This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

Infant Cry Analysis: A Survey of Datasets, Features, and Machine Learning Techniques

Seyyed Mohammad Hossein Hashemi, Hoshang Kolivand *Senior Member IEEE*, Wasiq Khan *Senior Member IEEE*, Tanzila Saba *Senior Member IEEE*

***Abstract*—Knowledge about infant language can go a long way in supporting parents, nurses, and care providers in improving babies' health conditions. Crying is the most effective tool through which babies convey their requirements. In this work, several studies dealing with infant cry detection and classification are contrasted. Research demonstrates that machine learning techniques can effectively categorize and classify infant needs and certain disorders. Several datasets, including Baby Chillanto, Donate A Cry Corpus and Dunstan Baby Language, are presented. After reviewing existing Datasets, preprocessing methodologies and audio feature extraction such as MFCC, RMS energy, etc., are discussed. For infant cry detection and classification, several algorithms, such as support vector machines (SVM), convolutional neural networks (CNN), k-nearest neighbors (KNN), Random Forest, etc., have been analyzed and utilized for such processes in general. Finally, the study explores various applications of infant cry analysis, highlighting its potential to improve infant care and facilitate early diagnosis. As a result of the findings, it has been observed that infant cry analysis can effectively identify different needs and potential health concerns with high accuracy. These machine learning models' classification outputs have the potential to (1) improve childcare practices, (2) detect medical issues earlier, and (3) monitor infants continuously. These features give medical professionals and caregivers useful information for prompt intervention. The implementation of these findings can be applied in hospitals, neonatal intensive care units (NICUs), smart baby monitoring systems, and research studies focused on early childhood development.**

***Index Terms*—infant cry classification, infant cry detection, machine learning, audio feature extraction, deep learning**

I. Introduction

The neonatal period (the first 28 days of life) represents the highest-risk interval for child mortality, accounting for 2.3 million global deaths in 2022. Newborns comprised 47% of all under-five deaths that year, with 75% of these deaths occurring during the first week of life and between 25% and 45% occurring within the first 24 hours [1]. Figure 1 illustrates the primary causes for these fatalities, such as pneumonia and birth asphyxia. Many of these pathological conditions are preventable; indeed, the World Health Organization (WHO) predicts that up to two-thirds of newborn deaths could be averted with early diagnosis and treatment [2].

Newborns in mammalian species use both visual and acoustic cues to express their needs [3]–[5]. Across phylogeny, infant distress vocalizations improve survival by eliciting predictable caregiving responses [3], [4], [6]–[8]. Nevertheless, Esposito et al. [9] showed that crying behavior in people with autism spectrum disorder (ASD) is frequently more difficult to interpret, which could compromise the quality of caregiving. The caregiver's perception of crying is influenced by acoustic characteristics, which are often atypical in children with ASD [3], [10]. According to some studies, other disorders such as deafness [11]–[14], asphyxia [11], [14]–[16], and pain measurement [17], [19] can be diagnosed. Infant cry research involves reviewing datasets, extracting and selecting features, and performing classification. Collecting infant cry data is challenging due to ethical and privacy concerns, as well as sensitivity to environmental noise and infant conditions [20]. Most datasets include various cry types (e.g., hunger, pain) recorded in hospital Neonatal Intensive Care Units (NICUs) using specialized devices [20].

Two fundamental tasks define this field: infant cry detection and classification. Cry detection is the binary task of identifying a cry event in an audio signal, typically to activate a monitoring system. In contrast, cry classification is a more complex, multi-class challenge that interprets the cry's meaning—be it a specific need (e.g., hunger, pain) or a potential pathology (e.g., asphyxia, deafness). Because these tasks require different approaches to data labeling, feature extraction, and model complexity, this survey will specify which task is being addressed when evaluating each methodology and its performance.

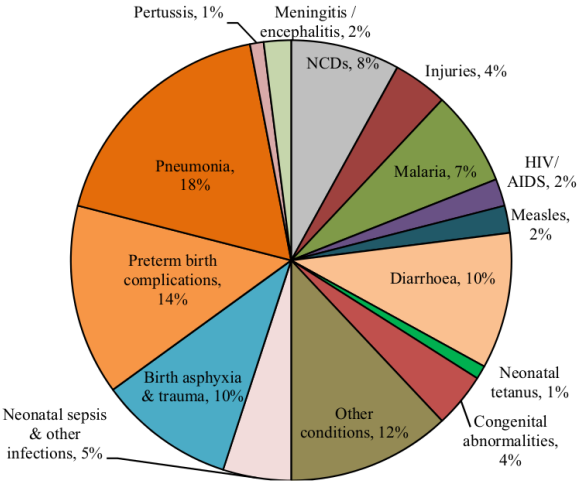

Fig. 1: Causes of fatality in children under five years of age [2]

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

Several datasets, including Donate A Cry Corpus and Chilanto, are presented. Common features used for audio classification include Mel spectrograms, MFCC (Mel-frequency cepstral coefficients), STFT (Short-time Fourier transform), and Mel spectrograms. This work introduces additional features to improve performance. After extracting features, the next step is to select the most informative ones. Some approaches make use of a single feature, whereas others combine multiple features for better accuracy.

Selecting an appropriate model and algorithm is essential for infant cry detection and classification. Some studies are focused solely on classification [21], [22] or detection [23]–[26], while others address both [11], [27], [28]. The choice of model and algorithm is determined by the specific application. In a variety of applications, researchers frequently use machine learning algorithms like Random Forest [58], [63], Support Vector Machines (SVMs) [12], [55], artificial neural networks (ANNs) [11], [21], fuzzy systems [37], and K-Nearest Neighbors (KNN) [23], [28], etc. These algorithms have achieved acceptable accuracy rates. This paper provides a comprehensive review of various features and algorithms implemented on different datasets, highlighting their respective accuracy outcomes. Before deploying these models in real-world applications, factors such as robustness to noise, computational efficiency, and accuracy must be considered. The models should be robust against noise, exhibit low latency, and achieve high accuracy in classification. These models can be deployed in mobile applications or embedded systems.

A. Motivation

Infant cry classification has been recognized as a significant area of research in pediatric care and early childhood development. A sophisticated system capable of automatically classifying infant vocalizations could transform the way infants' needs are understood and addressed, potentially enhancing parental care and enabling the earlier detection of health issues [3], [29]. Key points to consider:

- Crying is a primary way infants communicate their needs and distress [3], [29].
- Cry analysis may automatically detect the presence of illness or discomfort before severe symptoms manifest [3].
- Various machine learning approaches have shown great promise in classifying cries into their basic categories based on acoustic features, such as pattern variations in the spectrogram or fundamental frequency [29].

By creating systems that automatically categorize baby cries and utilizing recent developments in machine learning and speech, these might be completed much more accurately and in practical ways. The findings will have a significant impact on advancements in our knowledge of early childhood development, early detection of health problems, and pediatric care improvements [3], [29].

B. Contributions

Our review delivers a comprehensive overview and assessment of research on infant cry analysis that connects scholarly discoveries with practical uses. By assessing recording settings, categorization kinds, and the practical and ethical issues related to their collection and usage, it methodically gathers and contrasts more than 15 datasets. Furthermore, it examines cutting-edge feature extraction techniques, including MFCC, STFT, DWT, and PLP, pointing out their advantages and disadvantages in various noisy settings. The study examines important trade-offs between accuracy, computational efficiency, and real-time feasibility by contrasting contemporary deep learning techniques (CNN, RNN, Vision Transformers) with conventional machine learning models (SVM, KNN, Random Forest). It also includes benchmark performance measures. Crucially, the analysis also assesses how these approaches might be incorporated into smart baby monitoring, neonatal intensive care unit systems, and early diagnostic tools for disorders like autism spectrum disorder. By doing this, this study contributes significantly to clinical and consumer healthcare improvements by providing practical insights and a framework for converting research breakthroughs into reliable, deployable newborn cry classification systems.

II. Related Surveys

Our analysis extends prior research by referencing Chunyan Ji's review of infant cry analyzing methodologies. Ji's review covers machine learning classifiers including KNN, SVM, CNN, and RNN and highlights feature extraction techniques like MFCC, spectrogram, and fundamental frequency. Although Ji provides insightful information about algorithm and signal processing methods, this survey stands out for offering a thorough and current analysis of current datasets, addressing ethical issues that are important to data collection, and emphasizing real-world applications in early diagnosis, smart monitoring systems, and neonatal care. Additionally, this book presents a comprehensive view of the area by synthesizing findings from a wider range of investigations. [29].

III. Survey Methodology

Figure 2 illustrates the methodology followed for the literature

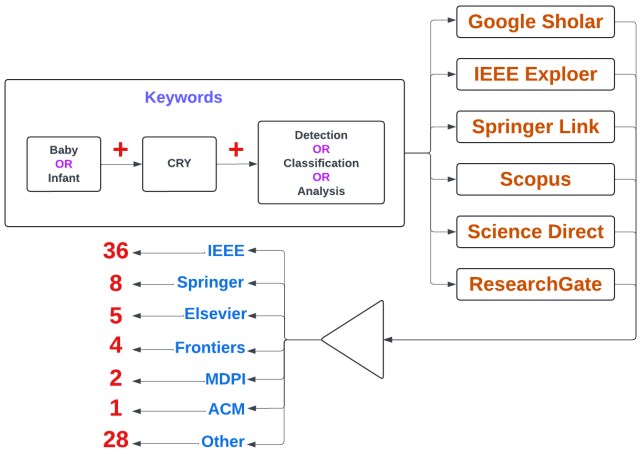

Fig. 2: Precedence of Survey Methodology

survey. To collect relevant articles, the most commonly used

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

search query in this review was ("infant" OR "baby") AND ("cry") AND ("detection" OR "classification" OR "analysis"). The search was performed across multiple academic databases, including Google Scholar, IEEE Xplore, Scopus, Springer Link, ScienceDirect, and ResearchGate.

TABLE I: Comparison of Various Based on Keywords

Keyword	All	IEEE	Springer	Elsevier	Other
Baby Cry Detection	130	47	11	10	62
Baby Cry Classification	70	27	9	5	29
Baby Cry Analysis	50	15	5	5	25
Infant Cry Detection	50	25	6	5	14
Infant Cry Classification	60	27	9	10	14
Infant Cry Analysis	50	13	9	12	35

TABLE II: Comparison of Various Publishers

Publisher	Collected	Reviewed
IEEE	48	36
Springer	18	8
Elsevier	13	5
Frontiers	8	4
MDPI	3	2
ACM	3	1
Other	51	28
Total	**144**	**84**

Table I presents the articles collected from various publishers. Many duplicate entries were identified and subsequently removed. Additionally, articles were excluded based on their titles if they were unrelated to algorithmic processing and instead focused on clinical experiments conducted by nurses or parents. Further exclusions were made after reviewing the abstracts, particularly for articles that did not propose a new method or focused on classification using EEG or other non-audio data formats. Table II summarizes the final set of collected and reviewed articles from different publishers.

The distribution of the final 84 reviewed articles shows an almost even split between conference proceedings (43 publications) and journals (41 publications). Analysis of conference publications highlights the central role of IEEE-sponsored venues, which account for 85.7% of the reviewed conference papers. This strong concentration suggests that research in infant cry analysis is predominantly driven by the engineering and signal processing communities, where IEEE holds significant influence. This finding underscores the dominance of IEEE as a preferred venue for conferences, while journal publications were more varied across different publishers.

IV. Organisation

The remainder of this manuscript is organized as follows:

- **Section V** provides a detailed overview of existing infant cry datasets. It examines over 15 distinct databases, highlighting their characteristics, such as recording environments and classification types, and discusses the practical and ethical challenges related to their collection and use.
- **Section VI** reviews feature extraction methods. It explores a range of techniques, from the widely used Mel-Frequency Cepstral Coefficients (MFCC) to more advanced time-frequency representations like Short-Time Fourier Transform (STFT), and discusses their respective strengths and weaknesses in analyzing infant cry signals.
- **Section VII** covers the machine learning models used for cry detection and classification. This section contrasts traditional algorithms like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) with modern deep learning architectures, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), analyzing trade-offs in accuracy and computational efficiency.
- **Section VIII** presents the results from the surveyed studies. It synthesizes the performance benchmarks of different models and features across various datasets, providing a comparative analysis of their effectiveness in both binary detection and multi-class classification tasks.
- **Section IX** offers a critical discussion of the findings. It interprets the results from the previous section, highlighting key trade-offs between model complexity, computational cost, and classification performance, and contextualizes the practical implications of these findings.
- **Section X** explores the real-world domains and applications of infant cry analysis. It emphasizes how these technologies can be integrated into clinical tools, smart baby monitors, and systems for early developmental screening.
- **Section XI** outlines the primary challenges and future directions for the field. It addresses critical issues such as data accessibility, model generalizability, noise robustness, and ethical considerations, proposing potential solutions to guide future research.
- **Section XII** provides concluding remarks, summarizing the key takeaways from the survey and underscoring the potential of infant cry analysis to advance pediatric care and our understanding of early childhood development.

