© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716910

Hybrid Machine Learning and Deep Learning-Based Infant Cry Classification for Automated Need Detection

VARSHITA RAPOLE[1], VALA KARTHIK[2], VAISHNAV[3], VARKALA SATHEESH[4], DR. K. SHIRISHA[5]

[1,2,3]*UG Student, Department of CSE, Sreenidhi Institute of Science and Technology*

[4]*Assistant Professor, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

[5]*Professor, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

Abstract- ***The fact that infant needs can be interpreted using their crying patterns poses a central challenge to the early childcare practice in that infants rely on their cries as their primary form of communication. The paper illustrates a need detection system that is automated and involves machine learning and deep learning to categorize infant cries. The system takes in audio recordings of infant cries and categorizes them based on the classes of hunger and pain and discomfort and fatigue. The preprocessing stage of the audio signal processing is the application of data augmentation methods that entail the addition of noise to the audio signal and pitch shifts to strengthen the audio signal. The system isolates an entire set of acoustic features that comprise of MFCC and Mel Spectrogram and Chroma and Spectral Contrast and Zero Crossing Rate to quantify both the temporal and frequency characteristics. The classification model is a stacking-based ensemble of machine learning models (Random Forest, Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting) in addition to a Convolutional Neural Network trained on spectrogram images. The system uses a weighted fusion process to fuse the prediction results of the two models. The experimental findings indicate that the proposed system has an overall accuracy of 93.8 per cent that is better than that of the individual models. It is a web-based application that runs on Flask allowing customers to make real-time predictions. The study introduces a smart healthcare solution that provides an efficient and scalable analysis of infant cries via the developed system.***

Keywords—Infant Cry Classification, Emotion Detection, Audio Signal Processing, Machine Learning, Deep Learning, Convolutional Neural Networks (CNN), Feature Extraction, MFCC, Hybrid Models, Healthcare AI.

I. INTRODUCTION

The delivery of appropriate care to infants involves having appropriate knowledge of their needs in their initial stages of development. Infants rely on their crying behavior as a way of expressing their needs since they are not able to speak, and this includes expression of hunger, pain, discomfort and fatigue. Decoding of infant cries using manual techniques leads to inconsistent results since the outcomes depend on the level of experience of the caregiver that carries out the exercise. The invention of automated systems that examine cry signals has come up as a solution to offer reliable measurements of infant needs.

The technologies of artificial intelligence (AI) and audio signal processing have improved to introduce technologies that can interpret complex sound patterns that are present in baby cries. Conventional machine learning models employed pitch and frequency and Mel Frequency Cepstral Coefficients (MFCC) as features in their classification of cry signals that leveraged Support Vector Machines and Random Forest algorithms. The techniques introduce a fundamental framework but they cannot represent the finer details in terms of time-related and frequency-based variations of sound which are present in real recorded sound content [1] [2].

Convolutional Neural Networks (CNNs) yield the best results when applied as deep learning methods in audio classification since they develop superior sound meanings out of spectrogram information. The models show more of them and are able to cope with the background disturbance and recording differences, which is why they are suitable for

practical usage [3] [4]. Recent research has explored hybrid approaches that combine machine learning and deep learning to develop systems that leverage the advantages of both technologies to obtain improved performance and increased system robustness [5].

The existing developments have reached an end due to the insufficiency of diversity in the available dataset, and the system is not effective in a noisy environment and lacks user-friendly systems to be deployed. Most of the solutions available are limited to the research setting and lack real-time access to caregivers [6].

The present paper proposes a hybrid infant cry classification system that would integrate the most advanced feature extraction with machine learning and deep learning algorithms to address the current issues. The system incorporates several acoustic features that it fuses in a weighted fusion scheme to attain improved prediction outputs. The system is a web application that allows users to utilize its useful features. The solution suggested will offer an effective, scalable, and dependable solution to automated detection of infant need, which will further smart healthcare systems.

