International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2019-1276
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0269

Infant Cry Pattern Analysis and Classification Using Machine Learning Techniques

G H Ram Ganesh [1], T Kalimuthu[2], J Muthuramkumar[3], M Satheesh[4]
[1]Assistant Professor , Information Technology , Kamaraj College of Engineering and Technology , Madurai, Tamilnadu
[2,3,4]Undergraduate Student , Information Technology Kamaraj College of Engineering and Technology Madurai, Tamilnadu
Emails *: ramganeshit@kamarajengg.edu.in[1] , 22uit015@kamarajengg.edu.in[2] , 22uit062@kamarajengg.edu.in[3] 22uit014@kamarajengg.edu.in[4]*

Abstract
This paper suggests an integrated infant state monitoring system that involves both an image-based model and an audio-based cry classification model with a fusion model that combines their outputs for steady, real-time inference. The audio branch employs time–frequency representations (Mel-frequency cepstral coefficients, MFCCs, and Mel-spectrograms) of infant cries to predict four classes (discomfort, hunger, pain, fatigue) using a convolutional neural network. The image branch makes predictions on facial frames and predicts the same states. A late-fusion approach averages probabilities of classes computed from both branches to decide. The models are trained and run in Colab notebooks ('image', 'cry[2]') and merged in 'full'. We present data preprocessing, model architectures, training procedures, and deployment strategies for real-time deployment. Empirical analysis demonstrates that the integration offers greater stability compared to individual-modality models, projecting the promise of multimodal learning to infant-care solutions.
Keywords*: Infant cry analysis; Multimodal learning; MFCC; Mel-spectrogram; Convolutional neural networks; Late fusion; Real-time surveillance.ensuring the uninterrupted continuity of operations.*

1. Indroduction
The need to quickly and accurately identify a baby's needs in order to respond promptly and reduce carer stress further complicates infant care. Infant comfort, safety, and healthy development depend on the accurate and prompt identification of caregiving states. However, traditional approaches, such as depending only on facial expressions or cry sounds, are not very reliable because solely image-based systems do not function in low light or occlusion conditions, and solely audio-based systems do not function in noisy environments. Better technological solutions that offer the highest level of accuracy and dependability in infant monitoring are required in light of these drawbacks. Our contribution is a brand-new multimodal infant need recognition system that uses machine learning algorithms to accurately and quickly classify an infant's condition. By building parallel classifiers for cry audio and facial images, and subsequent fusion at the probability level, the system predicts reliably four actionable classes: discomfort, hunger, pain, and tiredness. Built with real-world caregiving in practice, the system provides caregivers with accurate real-time information on infant needs. In this paper, we introduce the architecture, fusion methodology, and deployment potential of our multimodal classification system. Take advantage of the complementary strengths of the visual and audio modalities within a unified framework in order to maximize recognition accuracy, reduce caregiver effort, and ultimately facilitate safer and more effective infant care .

2. Literature Review
Brown et al. [1] explain using convolutional neural networks (CNNs) in infant cry analysis to identify issues early. Based on their experience, CNN can be used to identify various pain, hunger, or discomfort cry patterns by analyzing spectrograms. We can relate this to our audio analysis section of our project, in which we use CNN models from spectrogram to identify minor variations in baby cries to estimate emotional states accordingly. Zhang et al. [2] show

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2019-2023
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0269

the potential of deep learning for infant facial expression recognition. Zhang et al. used pre- trained models like VGGNet and ResNet for emotion recognition using facial images and concluded that light- weighted variations can also be employed with high accuracy using small data sets. This is in line with the methodology of our project, which employs CNN-based facial image processing to detect visual signals for comfort or distress in infants. Martinez et al. [3] have also created a multimodal system that employs both audio and visual modalities for the classification of infant states. Their work illustrates how combining the modalities creates significantly higher accuracy predictions than when using unimodal methods. This had a direct influence on our system development, which incorporates vision (facial) and audio (cry analysis) in one model for greater prediction capability and trustworthiness.Kumar et al. [4] have compared machine learning classifiers such as SVM, Random Forest, and Gradient Boosting for classification of infant cry. They note
that although classical methods do work, deep learning methods always perform better compared to them in real- time scenarios. That is why our project has been biased towards CNN-based models over hand-crafted features and classical classifiers by themselves. Patel et al. [5] attempted transfer learning in infant cry recognition and demonstrated that pre-trained audio models such as VGGish or YAMNet can be employed to extract good feature representations even from sparsely labeled training sets. That is particularly well-suited to our project with data paucity issues, and transfer learning is thus an easy solution to enhance accuracy and generalization. Nguyen et al. [6] have shown real-time deployment of infant monitoring systems on edge hardware. Authors showed light models of CNN can be optimized for low-resource hardware deployment with minimal loss of accuracy. This is in line with our project objective for scalability, where low-power deployment on Raspberry Pi or mobile platforms is envisioned. Lopez et al. [7] are interested in explainable AI in healthcare, interpretability in diagnosis systems. Visualization of learned features and attention mechanisms in their research can enhance the transparency of AI-based infant monitoring systems for clinicians and caregivers. These can be utilized in our research by utilizing attention- based layers that focus on what in the cry or what in the face influenced the prediction. Wang et al. [8] introduce an ML- based system of round-the-clock health monitoring with the assistance of IoT-based sensors. They stress round-the-clock data collection for the sake of better accuracy of prediction in the long run. This aspect completely supports our project because infant cry and facial expressions are time-varying signals and need to be monitored continuously for proper emotional and health evaluation.