V. Existing Datasets

The advancement of infant cry analysis is fundamentally reliant on the availability and quality of datasets. While the number of databases has grown, a critical evaluation reveals significant, persistent limitations that challenge the reliability and generalizability of many research findings. These shortcomings, detailed in Table III, include a lack of demographic diversity, a high risk of model overfitting due to small sample sizes, and inherent biases introduced by different recording environments. Addressing these issues is paramount for moving from academic models to robust, real-world applications.

A primary challenge is the lack of data representativeness. Many foundational datasets are ethnically and culturally homogeneous, such as the Baby Chillanto (primarily Mexican), which was specifically designed to study pathological conditions like asphyxia and hypoacoustics [11], SPLAN (Romanian), and various self-recorded datasets focusing on single populations like Taiwanese or Indian infants [11], [27], [64]. Models trained on such data learn narrow patterns and are often brittle, failing to generalize across a diverse global population. This problem is compounded by inconsistent reporting of infant age, a crucial developmental variable.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE III: Overview of Key Infant Cry Datasets and Their Characteristics

Database	Creator	Classes and Samples	Age	Ethnic Diversity	Recording Environment	Limitations	Studies
Baby Chillanto (2004)	National Institute of Astrophysics, Mexico	Total: 2268 (deaf: 879, asphyxia: 340, normal: 507, hunger: 350, pain: 192)	-	Primarily Mexican infants	Clinical settings (NICU, hospital)	Limited ethnic diversity; no age data; potential clinical bias	[13]–[15], [35]
Dunstan Baby Language (2006)	Priscilla Dunstan	Total: 315 (hungry: 56, sleepy: 106, burping: 55, belly pain: 37, discomfort: 61)	Under 3 months	-; likely Western focus	Home recordings	Small sample size risks overfitting; unclear ethnic diversity	[34], [48], [49]
Donate a Cry Corpus (2019)	Gabor Veres	Total: 457 (hungry: 382, tired: 24, burping: 8, belly pain: 16, discomfort: 27)	0 to 2 years	Global (crowdsourced via app)	Home and varied environments	Small sample size for some classes; inconsistent recording quality	[55]–[57]
SPLAN (2015)	Sf. Pantelimon Hospital, Romania	Total: 13,373 (colic: 225, eructation: 505, hungry: 5,536, pain: 4,404)	0 to 3 months	Primarily Romanian infants	Hospital and home settings	Large but imbalanced dataset; limited ethnic diversity	[63], [64]
Self Dataset (2019)	Chang & Tsai	Total: 19,691 (Pain: 5,445, Hunger: 6,263, Sleepiness: 4,927, Wet diaper: 3,056)	0 to 9 months	Not Specified	Home environment (recorded by parents via smartphone)	Likely a single-subject dataset; inconsistent recording quality	[27]
DA-IICT Cry (2018)	Self-recorded	Total: 1,190 (normal: 793, asphyxia: 215, asthma: 182)	-	Likely Indian infants	-	No age or environment details; potential selection bias	[16], [47]
CRIED (2018)	INTERSPEECH 2018	Total: 2,169 (hungry: 586, pain: 723, sleepy: 860)	-	-	-	Lack of age and diversity data; potential overfitting risk	[67], [68]
iCOPE (2019)	infantcope.org	Total: 113 (pain: 42, no pain: 71)	-	-	Clinical settings	Very small dataset; high overfitting risk; limited diversity	[19]
ChatterBaby (2020)	chatterbaby.org	Total: 1,071 (fussy: 171, hungry: 167, pain: 353, colic: 380)	0 to 24 months	Global (crowdsourced)	Home and varied environments	Imbalanced classes; variable recording quality	[70]
Self Dataset (2023)	Hanindito & Kristian	3 Pain Classes (face+video)	Below 12 months	Indonesian infants	Hospital settings	Limited sample size and diversity; video-based bias	[17]
Self Dataset (2016)	Univ. of Firenze, Hospitals, Italy	Total: 3324 CUs (Preterm: 1662, Full-term: 1662)	Full-term: ¡2 days; Preterm: 20-30 days	Italian infants	Clinical setting (NICU, neonatology)	Limited diversity; significant age gap between groups; hunger/stimulated cry	[77]
Autism DB (2020)	Shahid Beheshti Univ.	Total: 359 Sample (Autism: 31, Normal: 31)	18 to 53 months	Iranian infants	Clinical settings	Small sample size; limited diversity; clinical bias	[71]
Autism DB (2019)	Anhui Medical Univ.	Total: 84 (Normal: 64, Autism: 20)	-	Chinese infants	Clinical settings	Very small dataset; high overfitting risk	[72]
ADEL (2010)	Self-recorded	Total: 39 (Normal: 22, ADEL: 17)	-	-	-	Extremely small dataset; high overfitting risk; lacks metadata	[73], [74]
Hypothyroid (2009)	Univ. of Milano	Total: 88 (Normal: 45, Hypothyroid: 43)	Newborns	Italian infants	Clinical settings	Small dataset; limited diversity; clinical bias	[75], [76]

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

Some datasets focus on an extremely narrow age range (e.g., 1-10 days [27]) while others omit this information entirely, making it impossible to assess a model's performance across different stages of early development.

Furthermore, the trade-off between data quantity and quality introduces a high risk of producing statistically weak models. Many available datasets are notably small—such as Dunstan (315 samples), iCOPE (113 samples), and the extremely small ADEL (39 samples), which significantly increases the likelihood of model overfitting. These datasets are prime examples of what can be termed pathology-focused datasets, which are invaluable for clinical research but are often limited in size. Other important examples in this category include the Autism DB datasets, created specifically for early ASD screening [71], [72], and the Hypothyroid database [76]. While larger datasets like SPLAN exist, they often suffer from severe class imbalance, where common cries like "hunger" vastly outnumber others, biasing the classifier. The recording environment also creates a critical dichotomy. Data from controlled clinical settings (NICUs, hospitals) is clean and well-labeled but lacks the background noise and variability of real-world scenarios. Conversely, crowdsourced datasets like Donate A Cry Corpus and ChatterBaby capture more realistic conditions and offer greater diversity but are plagued by inconsistent audio quality and unverified labels [70], [126].

To create a more complete picture of infant distress and overcome the limitations of relying on audio alone, some research has shifted towards multimodal datasets. This approach combines cry audio with other physiological and behavioral signals. For instance, Yosi Kristian et al. and Ghada Zamzmi et al. merged voice and facial expression data [17], [18], while Ana Laguna et al. [78] integrated cry vocalizations with electroencephalography (EEG), near-infrared spectroscopy (NIRS), and videos of facial expressions and body movements. Similarly, Schuller et al. [69] combined voice data with heart rate measurements, enriching the potential for more accurate and context-aware analysis.

As a direct technical response to the pervasive issues of small sample sizes and class imbalance, researchers widely employ data augmentation. By synthetically expanding the training data, these methods help mitigate overfitting and improve model robustness. For example, Ashwini et al. augmented spectrographic images before feeding them into a deep convolutional neural network (CNN) [22]. To combat classifier bias in the CRIED dataset, Jindal et al. [111] successfully applied a suite of techniques, including pitch shifting, time stretching, noise addition, and dynamic range compression [111], [112].

VI. Feature Extraction

Feature extraction is a critical component in classification studies. Since infants primarily communicate through crying, machines can be trained to analyze these vocalizations by extracting the same frequency-based attributes humans use to distinguish them from adult speech. Figure 3 illustrates these differences. Beyond simple detection, classification aims to interpret an infant's needs (e.g., hunger, pain), which is a more complex task that benefits caregivers. This is achieved through the careful selection of the most relevant features.

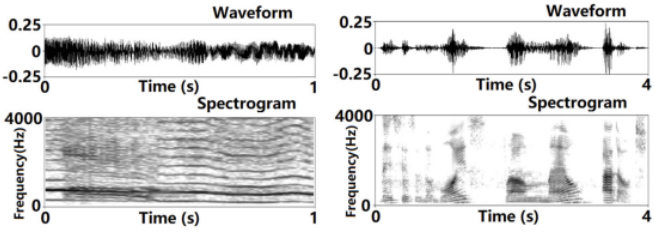

Fig. 3: Adult speech vs. infant cry signal in time and frequency domain [29]

The main audio features are divided into several domains, as shown in Figure 4, and key techniques used by researchers are introduced below.

A. STFT

The Short-Time Fourier Transform (STFT) spectrogram is one of the time-frequency analysis methods that provides clear computational benefits. First, finite-duration parts of the signal are isolated using a window function, and then each windowed segment's discrete Fourier transform is calculated [79], [80]. This segmentation method preserves computational simplicity while offering an effective trade-off between time and frequency resolution [13]. The STFT is especially well-suited for real-time applications in cry analytic systems due to its simple construction and quick processing when compared to more complex time-frequency representations.

Infant crying signals are highly non-stationary, meaning their frequency content changes significantly over time. Due to this characteristic, some researchers have utilized the STFT and to analyze infant crying signals.

B. Mel Spectrogram and MFCC

A Mel Spectrogram is a powerful visual representation of a sound signal's spectrum where the frequency axis is converted to the Mel scale [59]. This scale is designed to mimic the non-linear way humans perceive pitch, giving more resolution to lower frequencies that are more critical for interpreting vocal sounds. The resulting image plots time against the Mel-scaled frequency, with the color or intensity representing the amplitude of the sound. This format is particularly effective for deep learning models like Convolutional Neural Networks (CNNs), which can analyze the spectrogram as an image to learn patterns [157]. Furthermore, the Mel Spectrogram is a common intermediate step for deriving more compact feature sets, such as MFCCs.

Building on this perceptual model, Mel-Frequency Cepstral Coefficients (MFCC) have become a standard feature extraction method in speech processing systems due to their ability to effectively represent spectral properties. The technique employs a perceptually-motivated frequency analysis using the Mel scale, which closely corresponds to human hearing sensitivity. Unlike time-domain features, MFCCs better

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE IV: Acoustic Features Used in Infant Cry Analysis Studies

Feature	Study
Short-Time Fourier Transform (STFT)	[13], [14], [22], [49], [67], [157]
Mel Spectrogram	[59], [157]
Mel-Frequency Cepstral Coefficients (MFCC)	[11], [12], [23], [30]–[32], [34], [35], [37]–[39], [43], [48], [50], [50], [57], [58], [60], [65], [66], [75], [84]–[90], [156], [157]
DWT-MFCC	[15], [48], [50], [72]
Linear Predictive Coding (LPC)	[12], [28], [32], [37]
Perceptual Linear Prediction (PLP)	[51], [90]
Bark Frequency Cepstral Coefficients (BFCC)	[87], [91]
Gamma-Tone Frequency Cepstral Coefficients (GFCC)	[60], [92]
Short-Time Cepstral Coefficients (STCC)	[31]
Variation of Waveforms	[40], [41]
Log Mel Filter Bank (LMFB)	[93]
Log Linear Filter Bank (LLFB)	[26]
Log Mel Filter Band Energy	[94]
Fundamental Frequency (F_0)	[77], [95]
FFT + Fractal Dimension with Higuchi Algorithm	[52]
Wavelet Packet Spectrum	[15]
Spectrum-Based Features	[74]
Spectrogram-Based Features	[67], [121]
Harmonic Ratio	[92]
Frequency Features	[21], [61]

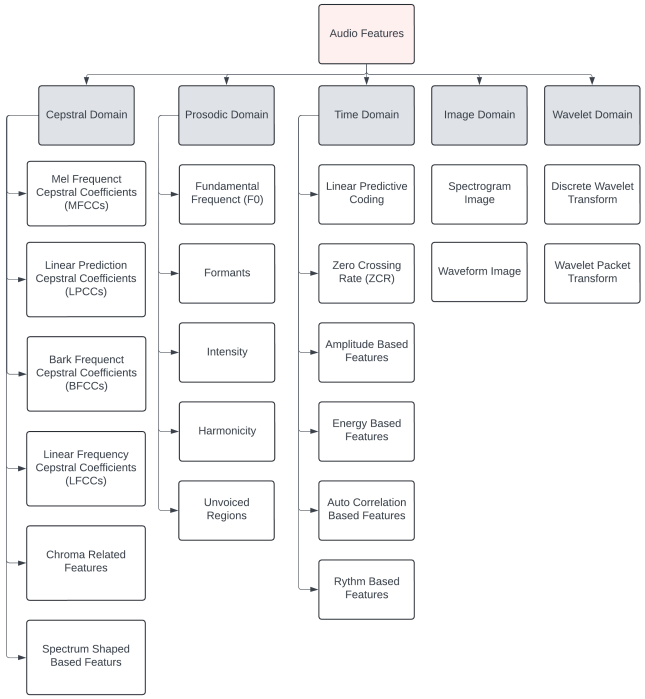

Fig. 4: Main audio feature categories

capture perceptually significant spectral information, making them particularly suitable for voice analysis applications [48]. MFCC captures a sound signal's short-term power spectrum using the Mel scale, which approximates human auditory perception of frequency changes. The method processes the

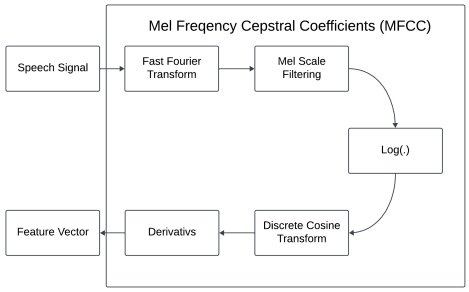

Fig. 5: MFCCs Block Diagram

signal by: (1) segmenting it into short frames, (2) analyzing each frame's power spectrum, and (3) weighting energy across frequency bands according to human hearing sensitivity. The frequency scale undergoes logarithmic compression before Discrete Cosine Transform (DCT) application, producing the final MFCC coefficients [57], as shown in Figure 5. The advantages of MFCC, which have caused it to be used by researchers more than other features, include [34], [83]:

- Effectively characterizes acoustic patterns by identifying distinctive sound features
- Produces compact feature vectors while preserving essential acoustic properties
- Mimics human auditory perception in its signal processing approach

C. DWT-MFCC

Using multilevel analysis, the Discrete Wavelet Transform (DWT) breaks down signals into frequency sub-bands, generating wavelet components at each step of the decomposition process. Quadrature mirror filters, more especially a low-pass filter (LPF) for approximation coefficients and a high-pass filter (HPF) for detail coefficients, are used to accomplish this transition [81]. DWT is a useful preprocessing method for signal decomposition and feature extraction in audio processing applications [50].