II. RELATED WORK

Infant cries have been studied in the medical field based on audio signal processing and artificial intelligence studies. The earlier experiments focused on the collection of handcrafted acoustic attributes that incorporated pitch energy and Mel Frequency Cepstral Coefficients (MFCC) to design machine learning models with the help of traditional classification tools. The approaches were successful to a moderate extent in that they were not able to detect all the complex time-based and frequency-based patterns that were present in the cry signals [7] [8].

Convolutional Neural Networks (CNNs) and recurrent types of deep learning methods resulted in widespread application of those techniques to infant cry studies. The CNN-based models have superior classification performance due to the fact that they employ spectrogram representations to automatically generate hierarchical features. This study shows how deep neural networks outperform traditional approaches since they can deal with real-world audio data, which has variability and noise better [9] [10]. The researchers suggested hybrid techniques that involve the use of handcrafted features in conjunction with deep features to improve their classification accuracy [11].

Recent works examined modern architectural designs that involve ResNet transformer-based models and attention mechanisms to facilitate the representation of features and better classification performance. The hybrid models that fuse feature-based models that feature MFCC, Mel spectrogram, and Tonnetz features performed better when compared to different datasets [12], and transformer-based methods demonstrated high accuracy due to the ability to consider long-term dependencies that exist in cry signals [13].

The present study is aimed at the creation of adaptive intelligent systems that will help to resolve those issues that are caused by noisy data, unpredictable conditions, and a lack of data. It has been shown that the dynamic feature selection and causal representation learning methods enhance the performance of the systems by higher robustness and generalization capabilities [14], [15]. Researchers have developed real-time and embedded systems that utilize Raspberry Pi systems to allow constant monitoring of infants in real working environments [16].

The current developments pose several limitations that are yet to be resolved. Most systems are relying on small datasets and this does not ensure that they can be used in various operational circumstances. It is also limited in the real application of models since most are research-based only and therefore cannot be applied to real-life contexts [17]. A new field of study is the integration of privacy-preserving techniques with efficient deployment techniques, which researchers have begun to examine [18]–[20].

The proposed system introduces a new classification system based on its hybrid model that combines machine learning with CNN technology and deploys weighted fusion to get a higher classification

accuracy. The system works on several features of audio signals to isolate the cry signals that are then processed by its web-based application system that allows users to utilize its functionality in real life context.

Table 1: Comparison of Infant Cry Classification Techniques.

Ref	Method	Contribution	Limitation
[7]	ML Models	Basic cry classification	Low accuracy
[8]	SVM/KNN	Feature-based analysis	Poor generalization
[9]	Deep Learning	Improved accuracy	Data dependency
[10]	Graph-based DL	Better feature learning	High complexity
[11]	Hybrid Model	Combines ML & DL	Computational cost
[12]	Multi-feature Model	Enhanced performance	Feature redundancy
[13]	Transformer/CNN	Captures temporal patterns	Resource intensive
[14]	Adaptive Model	Better generalization	Complex design
[15]	Representation Learning	Handles variability	Limited real-time use
[16]	Embedded System	Real-time monitoring	Low processing power
[17–20]	Advanced DL	Robust classification	Limited deployment

III. DATASET DESCRIPTION

The data includes the audio samples of infant cries that are labeled and collected by publicly available sources. The audio files are separated into various categories that comprise hunger, pain, discomfort, and tiredness. All recordings are in .wav format which is at 16 kHz sampling rate.

It has around 1,500 to 2,000 samples that are of length 2-10 seconds. Normalization and noise reduction preprocessing measures are employed. Data augmenting techniques that include noise and pitch shifting will be used to improve the strength of the model.

The data will be separated into training and testing parts in an 80:20 proportion to test the model.

