research and analysis. *J. Voice* **2017**, *31*, 252.e15–252.e26. [\[CrossRef\]](#) [\[PubMed\]](#)
86. Hoq, E.; Nwogu, I. Exploring the Differences between Deaf and Hearing Infant Cries. In *Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.
87. Anders, F.; Hlawitschka, M.; Fuchs, M. Comparison of artificial neural network types for infant vocalization classification. *Ieee/Acm Trans. Audio Speech Lang. Process.* **2020**, *29*, 54–67.
88. Lin, S.; Zhu, H.; Tong, K.; Zhao, Y.; Chen, S.; Pan, Y. Label Uncertainty Suppression Based on Heterogeneous Siamese Network for Neonatal Pain Assessment in Uncontrolled Conditions. In *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics 2025*, Atlanta, GA, USA, 26–29 October 2025.
89. Hossain, M.I.; Zamzmi, G.; Mouton, P.; Sun, Y.; Goldgof, D. Enhancing Neonatal Pain Assessment Transparency via Explanatory Training Examples Identification. In *Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*; IEEE: Piscataway, NJ, USA, 2023; pp. 311–316.
90. Hossain, M.I.; Zamzmi, G.; Mouton, P.R.; Salekin, M.S.; Sun, Y.; Goldgof, D. Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions. *Acm Comput. Surv.* **2023**, *57*, 1–46 [\[CrossRef\]](#)

91. Ghorbani, A.; Wexler, J.; Zou, J.Y.; Kim, B. Towards automatic concept-based explanations. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9277–9286.
92. Gkikas, S.; Tsiknakis, M. Automatic assessment of pain based on deep learning methods: A systematic review. *Comput. Methods Programs Biomed.* **2023**, *231*, 107365. [[CrossRef](#)]
93. Shi, Y.; Wu, W. Multimodal non-invasive non-pharmacological therapies for chronic pain: Mechanisms and progress. *BMC Med.* **2023**, *21*, 372. [[CrossRef](#)]
94. Abisheva, G.; Goranin, N.; Razakhova, B.; Aidynov, T.; Satybaldina, D. Specifics of Data Collection and Data Processing during Formation of RailVista Dataset for Machine Learning-and Deep Learning-Based Applications. *Sensors* **2024**, *24*, 5239. [[CrossRef](#)]
95. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.