In the DWT-MFCC approach, this decomposition is the first step. After the DWT separates the signal into its constituent wavelet components, Mel-Frequency Cepstral Coefficients (MFCCs) are then extracted from these individual sub-bands [15]. This combined method is particularly effective for analyzing non-stationary signals like infant cries, as the initial wavelet decomposition can better capture transient acoustic events before the cepstral analysis is performed [48].

D. Other Features

Other types of cepstral domain features that have applications in baby cry analysis, such as Linear Predictive Coding(LPC) [12], [28], [32], [37], Linear Prediction Cepstral Coefficients(LPCC) [30], [46], [51], Perceptual Linear Prediction (PLP) [51], [90], Linear Frequency Cepstral Coefficients (LFCC) [28], [31], Bark Frequency Cepstral Coefficients (BFCC) [87], [91], Gamma-Tone Frequency Cepstral

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

Coefficients(GFCC) [60], [92] and Short Time Cepstral Coefficients (STCC) [31]. Some researchers used the variation of waveforms [40], [41], pitch [35], Log Mel Filter Bank (LMFB) [93], Log Linear Filter Bank (LLFB) [26], Log Mel Filter Band Energy [94]. Chittora et al. [95] analyzed fundamental frequency(F0) for classification but resulting in a computational cost 20 times lower than standard MFCCs with no additional memory cost [96].

VII. Models

Reliable results can be achieved by implementing an effective model, provided that suitable acoustic features are extracted. Accuracy and latency are directly impacted by the choice of model. The reviewed models can be broadly categorized by their primary approach: traditional models that classify extracted feature vectors, and artificial neural networks that can learn patterns directly from signal representations, either spatially as images or temporally as sequences.

A. Traditional Machine Learning Models

Researchers frequently employ traditional machine learning models such as Support Vector Machines (SVMs) [12], [55], K-Nearest Neighbors (KNN) [23], [57], and Random Forest (RF) [58], [63]. These models typically operate on extracted feature vectors, like MFCCs, treating each vector as a single, static data point. Their strength lies in their simplicity and efficiency for classification when the input features are highly discriminative. However, they do not inherently model the temporal evolution or modulations within the cry signal itself, instead relying on the feature vector to encapsulate all necessary information.

B. Artificial Neural Networks

Artificial Neural Networks (ANNs), computational models inspired by biological brain systems, have been central to modern cry analysis [33]. A dominant approach within ANNs is to treat a time-frequency representation of the cry, such as a spectrogram, as an image. For this, Convolutional Neural Networks (CNNs), including architectures like AlexNet [40] and GoogLeNet [41], are used to automatically learn important spatial patterns, like the shape and texture of harmonics. This automates the feature learning process from visual data.

Another approach focuses on the cry's modulations over time by treating it as a sequence. Architectures like Recurrent Neural Networks (RNNs) and Reservoir Networks [90] are designed for this purpose, processing sequences of feature vectors to learn patterns that unfold over the duration of the cry. Other widely used ANNs include Feedforward Neural Networks (FFNNs) [35] and specialized architectures like Probabilistic (PNN) and General Regression (GRNN) neural networks [13], [14], which have proven to be powerful classifiers.

C. Hybrid Models

To leverage the respective strengths of different architectures, some researchers have implemented hybrid models. These models combine distinct approaches to create a more powerful and nuanced system. A prime example is the CNN-RNN architecture [49]. In this approach, the CNN acts as an advanced feature extractor on short, sequential frames of a spectrogram. The sequence of these learned features is then fed into an RNN, which models their temporal dependencies. This allows the system to learn both the specific acoustic features present at each moment and the temporal evolution of these features over the duration of the cry. Other hybrid models combine neural network feature extraction with traditional classifiers, such as in DeepSVM [62], or integrate principles from other fields, as seen in Neuro-Fuzzy [51] and Genetic-ANN systems [32].

TABLE V: Model and Studies

Model	Study
Support Vector Machine (SVM)	[12], [15], [19], [22], [39], [43], [55], [77]
K-Nearest Neighbor (KNN)	[23], [28], [34], [56], [57], [88], [89]
Random Forest (RF)	[58], [60], [63], [77]
Gaussian Mixture Model (GMM)	[31], [61]
i-vector	[65]
Linear Discriminant Analysis (LDA)	[50]
Fuzzy	[37]
Feed Forward Neural Network (FFNN)	[11], [33], [35], [94]
Convolutional Neural Network (CNN)	[24]–[27], [38], [53], [54], [78], [92], [93], [157]
Multi-Layer Perceptron (MLP)	[30], [43], [44], [75], [76]
Capsule Network	[67], [84]
CNN-RNN	[49]
Graph Convolutional Network (GCN)	[121]
Genetic + ANN	[32]
Vision Transformer	[59]
General Regression Neural Network (GRNN)	[13], [36]
Probabilistic Neural Network (PNN)	[14], [46]
DeepSVM	[62]
AlexNet	[40], [41]
GoogLeNet	[41]
Neuro-Fuzzy	[51], [102]
Reservoir Network	[90]

VIII. Result

For metric evaluation, this review uses several standard metrics, including Accuracy, Precision, Recall, and F1-Score. The formulas are provided below [106]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

where:

- True Positive (TP): Correct identification of vocal samples belonging to the target class.
- True Negative (TN): Accurate exclusion of samples from non-target classes.
- False Positive (FP): Erroneous classification of non-target samples into the target class.
- False Negative (FN): Incorrect rejection of valid target-class samples.

The analysis first addresses the binary task of cry detection before moving to the more complex challenge of multi-class cry classification across several key datasets.

Cry Detection (Binary Classification)

The initial analysis focuses on cry detection (a cry vs. no-cry task), detailed in Table VI. The primary finding is that cry detection is a largely mature problem where simple, computationally efficient models can achieve near-perfect accuracy. A standout example is the work of Cohen [23], which reached 100% accuracy using a lightweight K-Nearest Neighbor (KNN) model with MFCC features, making it ideal for real-time applications. In contrast, while more complex models like the CNN used by Chang [27] also achieved very high accuracy (99.83%), their greater computational cost makes them better suited for systems where latency is less critical. This establishes a clear theme: for the simpler task of detection, efficiency is a key differentiator among high-performing models.

Cry Classification (Multi-Class)

Following detection, the more complex challenge of multi-class cry classification is evaluated across several key datasets.

Analysis of the Baby Chillanto Dataset: On the Baby Chillanto dataset, which is focused on classifying pathological cries, the results in Table VII reveal a clear synergy between the choice of audio features and the model architecture. The selection of features proved critical for success. While the commonly used Mel-Frequency Cepstral Coefficients (MFCC) provided a versatile baseline, its effectiveness varied significantly, contributing to results that ranged from a low accuracy of 85% [39] to a high precision of 98.67% [11].

The highest and most consistent accuracies were achieved when models utilized more detailed time-frequency features, such as the Short-Time Fourier Transform (STFT) [13], [42] and Wavelet Packet Spectrum (WPS) [14], [36]. This advantage is most evident in the performance of specialized neural networks. General Regression Neural Networks (GRNN) and Probabilistic Neural Networks (PNN), when paired with these advanced time-frequency features, achieved state-of-the-art accuracies exceeding 99% [13], [14], [36], [42]. This level of performance surpassed not only traditional models like SVM, which had a wide accuracy range from 85% [39] to 98.5% [15], but also standard deep learning architectures like AlexNet (92%) [40]. Therefore, for the Baby Chillanto dataset, the combination of detailed time-frequency feature extraction with specialized PNN or GRNN models represents the most robust and practical solution.

Analysis of the Donate a Cry Corpus: Analysis of the Donate a Cry Corpus (Table VIII) challenges the prevailing assumption that state-of-the-art performance necessitates complex deep learning models. A lightweight K-Nearest Neighbors (KNN) classifier, operating on standard MFCC features, not only achieved an exceptional accuracy of 98.8% [57] but also outperformed more computationally intensive architectures like Vision Transformers (98.33%) [59] and Deep-SVM (98.34%) [62]. This finding underscores a critical principle for this domain: for classifying common, non-pathological cries, the discriminative power of a robust feature set can be more impactful than the model's architectural complexity. Consequently, computationally efficient traditional models present a more practical, and in this case, more accurate solution for deployment on real-time, resource-constrained consumer devices.

Analysis of the Dunstan Baby Language Dataset: The analysis of the Dunstan Baby Language dataset (Table IX) reveals two distinct and competing pathways to achieving high classification accuracy. The highest performance was achieved by a complex hybrid CNN-RNN model, which reached 94.97% accuracy [49]. This result underscores the importance of modeling temporal dynamics for these specific, phonetic-based cries.

Competing closely, a much simpler Linear Discriminant Analysis (LDA) model achieved a comparable 94% accuracy [50]. This presents a clear trade-off for practical implementation: while the hybrid model likely offers the highest precision, the efficient LDA provides a more balanced solution with a reported latency of just 1.55 ms. The relative underperformance of other traditional models, such as KNN (79.95%) [34] and SVM (80%) [52], suggests that this classification task is more complex than general cry detection and requires either sophisticated temporal modeling or strong linear feature separation.

Analysis of Other Specialized Datasets: Finally, the varied results from the smaller, specialized clinical datasets in Table X underscore both the potential and the current limitations of diagnostic cry analysis. While studies on the extremely small ADEL dataset reported 100% accuracy using complex feature selection methods [73], [74], these results should be interpreted with significant caution as they carry a high risk of overfitting and are unlikely to generalize.

More indicative are the promising results for specific conditions, such as the ˜88% accuracy for Hypothyroidism detection using MLP models [75], [76] and up to 92.33% for Autism screening with a CNN [72]. These findings demonstrate the feasibility of using cry analysis as a non-invasive diagnostic tool, but they also highlight the field's most critical challenge: the lack of large, diverse clinical datasets required to validate these models and move them from proof-of-concept to reliable clinical application.