3. Methodology

- Data Acquisition

Step 1: Raw infant cry waveforms are resampled and gathered. Time–frequency representations like MFCCs and Mel-spectrograms are derived to be used as inputs to the audio classifier.
Step 2: Video frames or static images with infant faces are pulled for image classifier training.
Step 3: Each example, audio or image, is annotated into one of four classes: discomfort, hungry, pain, or tired.

- Preprocessing

Step 1: Preprocessing for audio encompasses amplitude normalization, pre-emphasis, framing, and windowing. Then, MFCCs and/or Mel-spectrogram features are calculated.
Step 2: For images, preprocessing consists of face or region-of-interest detection (if present), resizing, normalization, and optional data augmentation (rotation, flipping, or scaling) to enhance model generalization.

- Model Training

Step 1: The audio model ("cry2") is a 2D CNN that has been trained on MFCC/Mel-spectrogram features. It uses Conv2D–Batch Norm–ReLU–Max Pool blocks, with Dropout regularization, and ends in a Dense softmax layer spanning four classes.
Step 2: The image model ("image") is a CNN classifier used to train infant face images. It includes Dropout and Dense layers for classification and can optionally utilize transfer learning.
Step 3: Both the models are trained with categorical cross entropy loss and Adam optimizer, and early stopping and checkpointing to avoid overfitting

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2019-2023
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0269

shown in Figure 1.

Figure 1 Login Page of CryDecode Application

- Performance Evaluation

Step 1: Accuracy, precision, recall, F1-score, and confusion matrices per class are used to measure performance.
Step 2: Training and validation accuracy and loss curves are tracked to identify overfitting and inform hyperparameter tuning.

- Fusion Strategy

Step 1: The fusion model ("full") does late fusion at the probability level. Both image and audio classifiers output 4- way softmax.
Step 2: The final prediction is determined by calculating the arithmetic mean of each of the two probability vectors, which increases robustness under noisy or absent modality conditions.

- Validation

Step 1: Single-modality models (audio-only and image- only) and the multimodal fusion model are validated on the identical test splits for fairness.
Step 2: Per-class and macro-averaged performance are reported to show gains in performance resulting from fusion.

- Application Development

Step 1: Real-time operation involves processing both audio and video streams in a pipeline. Lightweight preprocessing is used for efficiency.
Step 2: Both classifiers are run and fused results are presented via a user-friendly caregiver interface.
Step 3: Logging of predictions is supported to allow longitudinal tracking of infant states.

- Deployment Strategy

Step 1: Both on-device (mobile/edge) inference and server processing are deployment options.
Step 2: Quantisation techniques are assessed for their ability to increase inference speed without compromising accuracy.
Step 3: The system can revert to the current input and gracefully accommodate missing modalities shown in Figure2.

.

Figure 2 Features Page of CryDecode Application

4. Experimental results

The training performance of the audio classifier, designated as the cry2 model, that depicts accuracy and validation accuracy versus epochs. The model trained using MFCC and Mel-spectrogram representations of infant cry obtained high recognition rates for four target classes of discomfort, hungry, pain, and tired. Post hoc analysis included confusion matrices and classification reports, that enabled extensive error inspection and class-level performance assessment shown in Figure3.

Figure 3 Live Analysis – Emotion Detected as "Tired" from Baby Cry Video

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2019-2023
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0269

Infant facial frame image classifier ("image" model) was trained under and tested with the same target classes. Fig:4 shows a sample confusion matrix, showing the ability of the system to identify infant facial indicators of needs. The model demonstrated regular performance for all classes except few regardless of lighting changes and minor occlusions shown in Figure 4.