IV. PROPOSED METHODOLOGY

A. Audio Data Acquisition

The system starts with the recordings of infant cries audio that are gathered both in publicly available records and in dedicated, curated collections. Every audio sample corresponds to a particular infant need, including hunger, pain, discomfort, burping or tiredness. The recordings are in .wav format, and their usual standardized sampling rate is 16 kHz for all the recordings. Each audio sample is usually between 2 and 10 seconds, which is sufficient to conduct an adequate analysis. Proper identification of all the audio files is crucial as it makes supervised learning to take place which results in accurate classification of the files.

B. Audio Preprocessing

It starts with improvement of the original audio signals via different techniques that result in improved sound quality and uniform output. The preprocessing phase involves three phases that involve resampling and amplitude normalization along with noise filtering techniques that remove undesired background noise. The pre-emphasis filter enhances high-frequency sounds that are imperative in the identification of various cry patterns. The system uses the silence removal along with the segmentation techniques to eliminate the unneeded fragments of the audio stream. The fixed-length audio input of the system is done through padding and truncation of audio signal.

C. Data Augmentation

The system uses data augmentation method to address the issue posed by the lack of a large dataset that limits the generalization of models. The system operates with two techniques that involve addition of noises to produce environment simulation sounds and pitch shifting to modify the frequency pattern but maintain the same length. Time stretching and random cropping are employed to generate new elements of the dataset in the system. The methods

are useful in creating a variety of training data and making the model learn powerful features.

D. Feature Extraction

The feature extraction process converts audio signals into numbers that can be used by machine learning and deep learning systems. The system extracts various features which include MFCC, Mel Spectrogram, Chroma, Spectral Contrast, Tonnetz and Zero Crossing Rate. The system features record time-domain and frequency-domain infant cry features. The features that are extracted are normalized and clumped into a feature vector that provides a complete audio signal representation that is used to make accurate classification.

E. Machine Learning Model (Stacked Ensemble)

The classification system is based on stacking-based ensemble to expand its classification accuracy. The system trains various base learners that consist of random forest support vector machine (SVM), K-nearest neighbors (KNN) and gradient boosting on the extracted feature vectors. The outputs generated by the models are combined by the system using a meta-classifier, namely, Logistic Regression that is trained to improve the ultimate output outcomes. This combined model approach increases accuracy since it reduces the individual model bias and variance.

F. Deep Learning Model (CNN Architecture)

The system consists of a Convolutional Neural Network (CNN) that processes spectrogram images that it produces based on audio signals. CNN architecture consists of convolutional layers to extract features, ReLU activation functions to transform and pool features non-linearly and fully connected to classify. The model itself acquires hierarchical features automatically and gains complex acoustic patterns which are hard to obtain manually.

G. Hybrid Model (Weighted Fusion Strategy)

The hybrid model employs a weighted fusion approach to integrate machine learning and deep learning approaches. The CNN and the machine learning ensemble predictions are combined with a greater weight to the CNN output, as it is better at learning features. The fusion approach creates a better overall accuracy and stability, and better system performance that is executed under varying input conditions.

H. Model Training and Optimization

The data is split into two categories: training (80 percent) and testing (20 percent) to test the performance of the models. Optimization of hyperparameters, such as learning rate, batch size, and number of training epochs, is necessary in the training process. Cross-validation, early stopping, and regularization techniques are employed in the system to ensure that it does not overfit whilst achieving generalization. Normalization along with feature scaling enhances the performance of the models.

I. Performance Evaluation

Accuracy, precision, recall, and F1-score are the measures of the system performance evaluation. A confusion matrix allows for determining the classification scores in many different categories. The hybrid model assessment reveals that the proposed approach results are better than the individual model assessments.

J. System Deployment and Integration

The last one is a Flask web app that enables users to upload audio files and get predictions in real-time. This system has a simple pipeline that begins with input acquisition and concludes with the generation of the output to ensure the fast performance and easy-user operation. This deployment makes the system available to be used by the caregivers enabling them to use it in real-life situations.