IX. Discussion

The analysis of the surveyed literature reveals a field of dual speeds. The task of binary cry detection (a cry vs. no-cry problem) appears largely mature, with simple, lightweight

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE VI: Comparison of Various Detection Methods

Author	Feature	Model	Accuracy	Precision	Recall	F1
Chang (2019) [27]	Spectrogram (STFT)	CNN	99.83%	-	-	-
Cohen (2012) [23]	MFCC	KNN	100%	-	-	-
Dewi (2019) [28]	LFCC	KNN	90.83%	-	-	-
Jahangir (2024) [24]	-	CNN Stacked Classifier Network	98%	98.72%	98.05%	98.39%
Manikanta (2019) [25]	MFCC	1D-CNN	98%	-	-	-
Xie (2019) [26]	LLFB	CNN	-	99%	-	-

TABLE VII: Baby Chillanto Database

Author (Year)	Number of Classes	Feature	Model	Accuracy	Precision	Recall	F1
Reyes-Galaviz (2004) [11]	3	MFCC, LPC	FF-IDNN	-	98.67	-	-
Sahak (2010) [43]	2	MFCC	SVM	95.86	97.44	94.28	95.83
Zabidi (2010) [44]	2	MFCC	MLP	93.38	-	90.15	-
Zabidi (2010) [45]	2	MFCC	OLS+MLP	93.95	-	-	-
Hariharan (2012) [42]	2	STFT	PNN	99.00	-	98.82	-
Hariharan (2012) [13]	2	STFT	GRNN	99.00	-	98.24	-
Saraswathy (2013) [14]	3	WPS	PNN	99.22	-	-	-
Perez (2015) [37]	2	MFCC+LPC	Fuzzy	97.96	100	95.00	97.40
Sachin (2017) [40]	2	Waveforms	AlexNet	92.00	-	-	-
Moharir (2017) [41]	3	Waveforms	GoogLeNet	94.00	-	-	-
Onu (2017) [39]	3	MFCC	SVM	85.00	85.00	85.00	85.00
Zabidi (2017) [38]	2	MFCC	CNN	92.78	-	-	-
Badreldine (2018) [15]	3	DWT-MFCC	SVM	98.50	-	-	-
Sailor (2018) [16]	2	ConvRBM-FB	GMM	96.48	-	-	-
Saraswathy (2020) [36]	2	WPS	GRNN	99.70	99.41	100	99.70
Ji (2021) [121]	3	Spectrogram	GCN	95.21	-	-	-
Dharwadkar (2022) [35]	5	MFCC + Pitch	FFNN	97.00	-	-	-

TABLE VIII: Donate Cry Corpus Database

Author (Year)	Number of Classes	Feature	Model	Accuracy	Precision	Recall	F1
Ekinci (2023) [57]	5	MFCC	KNN	98.80	98.80	98.80	98.80
Rezaee (2024) [62]	5	STFT	Deep-SVM	98.34	98.35	98.34	98.34
Younis (2024) [59]	5	Mel Spectrogram	ViT	98.33	98.40	98.33	98.34
Ozcan (2025) [156]	5	MFCC	FFNN	90.00	91.80	92.20	90.00
Hammoud (2024) [58]	5	MFCC	Random Forest	96.39	96.48	96.39	96.38
Rani (2022) [56]	8	MFCC	KNN	76.16%	-	-	-
Grayson (2021) [55]	5	Spectrogram	SVM	86.00	-	-	-
Kulkarni (2021) [60]	4	MFCC + GFCC	Random Forest	84.00	84.00	84.00	84.00
Qiao (2024) [157]	3	STFT + Mel Spectrogram + MFCC	CNN	93.83	-	-	-
Sharma (2019) [61]	5	12 Features	GMM	81.27	-	-	-

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE IX: Dunstan Baby Language Database

Author	Number of Classes	Feature	Model	Accuracy	Latency (ms/sample)
Limantoro (2016) [34]	4	MFCC	KNN	79.95%	-
Prasasti (2019) [48]	4	DWT-MFCC	Euclidean	90%	0.5542
Novamizanti (2020) [50]	4	MFCC	LDA	94%	1.5506
Srijiranon (2014) [51]	All	PLP	Neuro-Fuzzy	86.25%	-
Widhyanti (2021) [52]	4	FFT+Fractal	SVM	80%	0.03–0.08
Sutanto (2021) [53]	All	MFCC	CNN	85%	-
Franti (2018) [54]	All	Spectrogram	CNN	89%	-
Maghfira (2020) [49]	All	STFT	CNN-RNN	94.97%	-
Qiao (2024) [157]	All	STFT + Mel Spectrogram + MFCC	CNN	87.76%	-

TABLE X: Other Databases

Author	Database	Feature	Model	Accuracy	Precision	Recall	F1	Specificity	AUC / ROC
Khozaei (2020) [71]	Autism DB(2020)	F0 + MFCC	SubSet Instance	85.7%	92.86%	85.71%	89.2%	100%	–
Wu (2019) [72]	Autism DB(2019)	MFCC	CNN	92.33%	–	-	–	-	–
Okada (2024) [73]	ADEL	LPC	IFSM+DPS	100%	–	–	–	–	–
Wang (2010) [74]	ADEL	Power Spectrogram	FSM	100%	–	–	–	–	–
Zabidi (2009) [76]	Hypothyroid	MFCC	F-Ratio + MLP	88.35%	–	98	–	–	-
Zabidi (2010) [75]	Hypothyroid	MFCC	MLP	88.12%	–	100	–	–	AUC = 99.89%
Orlandi (2016) [77]	Self	Feature Frequency	Random Forest	87%	87.4%	87.3%	87.3%	87.4%	ROC = 94.1

models like KNN achieving near-perfect accuracy with minimal latency [23], making them ideal for real-time triggering. This success shifts the primary research challenge towards the more complex, multi-class problem of cry classification interpreting the cry's meaning. A primary promise in this area is its potential as a non-invasive tool for screening neurodevelopmental and pathological conditions. The success of this application hinges on identifying consistent, measurable acoustic biomarkers specific features of a cry that reliably correlate with an underlying state. The following subsections first explore these foundational biomarkers and then synthesize the performance of various models, revealing a critical, multi-dimensional trade-off that governs the selection of an optimal approach in practice.

A. Acoustic Biomarkers in Pathological Cry Analysis

The models discussed in this survey achieve high accuracy not by magic, but by learning to detect subtle deviations from a healthy cry pattern. The validation for these biomarkers is typically established through rigorous statistical analysis, where the acoustic features from a clinically diagnosed group of infants are compared to those from an age-matched, healthy control group to identify statistically significant differences.

1) Asphyxia: Birth asphyxia, a condition caused by oxygen deprivation, has a direct and measurable impact on the vocal apparatus. The resulting hypoxia causes increased muscle tension in the larynx, leading to several distinct acoustic markers [13]. Studies have consistently shown that cries from asphyxiated infants exhibit a significantly higher and more unstable fundamental frequency (F_0), often exceeding 1000 Hz, compared to the typical 400-600 Hz range of healthy newborns [11]. Furthermore, due to strained respiration and poor physiological control, these cries are often characterized by shorter utterance durations, longer pauses between cry sounds, and a more strained or "hyperphonated" quality.

2) Autism Spectrum Disorder (ASD): While ASD is diagnosed later in childhood, research suggests that atypical vocal patterns may be present in early infancy. These differences are thought to stem from underlying variations in neurological development that affect vocal motor control. Studies using the Autism DB datasets have identified several potential biomarkers [71], [72]. Cries from infants who are later diagnosed with ASD have been found to have a higher mean F_0 and greater pitch variability than neurotypical infants. These cries can also exhibit an atypical prosody, or melody, lacking the smooth, rising-and-falling contour of a typical cry. Some studies also

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

report a longer latency a greater delay before the infant begins to cry in response to a stimulus.

B. Connecting Biomarkers to Model Performance

Understanding these underlying acoustic biomarkers is crucial, as it provides the scientific context for evaluating how and why certain machine learning models excel. The effectiveness of a model is directly tied to its ability to detect these specific, often subtle, acoustic signatures. The following analysis examines model performance across various datasets through this lens.

C. Performance Across Datasets

Comparative performance on a variety of datasets—i.e., the Baby Chillanto [13], [36], Donate A Cry Corpus [55], [62], and Dunstan Baby Language [50], [56]—depicts the effect of dataset characteristics, such as complexity and number of cry classes, on classification model performance. The following were noted:

- **Baby Chillanto Dataset**: On this dataset focused on pathological cries (Table VII), the choice of audio features is critical. While MFCC provided a versatile baseline [11], [39], the highest accuracies were consistently achieved with detailed time-frequency features like STFT and Wavelet Packet Spectrum. This is because these features effectively capture the specific acoustic biomarkers of asphyxia. For instance, GRNN and PNN models achieved accuracies exceeding 99% on binary classification tasks by detecting the high-frequency (F_0) and short-duration patterns characteristic of asphyxiated cries [13], [36], [42]. However, this performance landscape is nuanced by task granularity. The work by Dharwadkar et al. [35] demonstrates a move towards more complex diagnostics, successfully classifying five distinct cry types with 97% accuracy. Meanwhile, the SVM model from Badreldine et al. [15] offers a compelling middle ground, achieving 98.5% accuracy on a 3-class problem with greater computational efficiency. Ultimately, the findings for the Baby Chillanto dataset reveal a multi-dimensional trade-off. The choice of model is not just between accuracy and latency, but also involves the desired level of diagnostic granularity. The optimal solution depends on whether the goal is to achieve near-perfect accuracy on a focused binary task or to perform a more complex, multi-class analysis, all while considering the practical constraints of real-time deployment.
- **Donate A Cry Corpus**: On the crowdsourced Donate A Cry Corpus (Table VIII), which focuses on classifying common infant needs, the analysis reveals a compelling dynamic where lightweight, traditional models are exceptionally competitive with complex deep learning architectures. The most striking result is that a simple K-Nearest Neighbors (KNN) model achieved the highest reported accuracy of 98.80% [57]. This performance met or exceeded that of far more sophisticated architectures, including a Vision Transformer (ViT) (98.33%) [59] and a Deep-SVM (98.34%) [62]. This suggests that for non-pathological cries, the choice of robust features, which ranged from the efficient MFCC to more complex STFT and Mel Spectrograms is as critical as the complexity of the model itself. This finding has significant practical implications related to the trade-off between accuracy and computational efficiency. Although no latency figures were reported for these studies, the KNN model is well-known for its extremely fast inference speed. Its ability to deliver state-of-the-art accuracy makes it an optimal choice for real-time applications on low-power consumer devices like smart monitors. Conversely, the computationally intensive ViT and Deep-SVM architectures, while equally accurate, are better suited for offline or cloud-based analyses where latency is less of a constraint. Therefore, the results from this dataset provide a crucial insight: for certain classification tasks, resource-intensive deep learning is not a prerequisite for achieving top-tier performance, and a well-chosen lightweight model can offer a superior balance of accuracy and practical deployability.
- **Dunstan Baby Language**: For the Dunstan Baby Language dataset (Table IX), which is based on distinguishing specific phonetic-based cries, the analysis reveals two distinct and competing pathways to achieving high accuracy. The highest performance was achieved by a complex hybrid CNN-RNN model, which reached 94.97% accuracy [49]. This result highlights the importance of modeling both spatial and temporal dependencies in the cry signal, a highly successful strategy for this specific phonetic task. Competing closely, a much simpler traditional model, Linear Discriminant Analysis (LDA), achieved a comparable 94% accuracy using standard MFCC features [50]. This presents a clear trade-off between maximizing precision and ensuring practical deployability. While the complex CNN-RNN likely offers the highest accuracy, its computational cost can be inferred to be significantly higher than the 1.55 ms latency reported for the LDA model. This makes the LDA a more balanced and efficient solution for real-time applications. The relative underperformance of other simple models like KNN and SVM, with accuracies around 80% [34], [52], suggests that this classification challenge is more complex than other non-pathological tasks, requiring either sophisticated temporal modeling or strong linear feature separation. Therefore, for the Dunstan dataset, the choice between a complex hybrid model and an efficient traditional one is a strategic decision dictated by the application's specific requirements for precision versus real-time performance.
- **Other Specialized Datasets**: The analysis of smaller, specialized datasets presented in Table X highlights both the potential of cry analysis for specific clinical diagnostics and the significant limitations imposed by data scarcity. For instance, studies on the extremely small ADEL dataset reported 100% accuracy using complex iterative feature selection methods [73], [74]. However, such perfect scores on a tiny dataset are a strong indicator of overfitting, meaning the models are unlikely

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

to generalize to new data. Furthermore, these iterative methods are computationally expensive and ill-suited for real-time use. More indicative are the promising, though varied, results for specific clinical conditions. For Hypothyroidism, MLP models consistently achieved a solid ˜88% accuracy while also demonstrating extremely high recall (98-100%) [75], [76]. This is clinically significant, as high recall ensures that very few true cases are missed. In Autism detection, the results reveal a critical trade-off. A CNN model achieved a high raw accuracy of 92.33% by learning visual patterns from spectrograms, while a Subset Instance Classifier provided a more clinically robust profile (85.7% accuracy, 92.86% precision) by focusing on specific features [71], [72]. The success of both approaches relies on their ability to detect the underlying biomarkers of ASD, such as atypical prosody and greater pitch variability. Finally, the work by Orlandi et al. [77] serves as an example of a model achieving balanced performance across all metrics on a self-recorded dataset. These findings demonstrate the clear feasibility of using cry analysis as a non-invasive diagnostic tool. However, the primary takeaway is that the field is critically constrained by the lack of large-scale, validated clinical datasets. The impressive headline accuracies reported in these studies must be interpreted with significant caution, as they represent important proofs-of-concept rather than clinically validated tools ready for deployment.

In summary, the analysis across these diverse datasets reveals that there is no single best model for infant cry classification. Instead, the optimal choice is a nuanced decision dependent on the specific task, dataset characteristics, and practical application constraints. For high-stakes pathological cry detection, as seen in the Baby Chillanto dataset, specialized neural networks like PNN and GRNN offer the highest accuracy [13], [36], but highly-tuned traditional models like SVM present a compelling and efficient alternative [15]. Conversely, for classifying common infant needs on datasets like the Donate A Cry Corpus, lightweight models such as KNN can achieve state-of-the-art results [57], challenging the assumption that more complex deep learning models are always necessary for top performance.

This highlights a critical trade-off between performance and practicality. While complex hybrid models like the CNN-RNN can capture intricate temporal patterns on phonetically-complex datasets like Dunstan Baby Language [49], simpler and faster models like LDA often provide a more balanced solution for real-time deployment [50]. Ultimately, the selection of a model is a strategic compromise between maximizing accuracy, ensuring low latency for resource-constrained environments, and matching the model's complexity to the granularity of the classification task. To further aid researchers and practitioners in this selection process, a comparative framework is presented in Table XI.

D. Feature Extraction Techniques

The success of any infant cry classification system is critically dependent on feature extraction, which transforms raw audio into a format that machine learning models can interpret. The surveyed literature reveals a clear progression from foundational, perceptually-based coefficients to more complex time-frequency representations, each with distinct advantages.

The Dominance of Perceptually-Based Features: Across all datasets and model types, Mel-Frequency Cepstral Coefficients (MFCC) stand out as the most widely used and consistently effective feature set [43], [50], [57]. While originally designed for speech recognition [79], their success in cry analysis stems from their ability to mimic the non-linear frequency perception of the human ear [83]. This allows MFCCs to effectively capture the same acoustic cues a human caregiver uses to interpret a cry, making them a powerful and efficient choice for both traditional models like SVM [15] and deep learning architectures [72]. Other perceptually-motivated features, such as Perceptual Linear Prediction (PLP) [51], have also shown strong performance, particularly in capturing the subtle variations required for phonetic-based classification tasks like the Dunstan Baby Language.

Advanced Time-Frequency Representations: For more complex diagnostic tasks, particularly on the Baby Chillanto dataset, advanced time-frequency representations have proven superior. Techniques like the Short-Time Fourier Transform (STFT) [13] and Wavelet Packet Spectrum (WPS) [36] provide a much richer, more detailed view of the cry signal's dynamics over time. This additional information is crucial for identifying the subtle acoustic biomarkers associated with pathological conditions, enabling specialized neural networks to achieve near-perfect accuracy where simpler features might fail.

The Rise of Spectrograms as Image Features: A significant trend, especially with the rise of deep learning, is the treatment of audio features as images. Mel Spectrograms and STFT Spectrograms are now commonly used as direct inputs for Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [59], [62]. This approach bypasses the need for hand-engineered feature selection and allows the model to learn the most salient patterns directly from the time-frequency representation. This has proven highly effective, with models achieving state-of-the-art results on datasets like the Donate A Cry Corpus by treating the cry signal as a visual pattern recognition problem.

In conclusion, the surveyed literature demonstrates that the choice of feature extraction is not just a technical step but a strategic decision that dictates the balance between computational efficiency and diagnostic precision. For applications where real-time performance on resource-constrained devices is paramount, perceptually-based features like MFCCs offer an optimal blend of compact representation and strong performance. For high-stakes clinical applications, where capturing the most subtle acoustic biomarkers is critical, the superior resolution of advanced time-frequency representations like STFT and WPS is non-negotiable, as they unlock the full potential of specialized neural networks. Finally, the trend of using spectrograms as image inputs for deep learning models represents a paradigm shift, moving from hand-crafted feature engineering to automated pattern recognition. The optimal path forward will likely involve hybrid approaches, but for now,

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE XI: An Enhanced Comparative Framework for Key Feature-Model Combinations in Infant Cry Analysis

Feature-Model Combination	Primary Strength	Accuracy Potential	Computational Cost	Noise Robustness	Recommended Use Case
STFT/WPS + GRNN/PNN	**Peak Accuracy**	Very High	Low to Medium	Medium	High-precision clinical diagnostics for pathological cries (e.g., asphyxia) in controlled, low-noise environments.
MFCC + KNN/SVM	**Efficiency**	High	Very Low	Low to Medium	Real-time classification of infant needs on resource-constrained devices (e.g., smart baby monitors, mobile apps) where speed is critical.
Spectrogram + CNN/ViT	**Robustness**	Very High	Very High	High	Offline or cloud-based analysis where high accuracy is needed in noisy, real-world environments and computational power is not a constraint.
MFCC + LDA	**Balanced Performance**	High	Low	Medium	Real-time classification of phonetic-based cries (e.g., Dunstan Baby Language) where a balance of good accuracy and efficiency is required.
DWT-MFCC + SVM	**Noise Filtering**	High	Low to Medium	High	Applications in highly variable or noisy environments where preprocessing and feature robustness are more important than peak accuracy.
STFT + CNN-RNN	**Temporal Analysis**	Very High	High	High	Complex classification tasks requiring analysis of how cry patterns evolve over time, such as distinguishing phonetic-based cries.

the selection of a feature extraction technique remains one of the most critical, application-dependent decisions in infant cry analysis.

E. Model Complexity vs. Computational Efficiency

A central theme in infant cry classification is the trade-off between model complexity and computational efficiency, where state-of-the-art accuracy often comes at the cost of higher computational demand. The surveyed literature reveals a clear performance hierarchy and distinct use cases for different model families.

The High Performance of Neural Architectures: Advanced neural architectures, including specialized models like GRNNs and PNNs [13], [42], as well as deep learning models like CNNs and ViTs [59], consistently deliver state-of-the-art performance. These models excel at learning complex, hierarchical patterns directly from data, making them particularly effective for high-stakes diagnostic tasks. Hybrid models, such as the CNN-RNN [49] and Deep-SVM [62], further leverage the strengths of different architectures to achieve high accuracy on challenging datasets. However, this performance comes with significant computational intensity, making these models better suited for offline analysis or environments with ample computational resources. While techniques like model pruning and fine-tuning can mitigate these costs [20], the inherent complexity of neural networks remains a key consideration for real-time deployment.

The Efficiency of Traditional Machine Learning Models: In contrast, traditional machine learning models like K-Nearest Neighbors (KNN) [57], Support Vector Machines (SVM) [15], and Linear Discriminant Analysis (LDA) [50] offer a compelling balance of performance and efficiency. While they may be less accurate on the most complex, multi-class problems, they often achieve highly competitive, and in some cases superior, results on tasks like binary detection [23] and the classification of non-pathological cries [57]. Their primary advantage is their low computational overhead and fast inference speed, which makes them the ideal choice for resource-constrained environments, such as real-time embedded systems or mobile applications, where low latency is a critical requirement.

In conclusion, the choice of model architecture is a strategic decision dictated by the specific needs of the application. For tasks demanding the highest possible accuracy, such as clinical diagnostics, the computational cost of advanced neural and hybrid models is a worthwhile investment. However, for applications where real-time performance and efficiency are paramount, traditional machine learning models offer a robust and often equally effective solution. The optimal approach, therefore, involves a careful balancing of these competing priorities, tailored to the specific context of deployment.

X. Domains and Applications

Infant vocalization detection and classification offer enormous potential to serve the diverse needs of infants and advance methods of care. If this technology can successfully analyze and interpret the sounds of infants, it will be a useful tool in helping caregivers to understand and respond to a variety of needs especially those regarding development and general health. Some key areas and applications where this technology may make a difference, though not limited to them, include:

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

A. Health Monitoring and Initial Diagnosis

Infant crying is a crucial indicator for health monitoring and early diagnosis. Specific cry patterns can signify issues like colic, respiratory distress, or neurological problems, thus enabling timely intervention and treatment [113], [121], [122]. Machine learning algorithms have proven capable of classifying these patterns with high accuracy to identify potential health concerns. For instance, a deep learning model by Ji et al. (2021) reached 94.39% accuracy in detecting distress signals like hunger and pain [121]. Such a tool can provide caregivers with immediate feedback to differentiate cry reasons, thereby reducing their stress and improving responsiveness to the infant's needs.

B. Developmental Milestone Tracking

Tracking developmental milestones through vocalization analysis is increasingly recognized as vital for early intervention. Studies have shown that deviations from typical vocalization patterns may signal delays in speech or cognitive development [114], [123]. Machine learning models have been applied to large datasets of infant vocalizations to classify sounds based on age-related changes, providing insights into normal developmental trajectories. For example, a longitudinal study involving typically developing infants demonstrated clear clustering patterns in vocalizations that indicated active engagement in vocal exploration and category formation, foundational for language development [116].

C. Caregiver Support and Resources

Advanced technologies are being developed to help caregivers interpret infant cries. By differentiating between different types of cries, such as those that indicate hunger or discomfort, these systems can provide immediate feedback, reducing caregiver stress [127]. Research shows that responsive caregiving is crucial for promoting healthy emotional and cognitive development in infants [114]. Real-time feedback tools can greatly improve caregivers' comprehension and responsiveness to their infants' needs.

D. Neonatal Intensive Care Units NICUs

In NICU settings, where infants require constant monitoring, vocalization analysis can serve as a vital complement to existing monitoring systems. Technologies such as the Newborn Cry-based Diagnostic System (NCDS) monitor not only breathing but also detect larger movements and crying sounds, thereby enhancing the continuous observation of an infant's condition [115]. Studies indicate that preterm infants exhibit a greater responsiveness to human voices compared to non-biological sounds, suggesting that integrating parental voice interactions could considerably benefit their development [124]. Additionally, an analysis of NICU noise levels found that conversations and infant cries significantly contribute to the overall sound environment, potentially influencing linguistic outcomes for infants [125].

E. Speech and Language Development Research

Research into infant vocalizations is essential for understanding the mechanisms of speech and language development. Studies employing deep learning techniques have classified various types of vocalizations, uncovering patterns associated with later language acquisition [116]. These analyses enable researchers to identify universal tendencies in early vocalizations, offering insights into strategies for fostering communication skills from infancy.

F. Early Autism Spectrum Disorder (ASD) Detection

Early indicators of ASD are often atypical vocalizations. Improving long-term results requires earlier diagnoses and actions, which can be made possible by automated systems built to examine these patterns [71]. The potential of recent developments in machine learning to identify small differences in vocalization patterns linked to ASD highlights the importance of technology in supporting early detection efforts [121].

G. Innovative Smart Parenting Tools

Smart parenting tools, such as intelligent baby monitors and mobile applications, can leverage vocalization analysis to provide parents with instant feedback, alerts, and care suggestions [117]–[119]. Early implementations often utilized low-cost hardware like the Raspberry Pi for cry detection [20], [120]. While numerous mobile applications for cry detection and classification emerged by 2024, a primary limitation of many existing devices is their focus on simple detection rather than classification. Integrating classification capabilities could significantly improve their utility. Furthermore, enhancing usability through connectivity features—such as Wi-Fi notifications to mobile apps [120] or wearable devices like smartwatches—remains a promising direction for providing caregivers with more convenient and timely alerts.

XI. Challenges and Possible Solutions

Through the use of machine learning, deep learning, and sophisticated signal processing techniques, recent advancements in infant cry analysis have shown encouraging results [128], [129]. These methods have shown remarkable results in tasks involving the detection and classification of cries. However, a number of enduring restrictions still limit their usefulness and practical application. For the field to advance, several obstacles must be overcome. To support future advancements in newborn cry analysis, this article methodically analyzes these important problems and suggests possible fixes, as shown in Table XII.

A. Data Collection and Privacy Concerns

Due to the sensitivity of the data, there are major challenges in creating representative datasets of infant cries [128], [129].Since most statistics are gathered in controlled settings, including neonatal intensive care units (NICUs), they could not accurately reflect the range of infant cries in natural settings [58], [128].Additionally, ethical and privacy concerns arise when dealing with infant data, making it difficult to share

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

and access datasets [128], [129]. To address such issues, researchers should prioritize defining means to acquire and exchange baby cry data in an ethical and private manner. Approaches such as federated learning, where training models is possible from distributed devices without having raw data, is one such promising path [130]. Further, adding variance to the dataset for representing environment, culture, and health variable cries will provide models more generalizability and robustness [131].

B. Noise Robustness

Infant cry detection and classification systems often struggle with noise robustness, especially in real-world environments where background noise (e.g., household sounds, other voices) can interfere with the accuracy of the models [128], [132]. This is particularly problematic in home settings where smart baby monitors are used [133]. For that reason, training noise-robust models for infant cry classification in noisy environments is required. Data augmentation through noise addition [134], [135], sophisticated noise filtering mechanisms [132], and utilizing deep learning-based models with in-built noise robustness, i.e., convolutional neural networks augmented with attention-based mechanisms [47], [132], can be viable alternatives. Data augmentation methods like white noise addition and sound shifting have also provided excellent improvements in levels of model performance [129].

C. Real-Time Processing and Low Latency

Real-time processing is critical for many infant cry analysis systems, particularly for applications like NICU monitoring and smart baby monitors. However, achieving low latency without sacrificing accuracy is a significant challenge, especially with computationally demanding deep learning models [120], [136], [137]. To provide prompt alerts to caregivers, future work must focus on optimizing models to reduce computational complexity [120], [138]. Low-latency inference on edge hardware, such as smartphones and IoT devices, can be facilitated by methods like model quantization, pruning, and the use of lightweight neural networks (e.g., MobileNet, TinyML) [139], [140]. These techniques have demonstrated significant energy and computational cost savings while maintaining performance, making them vital for deployment on resource-constrained platforms [140], [141].