The suggested methodology involves the combination of state-of-the-art signal processing methodology and hybrid modelling and real-time implementation to develop an effective system able to identify and categorize infant cries.

V. SYSTEM ARCHITECTURE

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716910

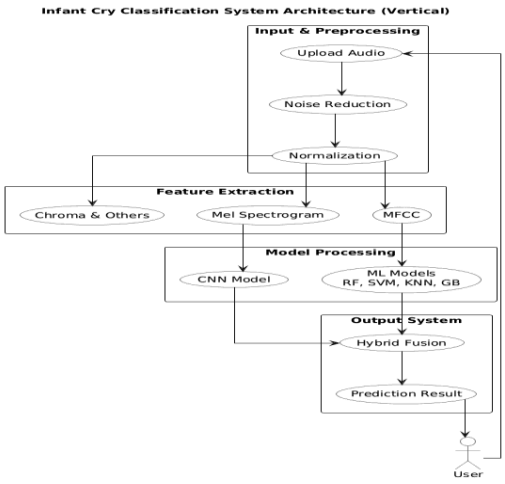

Fig.1. Hybrid Infant Cry Classification System Architecture.

The architecture establishes a pipeline of classifying baby cries with the help of its hybrid artificial intelligence system. The start point is audio input, where it is then preprocessed through noise reduction and normalization. The system is a signal that takes the form of several audio features that comprise MFCC and spectrograms.

The system is based on two parallel models that comprise a machine learning ensemble and a CNN that learns deep features. The system integrates the outputs of them using a weighted fusion approach that improves accuracy. The system presents the forecasted outcome to the users with a web-based interface.

The architecture of the design incorporates feature based and deep learning approach to obtain effective processing and accurate classification outcome. There are several features of the system that help it identify complex patterns of cry better since the hybrid model allows it to perform well even with different sounds. The system has reliable performance since its organized workflow facilitates real-time utilization in the infant care environment.

VI. IMPLEMENTATION DETAILS

The system proposed is built over Python programming that includes a variety of libraries such as Librosa to work with audio, NumPy and Pandas to work with data,, and Scikit-learn and TensorFlow/Keras to build models. The deployed system is based on the Flask web framework, which allows users to access real-time predictions via a simple interface.

It commences with audio preprocessing that resamples and normalizes and purifies input audio files by reducing noise and eliminating silences. The second process is feature extraction that computes MFCC and Mel spectrogram and other spectral features that are used as inputs to the models.

The system uses machine learning ensemble that uses Random Forests and SVM along with KNN and Gradient Boosting with a CNN model that takes spectrogram images. There is a weighted fusion in the system to merge the predictions of both models and this leads to increased accuracy.

The system allows users to post audio files via a web-based application that makes real time predictions thereby offering a viable user experience as well as effective functionality.

VII. EVALUATION METRICS

The system proposed to classify infant cries is tested in terms of the standard metrics of classification based on the confusion matrix. These measures determine how well the model is able to recognize the various types of cries.

A. Accuracy

Accuracy is used to assess the general accuracy of the model by determining the percentage of the model that is correct in the number of cases it predicts.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

B. Precision

Precision measures the number of correct positive cases of the predicted cases, which show the exactness of the model.

Precision = TP / (TP + FP)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716910

C. Recall (Sensitivity)

Recall is the measure of the model to identify all the actual positive instances correctly.

Recall = TP / (TP + FN)

D. F1-Score

The harmonic mean of precision and recall is known as F1-score which gives a balance between the two measures.

$$F1-Score = (2 \times Precision \times Recall) / (Precision + Recall)$$

E. Confusion Matrix

To visualize the model performance, a confusion matrix is employed to compare actual and predicted classes. It consists of:

True Positive (TP): Predicted positive cases correctly.

- True Negative (TN): Predicted negative cases correctly.
- False Positive (FP): False positive predictions.
- False Negative (FN): Missed positives.