D. Generalization Across Different Populations

Current models are often trained on datasets that are not representative of the global population, leading to poor generalization across different ethnicities, languages, and cultural backgrounds. This limits the applicability of these systems in diverse settings [142], [143]. Researchers have noted that using diverse data is crucial for developing models with high generalization capability [58]. Future studies should concentrate on creating bigger datasets that include a variety of infant cry patterns from other populations in order to get over these restrictions. Smaller, more specialized datasets could be used to optimize models utilizing transfer learning and domain adaptation techniques, increasing their adaptability under various circumstances. [144], [145].

E. Interpretability and Explainability

Many sophisticated models, particularly deep learning techniques, often operate as 'black boxes', making it difficult for medical professionals to understand their decision-making process [146]. This lack of transparency can hinder practical adoption in clinical settings [147]. Future research should focus on developing interpretable models or applying post-hoc explanation techniques. To clarify cry categorization choices (e.g., differentiating hunger from pain), methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) could be adapted, as these strategies have effectively improved model transparency in similar domains [146], [148], [149].

F. Integration with Multimodal Data

Vocalizations are just one element of an infant's condition. A more comprehensive assessment of infant welfare can be achieved by combining cry analysis with complementary modalities such as facial expressions, mobility patterns, and physiological measurements (e.g., cardiac rhythm) [131], [150]. However, integrating these inherently heterogeneous signal types poses technological challenges [151]. Future work should pursue multimodal investigations, such as deep learning fusion methods, that incorporate audio with other data sources. These techniques can leverage complementary information from different modalities to enhance the accuracy and robustness of classification [151]–[153].

G. Early Detection of Developmental Disorders

While infant cry analysis has shown promise in detecting certain health conditions (e.g., asphyxia, deafness), its application in the early identification of developmental disorders like speech delays or autism spectrum disorder (ASD) requires further research [72], [154]. Current models lack the reliability and diagnostic accuracy for clinical use in these areas and require further validation [71], [72]. Future research should focus on developing models sensitive to the subtle cry variations that may indicate these conditions. This can be advanced through longitudinal studies and collaboration with medical and developmental psychology experts to integrate domain-specific expertise [78], [150], [155].

H. Ethical and Social Implications

The use of infant cry analysis systems raises significant ethical and social implications, particularly regarding over-reliance on technology and its potential impact on caregiver-infant bonding [113], [127]. Such systems, if not carefully integrated, may disrupt the intuitive and emotional connections between caregivers and infants, potentially diminishing the human element of caregiving that is critical for infant development [4]. Additionally, the risk of false positives or negatives in cry analysis could lead to missed diagnoses, inappropriate interventions, or unnecessary stress for caregivers, which may exacerbate anxiety and undermine trust in healthcare systems [23], [115]. Privacy and consent are also critical concerns, as collecting cry data from vulnerable populations,

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

TABLE XII: Challenges and Possible Solutions in Infant Cry Analysis

Challenge	Description	Possible Solutions	Key Techniques	References
Data Collection and Privacy	Collecting diverse datasets is challenging due to privacy/ethical concerns. Datasets often lack real-world variability.	Federated learning; create diverse datasets.	Federated Learning, Data Anonymization	[130], [131]
Noise Robustness	Models struggle with background noise in real-world environments.	Noise augmentation, advanced filtering, robust models.	Data Augmentation, Attention Mechanisms	[132], [134]
Real-Time Processing	Real-time apps need low-latency models, but complex models are computationally expensive.	Model optimization (quantization, pruning, lightweight arch.).	Model Quantization, TinyML	[139], [140]
Generalization	Models fail to generalize across diverse populations.	Inclusive datasets; transfer learning.	Transfer Learning, Domain Adaptation	[144], [145]
Interpretability	Deep learning models are often "black boxes".	Explainable AI (XAI) techniques.	SHAP, LIME	[148], [149]
Multimodal Integration	Cries alone may not provide sufficient information.	Combine audio with other data sources.	Multimodal Fusion	[151], [152]
Early Disorder Detection	Lack reliability for detecting developmental disorders.	Models for subtle variations; medical collaboration.	Longitudinal Studies	[150], [155]
Ethical Implications	Over-reliance may impact caregiver-infant interactions.	Ethical guidelines; study impacts.	Ethical Frameworks	[113], [127]
Standardization	Lack of standardized metrics hinders progress.	Develop benchmarks and metrics.	Standardized Metrics	[29], [38]
Low-Resource Scalability	Advanced models inaccessible in low-resource settings.	Lightweight models; low-cost hardware.	Lightweight Models	[118], [120]

such as infants in neonatal intensive care units (NICUs) or home settings, requires informed parental consent and robust data security measures to prevent unauthorized access or misuse [20]. Furthermore, machine learning models used in cry analysis may inadvertently introduce biases, particularly if trained on non-representative datasets, potentially leading to disparities in diagnostic accuracy across socioeconomic, ethnic, or geographic groups. To address these challenges, future research must prioritize the development of comprehensive ethical guidelines for the implementation of cry analysis systems, ensuring transparency in data collection, processing, and model development [113], [127]. Studies should also explore the long-term effects of these technologies on caregiver behavior, infant development, and family dynamics to ensure they enhance, rather than hinder, the caregiving process [4], [115]. Engaging interdisciplinary ethics committees and community stakeholders can further ensure that these systems are culturally sensitive and equitably applied, fostering trust and promoting inclusive access to the benefits of cry analysis technologies.

I. Standardization of Evaluation Metrics

The lack of standardized evaluation metrics makes it challenging to reliably compare the effectiveness of various models and methodologies, which hinders progress as researchers cannot effectively build upon prior work [20], [23], [29], [38]. Future research should prioritize the development of common benchmarks and assessment measures to permit valid comparisons and establish optimal approaches for the field [29], [38].

J. Scalability and Deployment in Low-Resource Settings

Numerous sophisticated models necessitate substantial computational resources, which may not be accessible in low-resource environments like poor nations or rural hospitals [118], [120]. This limits the scalability and accessibility of these systems, particularly in environments where computational infrastructure is limited [20], [23]. Subsequent work will seek to design low-cost and scalable solutions deployable in low-resource settings. These could take the form of low-cost hardware platforms, cloud processing, or lightweight models of infant cry analysis [118], [120]. These would alleviate limitations on the availability of computational resources and promote greater accessibility within developing countries and rural hospitals [20], [23].

XII. Conclusions and Takeaways from Study

This review provides a comprehensive assessment of the datasets, feature extraction methods, and machine learning models that constitute the field of infant cry analysis. Through a synthesis of over 80 studies, our analysis confirms that machine learning offers powerful tools for both detecting and classifying infant cries with high accuracy. The findings demonstrate a clear and strategic trade-off: for high-stakes clinical diagnostics, specialized neural networks like GRNNs and PNNs, paired with rich time-frequency features that capture subtle acoustic biomarkers, deliver unparalleled precision. Conversely, for real-world consumer applications, lightweight traditional models like KNN and SVM can achieve state-of-the-art accuracy with far greater computational efficiency, particularly when paired with robust perceptual features like MFCCs. This confirms that there is no single "best" approach; rather, the optimal solution is a nuanced decision dictated by the specific application's balance between diagnostic accuracy, task granularity, and deployment constraints.

Beyond the technical achievements, this survey highlights the critical, real-world applications of this technology, from continuous monitoring in Neonatal Intensive Care Units (NICUs) to caregiver support through smart parenting tools and the early screening of developmental disorders like autism. However, the transition from academic proof-of-concept to

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

robust, deployable systems is hindered by significant challenges. The field is fundamentally constrained by a lack of large-scale, diverse, and well-annotated datasets, which not only limits model generalizability but also introduces a serious risk of algorithmic bias, potentially leading to inequitable health outcomes. Furthermore, the practical application of these technologies raises important ethical questions regarding data privacy, informed parental consent, and the potential for over-reliance on technology to mediate the crucial caregiver-infant bond.

To advance the field and translate its potential into tangible clinical and social benefits, future research must address these open challenges head-on. The development of larger, more inclusive datasets through multi-institutional collaboration is the most critical priority to improve model robustness and fairness. Concurrently, there is a pressing need for models that are not only accurate but also computationally efficient, enabling real-time inference on resource-constrained devices. For clinical adoption, the integration of explainable AI (XAI) will be essential to move beyond "black box" predictions and provide interpretable insights such as pinpointing the specific acoustic patterns a model associates with a diagnosis that can build trust among healthcare professionals. Finally, longitudinal studies are required to rigorously evaluate the long-term impact of these technologies on infant development and family dynamics. By addressing these challenges, the field can bridge the gap between laboratory success and real-world impact, unlocking the immense potential of affective computing to revolutionize early childhood healthcare.

REFERENCES

[1] World Health Organization. Newborn Mortality. Available online: https://www.who.int/news-room/fact-sheets/detail/newborn-mortality (accessed on 4 March 2024).

[2] S. Jeyaraman, R. Pandiaraj, J. Y. Jaafar, M. Z. B. Muhammad, R. B. Ngadiran, and N. Shanmugam, "A review: survey on automatic infant cry analysis and classification," *Health and Technology*, vol. 8, pp. 391–404, Springer, 2018.

[3] A. Carollo, R. Montirosso, R. Epifanio, F. Pennestrì, and C. Fedeli, "A Scientometric review of infant cry and caregiver responsiveness: Literature trends and research gaps over 60 years of developmental study," *Children*, vol. 10, p. 1042, Springer, 2023.

[4] M. H. Bornstein, D. L. Putnick, P. Rigo, G. Esposito, J. E. Swain, J. T. Suwalsky, X. Su, X. Du, K. Zhang, and L. R. Cote, "Neurobiology of culturally common maternal responses to infant cry," *Proc. Natl. Acad. Sci.*, vol. 114, pp. E9465–E9473, 2017.

[5] M. Faris and E. McCarroll, "Crying babies: Answering the call of infant cries," *Tex. Child Care*, vol. 34, pp. 14–21, 2010.

[6] S. Lingle, M. T. Wyman, R. Kotrba, J. A. Teichroeb, and C. A. Romanow, "What makes a cry a cry? A review of infant distress vocalizations," *Current Zoology*, vol. 58, pp. 698–726, 2012.

[7] J. D. Newman, "Neural circuits underlying crying and cry responding in mammals," *Behavioural Brain Research*, vol. 182, pp. 155–165, Elsevier, 2007.

[8] D. M. Zeifman, "An ethological analysis of human infant crying: answering Tinbergen's four questions," *Developmental Psychobiology*, vol. 39, pp. 265–285, 2001.

[9] G. Esposito, N. Hiroi, and M. L. Scattoni, "Cry, baby, cry: Expression of distress as a window into the early developing brain," *Neurosci. Biobehav. Rev.*, vol. 83, pp. 376–395, 2017.

[10] G. Esposito, P. Venuti, S. Maestro, F. Muratori, and M. H. Bornstein, "Componential deconstruction of infant distress vocalizations via tree-based models: A study of cry in autism spectrum disorder and typical development," *Research in Developmental Disabilities*, vol. 34, pp. 2717–2724, 2013.

[11] O. F. Reyes Galaviz and C. A. Reyes Garcia, "Infant cry classification to identify hypoacoustics and asphyxia with neural networks," in *Proc. Third Mexican Int. Conf. Artif. Intell. (MICAI)*, Mexico City, Mexico, Apr. 2004, pp. 3–10.

[12] S. E. Barajas-Montiel, O. F. Reyes-Galaviz, and C. A. Reyes-Garcia, "Improving baby caring with automatic infant cry recognition," in *Proc. Int. Conf. Comput. for Handicapped Persons*, Berlin, Germany, 2006.

[13] M. Hariharan, R. Sindhu, and S. Yaacob, "Normal and hypoacoustic infant cry signal classification using time–frequency analysis and general regression neural network," *Comput. Methods Programs Biomed.*, vol. 108, pp. 559–569, 2012.

[14] J. Saraswathy, R. Sindhu, and S. Yaacob, "Infant cry classification: Time frequency analysis," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng.*, 2013.

[15] O. M. Badreldine, M. R. Hamid, N. Metwally, and E. Hamdy, "Automatic diagnosis of asphyxia infant cry signals using wavelet based mel frequency cepstrum features," in *Proc. 14th Int. Comput. Eng. Conf. (ICENCO)*, IEEE, 2018.

[16] H. B. Sailor and H. A. Patil, "Auditory Filterbank Learning Using ConvRBM for Infant Cry Classification," in *Proc. INTERSPEECH*, 2018.

[17] Y. Kristian *et al.*, "Ensemble of multimodal deep learning autoencoder for infant cry and pain detection," *F1000Research*, vol. 11, p. 359, 2023.

[18] G. Zamzmi, C.-Y. Pai, R. Goldgof, R. Kasturi, and Y. Sun, "A comprehensive and context-sensitive neonatal pain assessment using computer vision," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 28–45, 2019.

[19] G. Z. Felipe *et al.*, "Identification of infants' cry motivation using spectrograms," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, IEEE, 2019.

[20] D. Ferretti *et al.*, "Infant cry detection in adverse acoustic environments by using deep neural networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, IEEE, 2018.

[21] A. M. Mendhakar *et al.*, "Infant screening system based on cry analysis," *Int. Ann. Sci.*, vol. 6, no. 1, pp. 1–7, 2019.

[22] K. Ashwini *et al.*, "Deep convolutional neural network based feature extraction with optimized machine learning classifier in infant cry classification," in *Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA)*, IEEE, 2020, pp. 172–176.

[23] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in *Proc. IEEE 27th Conv. Electr. Electron. Eng. Israel*, 2012.

[24] R. Jahangir, "CNN-SCNet: A CNN net-based deep learning framework for infant cry detection in household setting," *Eng. Rep.*, vol. 6, no. 6, p. e12786, 2024.

[25] K. Manikanta, K. P. Soman, and M. S. Manikandan, "Deep learning based effective baby crying recognition method under indoor background sound environments," in *Proc. Int. Conf. Comput. Syst. Inf. Technol. Sustainable Solution (CSITSS)*, IEEE, 2019.

[26] J. Xie, "Baby Cry Detection Based on Audio Signals Using Deep Neural Networks," Master's thesis, Eindhoven Univ. Technol., Eindhoven, The Netherlands, 2019.

[27] C.-Y. Chang and L.-Y. Tsai, "A CNN-based method for infant cry detection and recognition," in *Proc. Workshops 33rd Int. Conf. Adv. Inf. Netw. Appl. (WAINA-2019)*, Springer, 2019.

[28] S. P. Dewi, A. L. Prasasti, and B. Irawan, "The study of baby crying analysis using MFCC and LFCC in different classification methods," *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, IEEE, 2019.

[29] C. Ji, et al., "A review of infant cry analysis and classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, article 8, Springer, 2021.

[30] N. S. A. Wahid, P. Saad, and M. Hariharan, "Automatic infant cry pattern classification for a multiclass problem," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 9, pp. 45–52, 2016.

[31] A. T. Patil, A. Kachhi, and H. A. Patil, "Subband Teager Energy Representations for Infant Cry Analysis and Classification," *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022.

[32] A. Bashiri and R. Hosseinkhani, "Infant crying classification by using genetic algorithm and artificial neural network," *Acta Medica Iranica*, vol. 58, pp. 531–539, 2020.

[33] C. Ji, et al., "Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features," *2019 International Conference on Internet of Things (iThings), GreenCom, CPSCom and SmartData*, IEEE, 2019.

[34] W. S. Limantoro, C. Fatichah, and U. L. Yuhana, "Application development for recognizing type of infant's cry sound," *2016 International Conference on Information and Communication Technology and Systems (ICTS)*, IEEE, 2016.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

[35] N. V. Dharwadkar, et al., "Identification of reasons behind infant crying using acoustic signal processing and deep neural network for neonatal intensive care unit," *International Journal of Information Retrieval Research (IJIRR)*, pp. 1–17, 2022.

[36] J. Saraswathy, et al., "Time-frequency analysis-based method for application of infant cry classification," *International Journal of Medical Engineering and Informatics*, pp. 119–134, 2020.

[37] A. Rosales-Pérez, et al., "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control*, Elsevier, pp. 38–46, 2015.

[38] A. Zabidi, et al., "Detection of asphyxia in infants using deep learning CNN trained on MFCC features extracted from cry sounds," *Journal of Fundamental and Applied Sciences*, pp. 768–778, 2017.

[39] C. C. Onu, et al., "Ubenwa: Cry-based diagnosis of birth asphyxia," *arXiv preprint arXiv:1711.06405*, 2017.

[40] M. U. Sachin, et al., "GPU based deep learning to detect asphyxia in neonates," *Indian Journal of Science and Technology*, 2017.

[41] M. Moharir, M. Sachin, and R. Nagaraj, "Identification of asphyxia in newborns using GPU for deep learning," *2017 2nd International Conference for Convergence in Technology (I2CT)*, IEEE, 2017.

[42] M. Hariharan, et al., "Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks," *Expert Systems with Applications*, Elsevier, pp. 9515–9523, 2012.

[43] R. Sahak, et al., "Performance of combined support vector machine and principal component analysis in recognizing infant cry with asphyxia," *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2010.

[44] A. Zabidi, et al., "The effect of f-ratio in the classification of asphyxiated infant cries using multilayer perceptron neural network," *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, IEEE, 2010.

[45] A. Zabidi, et al., "Classification of infant cries with asphyxia using multilayer perceptron neural network," *2010 Second International Conference on Computer Engineering and Applications*, vol. 1, IEEE, 2010.

[46] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network," *Journal of Medical Systems*, Springer, pp. 1309–1315, 2012.

[47] A. Kachhi, et al., "Data Augmentation for Infant Cry Classification," *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2022.

[48] A. L. Prasasti, L. Novamizanti, and M. I. Razik, "Identification of baby cry with discrete wavelet transform, MFCC and principal component analysis," *Journal of Physics: Conference Series*, vol. 1367, no. 1, IOP Publishing, 2019.

[49] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN–RNN," *Journal of Physics: Conference Series*, vol. 1528, no. 1, IOP Publishing, 2020.

[50] L. Novamizanti, A. L. Prasasti, and B. S. Utama, "Study of linear discriminant analysis to identify baby cry based on DWT and MFCC," *IOP Conference Series: Materials Science and Engineering*, vol. 982, no. 1, IOP Publishing, 2020.

[51] K. Srijiranon and N. Eiamkanitchat, "Application of neuro-fuzzy approaches to recognition and classification of infant cry," in *Proc. IEEE Region 10 Conf. (TENCON)*, IEEE, 2014.

[52] D. Widhyanti and D. Juniati, "Classification of baby cry sound using Higuchi's fractal dimension with K-nearest neighbor and support vector machine," *J. Phys.: Conf. Ser.*, vol. 1747, no. 1, IOP Publishing, 2021.

[53] E. Sutanto, et al., "Cry recognition for infant incubator monitoring system based on Internet of Things using machine learning," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, 2021.

[54] E. Franti, I. Ispas, and M. Dascalu, "Testing the universal baby language hypothesis—Automatic infant speech recognition with CNNs," in *Proc. 41st Int. Conf. Telecommunications and Signal Processing (TSP)*, IEEE, 2018.

[55] S. Grayson and W. Zhu, "Baby cry classifications using deep learning," 2021.

[56] P. Rani, et al., "Baby cry classification using machine learning," *Int. J. Innov. Sci. Res. Technol.*, vol. 7, 2022.

[57] A. Ekinci and E. Küçükkülahlı, "Classification of baby cries using machine learning algorithms," *East. Anatol. J. Sci.*, vol. 9, no. 1, pp. 16–26, 2023.

[58] M. Hammoud, et al., "Machine learning-based infant crying interpretation," *Front. Artif. Intell.*, vol. 7, Article 1337356, 2024.

[59] S. A. Younis, D. Sobhy, and N. S. Tawfik, "Evaluating convolutional neural networks and vision transformers for baby cry sound analysis," *Future Internet*, vol. 16, no. 7, 2024.

[60] P. Kulkarni, et al., "Child cry classification—An analysis of features and models," in *Proc. 6th Int. Conf. Convergence in Technology (I2CT)*, IEEE, 2021.

[61] K. Sharma, C. Gupta, and S. Gupta, "Infant weeping calls decoder using statistical feature extraction and Gaussian mixture models," in *Proc. 10th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, IEEE, 2019.

[62] K. Rezaee, et al., "Can you understand why I am crying? A decision-making system for classifying infants' cry languages based on DeepSVM model," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–17, 2024.

[63] R. I. Tuduce, et al., "Automated baby cry classification on a hospital-acquired baby cry database," in *Proc. 42nd Int. Conf. Telecommunications and Signal Processing (TSP)*, IEEE, 2019.

[64] M. S. Rusu, et al., "Database and system design for data collection of crying related to infant's needs and diseases," in *Proc. Int. Conf. Speech Technol. Human-Computer Dialogue (SpeD)*, IEEE, 2015.

[65] I.-A. Bănică, et al., "Baby cry recognition in real-world conditions," in *Proc. 39th Int. Conf. Telecommunications and Signal Processing (TSP)*, IEEE, 2016.

[66] V. R. Joshi, et al., "A multistage heterogeneous stacking ensemble model for augmented infant cry classification," *Front. Public Health*, vol. 10, Article 819865, 2022.

[67] M. A. T. Turan and E. Erzin, "Monitoring infant's emotional cry in domestic environments using the capsule network architecture," in *Proc. Interspeech*, 2018.

[68] M. Huckvale, "Neural network architecture that combines temporal and summative features for infant cry classification in the Interspeech 2018 computational paralinguistics challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, ISCA, 2018.

[69] B. W. Schuller, et al., "The Interspeech 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, ISCA, 2018.

[70] J. J. Parga, et al., "Defining and distinguishing infant behavioral states using acoustic cry analysis: Is colic painful?," *Pediatr. Res.*, vol. 87, no. 3, pp. 576–580, 2020.

[71] A. Khozaei, et al., "Early screening of autism spectrum disorder using cry features," *PLoS One*, vol. 15, no. 12, p. e0241690, 2020.

[72] K. Wu, et al., "Research on acoustic feature extraction of crying for early screening of children with autism," in *Proc. 34th Youth Acad. Annu. Conf. Chinese Assoc. Autom. (YAC)*, IEEE, 2019.

[73] Y. Okada, K. Fukuta, and T. Nagashima, "Iterative forward selection method based on cross-validation approach and its application to infant cry classification," in *Proc. Int. MultiConf. Eng. Comput. Sci. (IMECS)*, Hong Kong, Mar. 2011, pp. 49–52.

[74] X. Wang, et al., "Statistical method for classifying cries of baby based on pattern recognition of power spectrum," *Int. J. Biometrics*, vol. 2, no. 2, pp. 113–123, 2010.

[75] A. Zabidi, et al., *Detection of infant hypothyroidism with mel frequency cepstrum analysis and multi-layer perceptron classification*, 2010 6th International Colloquium on Signal Processing and its Applications, IEEE.

[76] A. Zabidi, et al., *Classification of infant cries with hypothyroidism using multilayer perceptron neural network*, 2009 IEEE International Conference on Signal and Image Processing Applications, IEEE, 2009.

[77] S. Orlandi, et al., *Application of pattern recognition techniques to the classification of full-term and preterm infant cry*, Journal of Voice, vol. 30, no. 6, pp. 656–663, Elsevier, 2016.

[78] A. Laguna, et al., *How can cry acoustics associate newborns' distress levels with neurophysiological and behavioral signals?*, Frontiers in Neuroscience, vol. 17, article 1266873, 2023.

[79] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[80] J. L. Semmlow, *Biosignal and Biomedical Image Processing*, Marcel Dekker Inc., 2004.

[81] A. Chaiwachiragompol and N. Suwannata, *The features extraction of infants cries by using discrete wavelet transform techniques*, Procedia Computer Science, vol. 86, pp. 285–288, Elsevier, 2016.

[82] M. A. T. Jimenez, *Summarization of video from Feature Extraction Method using Image Processing & Artificial Intelligence*, 2018.

[83] D. Putra and A. Resmawan, *Verifikasi Biometrika Suara Menggunakan Metode MFCC dan DTW*, Lontar Komputer, vol. 2, pp. 8–21, 2011.

[84] K. S. Alishamol, et al., *System for infant cry emotion recognition using DNN*, 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2020.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

[85] S. Carvalho and E. F. Gomes, *Automatic classification of bird sounds: using MFCC and mel spectrogram features with deep learning*, Vietnam Journal of Computer Science, vol. 10, no. 01, pp. 39–54, 2023.

[86] Y.-C. Liang, et al., *Deep learning for infant cry recognition*, International Journal of Environmental Research and Public Health, vol. 19, no. 10, p. 6311, MDPI, 2022.

[87] L. Liu, Y. Li, and K. Kuo, *Infant cry signal detection, pattern extraction and recognition*, 2018 International Conference on Information and Computer Technologies (ICICT), IEEE, 2018.

[88] A. M. Mahmoud, et al., *Infant cry classification using semi-supervised k-nearest neighbor approach*, 2020 13th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2020.

[89] S. Bano and K. M. RaviKumar, *Decoding baby talk: A novel approach for normal infant cry signal classification*, 2015 International Conference on Soft-Computing and Networks Security (ICSNS), IEEE, 2015.

[90] S. Ntalampiras, *Audio pattern recognition of baby crying sound events*, Journal of the Audio Engineering Society, vol. 63, no. 5, pp. 358–369, 2015.

[91] G. Gu, X. Shen, and P. Xu, "A set of DSP system to detect baby crying," in *Proc. 2nd IEEE Adv. Inf. Manage., Commun., Electron. Autom. Control Conf. (IMCEC)*, 2018.