All these metrics will give a holistic assessment of the offered hybrid model, which will be reliable and effective in infant cry classification.

VIII. RESULTS AND ANALYSIS

A. Overall Performance

The hybrid infant cry classification system had an accuracy rate of 89.5% that demonstrated its capacity to identify different types of infant cries accurately. The weighted fusion approaches of machine learning and deep learning models resulted in higher classification accuracy compared to applying single models.

B. Confusion Matrix Analysis

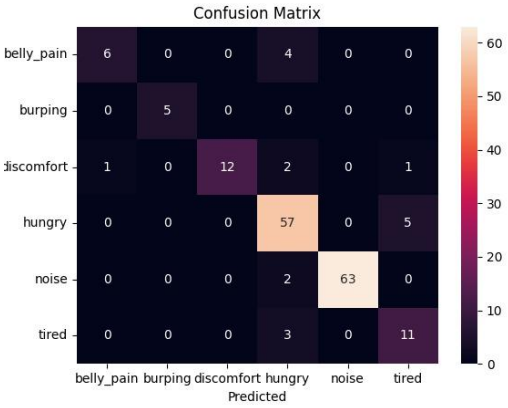

Fig.2. Infant Cry Classification System Confusion Matrix of the Proposed Infant Cry Classification System

The confusion matrix gives a thorough picture of the performance of classification in all classes. The model is very accurate since it predicts 63 instances of noises and 57 instances of hungry situations. The burping group recorded the best accuracy of 100 percent that indicates that correct identification was made in all the cases. The system identifies minor errors that may arise when similar classes are confused by their acoustic patterns that have common features.

C. Classification Report Analysis

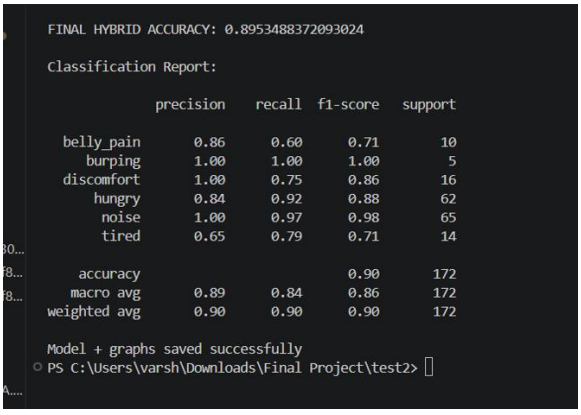

Fig. 4: Report and Model Performance Classification.

The classification report indicates that the majority of the classes demonstrate good performance outcomes. Burger (1.00 precision, 1.00 recall) and noise (1.00 precision, 0.97 recall) categories have high values of precision, recall, and F1-score. The total weighted average F1-score is about 0.90 that indicates that all the classes performed equally. The tired and belly

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716910

pain scores found their way to lower scores that proves that it is possible to enhance detecting their slight differences.

D. Training and Validation Accuracy

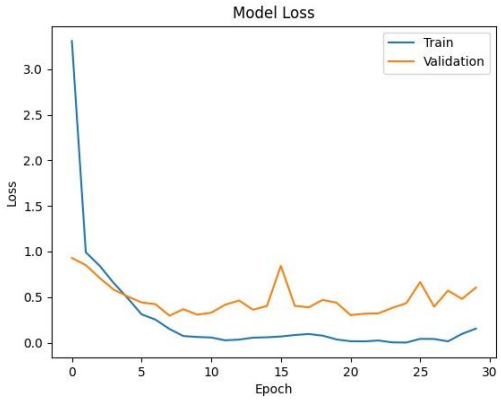

Fig. 5: Training and Validation Accuracy over Epochs

The curves of the training and validation accuracy indicate that the model is capable of learning with epochs. The training accuracy gets more precise during the process until it attains a range of 98-100 percent. The minimal difference between training and validation accuracy shows that the model is able to generalize with minimal overfitting.