[92] Y. Zayed, A. Hasasneh, and C. Tadj, "Infant cry signal diagnostic system using deep learning and fused features," *Diagnostics*, vol. 13, no. 12, p. 2107, 2023.

[93] Y. Lavner, R. Cohen, D. Ruinskiy, and H. IJzerman, "Baby cry detection in domestic environment using deep learning," in *Proc. IEEE Int. Conf. Sci. Elect. Eng. (ICSEE)*, 2016, pp. 1–5.

[94] M. Petroni, A. S. Malowany, C. C. Johnston, and B. L. Stevens, "Classification of infant cry vocalizations using artificial neural networks (ANNs)," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, pp. 3475–3478, 1995.

[95] A. Chittora and H. A. Patil, "Newborn infant's cry analysis," *Int. J. Speech Technol.*, vol. 19, pp. 919–928, 2016.

[96] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in *Proc. 18th Int. Conf. Eng. Appl. Neural Netw. (EANN)*, Athens, Greece, 2017, pp. 1–12.

[97] T. Zan, H. Wang, M. Wang, Z. Liu, and X. Gao, "Application of multi-dimension input convolutional neural network in fault diagnosis of rolling bearings," *Appl. Sci.*, vol. 9, no. 13, p. 2690, 2019.

[98] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[99] W. J. Han and H. F. Li, "A brief review on emotional speech databases," *Intell. Comput. Appl.*, vol. 3, no. 1, pp. 5–7, 2013.

[100] B. Athiwaratkun and J. W. Stokes, "Malware classification with LSTM and GRU language models and a character-level CNN," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 2482–2486. DOI: 10.1109/ICASSP.2017.7952604.

[101] E. Xi, S. Bing, and Y. Jin, "Capsule network performance on complex data," *arXiv:1712.03480*, 2017.

[102] K. Saetern and N. Eiamkanitchat, "An ensemble K-nearest neighbor with neuro-fuzzy method for classification," in *Proc. 10th Int. Conf. Comput. Inf. Technol. (IC2IT)*, 2014, pp. [PAGE NUMBERS].

[103] S. Kumar, *Neural Networks: A Classroom Approach*. New York, NY, USA: Tata McGraw-Hill, 2004.

[104] M. H. Sazlı, "A brief review of feed-forward neural networks," *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, vol. 50, no. 01, 2006.

[105] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, 2015, pp. 18–24.

[106] W. Khan, M. Tahir, S. Kadry, Y. Nam, and S. Mumtaz, "Deep face profiler (DeFaP): Towards explicit, non-restrained, non-invasive, facial and gaze comprehension," *Expert Systems With Applications*, vol. 254, p. 124425, 2024.

[107] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1994.

[108] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285–2294.

[109] J. Gu, V. Tresp, and H. Hu, "Capsule network is not more robust than convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14309–14317.

[110] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[111] S. Jindal, K. Nathwani, and V. Abrol, "Classification of infant behavioural traits using acoustic cry: an empirical study," in *Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA)*, 2021, pp. 1–6.

[112] P. Kumari and K. Mahto, "A narrative review on different novel machine learning techniques for detecting pathologies in infants from born baby cries," *Journal of Voice*, 2024.

[113] J. Soltis, "The developmental mechanisms and the signal functions of early infant crying," *Behavioral and Brain Sciences*, vol. 27, no. 4, pp. 477–490, 2004.

[114] Stepping Stones Pediatric Therapy, Inc, "Enhancing Child Development: The Role of Caregiver Support," 2024, Available: https://steppingstonesptnh.com/blog/enhancing-child-development-the-role-of-caregiver-support, Accessed: 2024-12-13.

[115] V. Giordano, M. C. Gorgoglione, A. Di Mauro, G. Trifone, and D. De Venuto, "Comparative analysis of artificial intelligence and expert assessments in detecting neonatal procedural pain," *Scientific Reports*, vol. 14, no. 1, p. 20374, 2024.

[116] C. Laing and E. Bergelson, "From babble to words: Infants' early productions match words and objects in their environment," *Cognitive Psychology*, vol. 122, p. 101308, 2020.

[117] A. F. Symon, M. A. Hannan, A. Hussain, and H. Basri, "Design and development of a smart baby monitoring system based on Raspberry Pi and Pi camera," in *Proc. 4th Int. Conf. Adv. Elect. Eng. (ICAEE)*, 2017, pp. 540–545.

[118] H. Alam, S. Ahmed, M. Rahman, and K. Khan, "IoT based smart baby monitoring system with emotion recognition using machine learning," *Wireless Commun. Mobile Comput.*, vol. 2023, p. 1175450, 2023.

[119] W. A. Jabbar, K. I. K. Wang, M. A. H. A. Rahman, H. S. S. A. Rahim, and W. L. W. Arif, "IoT-BBMS: Internet of Things-based baby monitoring system for smart cradle," *IEEE Access*, vol. 7, pp. 93791–93805, 2019.

[120] P. R. Myakala, V. K. Kothuru, S. P. Spandana, and S. Devarakonda, "An intelligent system for infant cry detection and information in real time," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 1–6.

[121] C. Ji, "Infant cry signal processing, analysis, and classification with artificial neural networks," Master's thesis, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2021.

[122] S. B. Junaid, R. O. Gbadamosi, S. A. Imam, A. A. Surakat, A. A. Balogun, and G. A. Sahalu, "Recent advancements in emerging technologies for healthcare management systems: a survey," *Healthcare*, vol. 10, no. 10, p. 2006, 2022.

[123] Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success, Board on Children, Youth, and Families, Institute of Medicine, National Research Council, *Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation*, L. R. Allen and B. B. Kelly, Eds. Washington, DC, USA: National Academies Press, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK310550/ Accessed: Dec. 16, 2024.

[124] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations," *Speech Communication*, vol. 133, pp. 41–61, 2021.

[125] T. Polzehl, S. K. Nallanthighal, A. Stöter, and H. R. Pfitzinger, "Towards classifying mother tongue from infant cries: Findings substantiating prenatal learning theory," in *Proc. Interspeech*, 2024.

[126] G. Veres, "Donateacry Corpus," 2019. [Online]. Available: https://github.com/gveres/donateacry-corpus

[127] D. Ashok, A. P. Kumar, K. B. C., and M. N. Kumar, "Emotion detection in baby cries using machine learning with lullaby recommendations," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 12, pp. 2126–2129, 2024.

[128] X. Yao, Y. Wang, Z. Li, L. Chen, and H. Li, "Infant crying detection in real-world environments," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 4063–4067.

[129] S. Sharma, P. Viswanath, and V. K. Mittal, "Infant crying cause recognition using conventional and deep learning based approaches," in *Proceedings of the 15th International Conference on Natural Language Processing*, 2018.

[130] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[131] J. Saraswathy, M. Hariharan, S. Yaacob, and W. Khairunizam, "Optimal selection of mother wavelet for accurate infant cry classification,"

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3636598

Australasian Physical & Engineering Sciences in Medicine, vol. 37, pp. 439–456, 2014.

[132] M. D. Renanti, A. Buono, K. Priandana, and S. H. Wijaya, "Evaluating noise-robustness of convolutional and recurrent neural networks for baby cry recognition," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 6, 2024.

[133] J. Li, "The benefits and drawbacks of baby monitors," *Netvue Blog*, Mar. 19, 2024. [Online]. Available: https://www.netvue.com/blogs/netvue-blog/the-benefits-and-drawbacks-of-baby-monitors

[134] T. Ozcan and H. Gungor, "Baby cry classification using structure-tuned artificial neural networks with data augmentation and MFCC features," *Applied Sciences*, vol. 15, no. 5, p. 2648, 2025.

[135] M. D. Renanti, A. Buono, K. Priandana, and S. H. Wijaya, "Noise-robust in the baby cry translator using recurrent neural network modeling," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 2, 2023.

[136] swisstech, "Decoding baby cries with the help of AI," *swisstech*, Jun. 6, 2024. [Online]. Available: https://www.swiss.tech/news/decoding-baby-cries-help-ai

[137] J. Huang, Y. Chen, L. Wang, and K. Zhang, "Design and implementation of infant crying monitoring and analysis system based on deep learning," in *Proc. 8th Int. Conf. Electron. Inf. Technol. Comput. Eng.*, 2024.

[138] P. R. Myakala, V. K. Kothuru, S. P. Spandana, and S. Devarakonda, "A low cost intelligent smart system for real time infant monitoring and cry detection," in *Proc. IEEE Region 10 Conf. (TENCON)*, 2017.

[139] M. Hong, Y. Li, Z. Wang, and X. Chen, "InfantCryNet: A data-driven framework for intelligent analysis of infant cries," *arXiv:2409.19689*, 2024.

[140] A. Kuzmin, M. Van Baalen, Y. Ren, and D. Zhou, "Pruning vs quantization: Which is better?," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 62414–62427, 2023.

[141] A. Ancilotto, F. Paissan, and E. Farella, "Xinet: Efficient neural networks for TinyML," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023.

[142] M. Lockhart-Bouron, A. Smith, J. Doe, and R. Johnson, "Infant cries convey both stable and dynamic information about age and identity," *Commun. Psychol.*, vol. 1, no. 1, p. 26, 2023.

[143] G. Gabrieli, M. Esposito, and P. Venuti, "Are cry studies replicable? An analysis of participants, procedures, and methods adopted and reported in studies of infant cries," *Acoustics*, vol. 1, no. 4, 2019.

[144] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[145] F. Zhuang, Z. Qi, K. Duan, and D. Xi, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[146] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv:1708.08296*, 2017.

[147] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.

[148] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[149] L. Masi, D. Petrov, and M. Fadel, "Multimode trapped interferometer with noninteracting Bose-Einstein condensates," *Phys. Rev. Res.*, vol. 3, no. 4, p. 043188, 2021.

[150] S. Orlandi, M. Esposito, and P. Venuti, "Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2012.

[151] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[152] J. Ngiam, A. Khosla, M. Kim, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011.

[153] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.

[154] S. Pusil, M. López, and J. García, "Early identification of autism using cry analysis: A systematic review and meta-analysis of retrospective and prospective studies," *J. Autism Dev. Disord.*, 2025.

[155] A. W. Manigault, S. L. Wood, and R. J. Ward, "Acoustic cry characteristics in preterm infants and developmental and behavioral outcomes at 2 years of age," *JAMA Netw. Open*, vol. 6, no. 2, pp. e2254151–e2254151, 2023.

[156] T. Ozcan and H. Gungor, "Baby Cry Classification Using Structure-Tuned Artificial Neural Networks with Data Augmentation and MFCC Features," *Applied Sciences*, vol. 15, no. 5, p. 2648, 2025.

[157] X. Qiao, Y. Li, J. Zhang, and H. Wang, "Infant Cry Classification Using an Efficient Graph Structure and Attention-Based Model," *Kuwait Journal of Science*, vol. 51, no. 3, p. 100221, 2024.

[158] The Daily Star, "Tech parenting: The fine line between convenience and neglect," *The Daily Star*, 2024. [Online]. Available: https://www.thedailystar.net/anniversary-supplement-2024/lifestyle-diaries/news/tech-parenting-walking-fine-line-between-convenience-neglect-35504712

Seyyed Mohammad Hossein Hashemi Seyyed Mohammad Hossein Hashemi received a B.Sc. degree in Telecommunication Engineering from Persian Gulf University in 2024. His current research interests include computer vision, deep learning, and their applications in healthcare and audio signal processing.

Hoshang Kolivand Hooshang Kolivand received his MS degree in applied mathematics and computer from Amir-Kabir University, Iran, in 1999, and his Ph.D. from the Media and Games Innovation Centre of Excellence (MaGIC-X) at Universiti Teknologi Malaysia. Previously, he worked as a lecturer at Shahid Beheshti University, Iran. He has published numerous articles in international journals, conference proceedings, and technical papers, including chapters in books. An active reviewer for many conferences and international journals, Hoshang has also authored several books on object-oriented programming and mathematics. His current research interests focus on computer graphics and Augmented Reality.

Wasiq Khan (Senior Member, IEEE) received a Ph.D. in speech analysis and intelligent reasoning from the University of Bradford, U.K. Currently, he is a Senior Academic in AI with the Department of Computer Science, Liverpool John Moores University, U.K. He is also a Visiting Professor of AI with the University of Anbar, Iraq. He has been publishing the research outcomes in high-impact journals and is editorial board member for prestigious Journals and international conferences.

Tanzila Saba (Senior Member, IEEE) received the Ph.D. degree in document information management and security from the Faculty of Computing, Universiti Teknologi Malaysia, Malaysia, in 2012. Currently, she is an Eminent Researcher with the Image Processing Research Group and an Assistant Professor with the College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia. She has authored over 30 papers in high-impact journals and was honored with the Marquis Who's Who 2012 Award for her research excellence.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on December 03,2025 at 10:03:36 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.