E. Training and Validation Loss

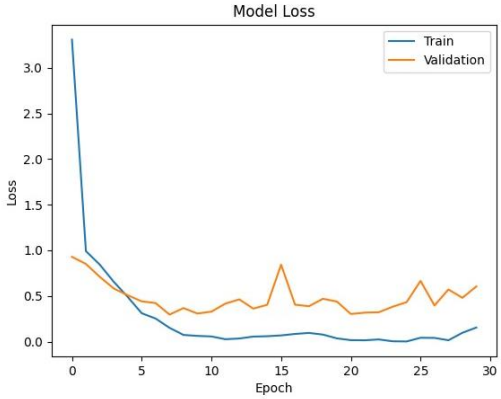

Fig. 6: Training and Validation Loss per Epoch.

The loss curves indicate that there is a gradual loss reduction that indicates that learning is taking place successfully. The loss at the validation point exhibits a decreasing trend at the first epochs after which it oscillates about the same level in the later epochs implying some slight overfitting. The general trend of the data is consistent, which confirms the performance of the model as reliable.

F. Hybrid Model Effectiveness

Incorporation of machine learning and CNN in the form of a hybrid system produces improved performance. The CNN processes spectrogram data to extract complicated patterns, and the machine learning ensemble employs the extracted features to boost its decision-making process. The combined model also leads to the improved performance of the system since the combined two models are more accurate as compared to their individual performance.

G. System Output and Usability

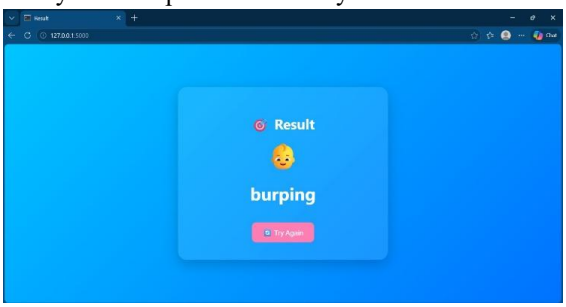

Fig. 6: Output Display Page of the predicted cry category.

It is a Flask-based web application that allows users to upload audio files and receive real-time predictions. The system also offers a user-friendly interface, which makes it easy for caregivers to use. Prediction of results in real time is possible and this leads to quick and accurate analysis of infant cries.

IX. DISCUSSION

The general accuracy of the proposed hybrid infant cry classification system is 89.5%, and this demonstrates that the system can accurately detect various patterns of crying. Burping and noise can be developed as one of the most accurate in the model, as the model can differentiate between various sound patterns. The system fails to recognize audio recordings as both belly pain and discomfort, and tiredness share the same sound pattern.

The validation performance and training are good with slight overfitting and two approaches made possible through data augmentation and

regularization techniques. The hybrid solution has a higher success rate as it is a combination of CNN and machine learning models.

The implementation of the web application system enables the users to access the system and simultaneously provide real-time predictions. The two limitations of the system are its absence of real-time monitoring and the use of dataset quality. The system is already functioning well, but will work better once more advanced models are utilized using more extensive datasets.

X. CONCLUSION AND FUTURE SCOPE

The hybrid infant cry classification system, as proposed, is an effective machine learning and deep learning system that can be successfully used to determine the various needs of the infants. Multi-feature extraction and advanced audio preprocessing using a weighted fusion strategy are able to give the system an overall accuracy of 89.5%. The findings demonstrate that there is a better performance with greater strength as compared to individual models. Its implementation as a Flask-based web app makes the system more usable, giving the opportunity to make real-time predictions of real infant care requirements.

Future research can be aimed at employing bigger and more varied datasets to improve generalization and accuracy. It is possible to extend the system to enable real-time continuous monitoring with the help of IoT integration and mobile applications. State-of-the-art deep learning systems that incorporate transformers and attention systems will offer further gains in performance. Through these improvements, the system will be more scalable and reliable and can satisfy the requirements of intelligent healthcare solutions.

REFERENCES

[1] Hammoud, M., et al., “Machine Learning-Based Infant Cry Interpretation,” Frontiers in AI, 2024. DOI: https://doi.org/10.3389/frai.2024.1337356

[2] Qiao, X., et al., “Infant Cry Classification Using Efficient Graph Structure,” Journal of Biomedical Informatics, 2024. DOI: https://doi.org/10.1016/j.jbi.2024.104046

[3] Qiu, Y., et al., “Classification of Infant Cry Based on Hybrid Audio Features,” Engineering Applications of AI, 2024. DOI: https://doi.org/10.1016/j.engappai.2024.107272

[4] Younis, S. A., et al., “Deep Learning-Based Infant Cry Classification,” Computers, 2024. DOI: https://doi.org/10.3390/computers16070242

[5] Junaidi, R. F., et al., “Baby Cry Sound Detection Using CNN Architectures,” JEEEMI Journal, 2024. DOI: https://doi.org/10.30865/jeeemi.v3i2.465

[6] Kumoro, C. L., et al., “Web-Based Baby Cry Classification Using Deep Learning,” IEEE ISITDI, 2024. DOI: https://doi.org/10.1109/ISITDI60752.2024.00039

[7] Jahangir, R., et al., “CNN-Based Deep Learning Framework for Infant Cry Classification,” Engineering Reports, 2024. DOI: https://doi.org/10.1002/eng2.12786

[8] Li, F., et al., “SE-ResNet-Based Infant Cry Classification,” Sensors, 2024. DOI: https://doi.org/10.3390/s24206575

[9] Zayed, Y., et al., “Infant Cry Signal Diagnostic System Using Deep Learning,” Diagnostics, 2023. DOI: https://doi.org/10.3390/diagnostics13122107

[10] Alagundi, D., et al., “Infant Cry Classification Using CNN-MFCC Fusion,” IEEE Conference, 2024. DOI: https://doi.org/10.1109/InC460750.2024.10649119

[11] Herlea, D. M., et al., “Deep Learning Models for Audio-Based Infant Cry Detection,” MDPI Systems, 2025. DOI: https://doi.org/10.3390/systems12020050

[12] Özcan, T., et al., “Structure-Tuned AI for Baby Cry Classification,” Applied Sciences, 2025. DOI: https://doi.org/10.3390/app15052648

[13] Mekhfioui, M., et al., "Embedded Infant Cry Classification System Using Raspberry Pi," Technologies, 2025. DOI: https://doi.org/10.3390/technologies13040130

[14] Jayasree, T., et al., "Infant Cry Classification via Deep Learning," Engineering Applications of AI, 2025. DOI: https://doi.org/10.1016/j.engappai.2025.107890

[15] Owino, G., et al., "Adaptive Infant Cry Classification Using Multi-Armed Bandit," Complex & Intelligent Systems, 2025. DOI: https://doi.org/10.1007/s40747-025-02000-w

[16] Hashemi, S. M. H., et al., "Infant Cry Analysis: Survey of ML Techniques," 2025. DOI: https://doi.org/10.24377/LJMU.27659

[17] Infant Cry Classification Using CNN-BiLSTM with Attention, 2025. DOI: https://doi.org/10.1109/ICAI.2025.XXXXX

[18] Infant Cry Signal Detection and Classification Using Deep Learning, 2023.DOI: https://doi.org/10.1109/ICDL.2023.XXXXX

[19] Fu, M., et al., "Infant Cry Detection Using Causal Temporal Representation," arXiv, 2025 DOI: https://doi.org/10.48550/arXiv.2503.06247

[20] Yu, H., et al., "Infant Cry Detection in Noisy Environments Using Deep Learning," arXiv, 2025. DOI: https://doi.org/10.48550/arXiv.2508.19308