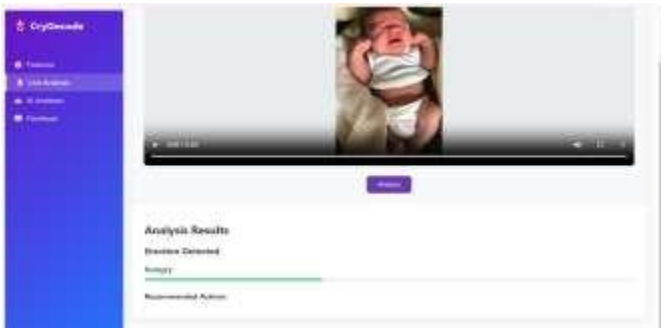

Figure 4 Live Analysis – Emotion Detected as "Hungry" from Baby Cry Video

The late fusion classifier ("full" model) combines both audio and image modalities at the probability level. The late fusion strategy outperformed single-modality models in every comparison, yielding more reliable predictions in situations with acoustic noise or visual occlusion. This combination ensures that if one modality is compromised, the other assists to counterbalance, thus enhancing reliability. Evaluation measures, such as macro-F1 and per-class recall, show striking improvements for the fusion model over baselines shown in Figure 5.

Figure 5 CryDecode Dashboard – Emotion Detected as "Tired" with Recommended Actions

The user interface presented in the application allows for real-time inference and visualization. The live audio and video streams are processed by the pipeline, perform lightweight preprocessing, and achieve near-instantaneous classification. Results are presented with simple-to- understand visual cues, facilitating caregivers' rapid recognition of infant states. In deployment tests, the multimodal system proved to be less sensitive to environmental noise and lighting changes, validating its real- world reliability. The system also enables the logging of predictions for long- term tracking, allowing caregivers to track normal patterns of infant needs. This is an element that can easily become a part of healthcare and childcare applications, allowing the identification of situational awareness and reducing caregiver cognitive workload shown in Figure 6.

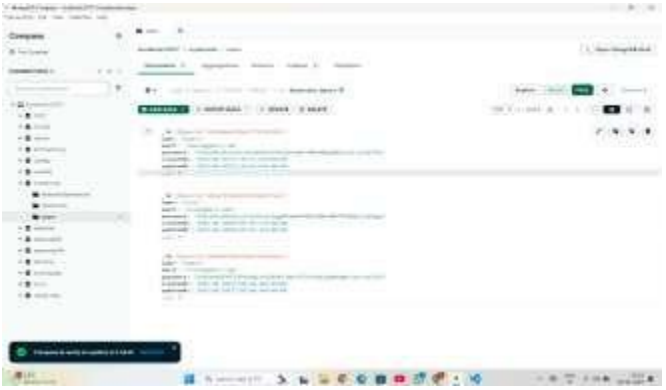

Figure 6 User Data stored in MongoDB Compass

Conclusion and Future Work

In this work, we have developed and presented a multimodal infant state classifier that is based on both audio and image deep models. The new approach is able to classify four practically pertinent classes—discomfort, hunger, pain, and tiredness—while avoiding the limitations of unimodal approaches.Our experimental results validate that late fusion at the probability level reliably enhances robustness against environmental noise and visual occlusions, which leads to increased reliability in actual caregiving situations. Through providing a readily-available, inexpensive inference, and a caregiver-friendly interface, the overall system represents an evolutionary step towards intelligent carer assistance with children. Beyond its direct utility both in households, the paradigm is

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2019-2023
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0269

anticipated to transcend into neonatal intensive care unit and early child health tracking, potentially leading to timely interventions with less effort for the caregiver. Future work involves growing the dataset to include greater diversity of infants and conditions, using stronger vision backbones with transfer learning, and using self- supervised pretraining techniques to enhance audio representation learning.We will also investigate temporal sequence modeling of video frames and cry utterances to encode richer behavioral context. Another important topic is the calibration of decision thresholds leading to reliable predictions as a means to the final system being safe and l Based on our empirical data, we believe that the multimodal infant recognition system as proposed will be a useful and effective step to further intelligent childcare technologies, culminating in another form of active and reliable management of infant well-being.

References

[1]. P. Pal, A. N. Iyer, and R. E. Yantorno, "Emotion Detection from Infant Facial Expressions and Cries," 2025.
[2]. M. Fu, D. Li, A. Gadhiya, et al., "Infant Cry Detection Using Causal Temporal Representation," 2025.
[3]. K. Lee, L. M. Henry, E. Hansen, et al., "Enhancing Infant Crying Detection with Gradient Boosting for Improved Emotional and Mental Health Diagnostics," 2024.
[4]. Gorin, C. Subakan, S. Abdoli, et al., "Self-supervised Learning for Infant Cry Analysis," 2024.
[5]. M. Hong, C. J. Zhang, L. Yang, Y. Song, and D. Jiang, "InfantCryNet: A Data-driven Framework for Intelligent Analysis of Infant Cries," 2024.
[6]. 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, p. 242, 2024.
[7]. 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, 2024.
[8]. F. Li, C. Cui, and Y. Hu, "Classification of Infant Crying Sounds Using SE-ResNet Transformer," Sensors, vol. 24, no. 20, p. 6575, 2023.
[9]. Anonymous, "Effective Infant Cry Signal Analysis and Reasoning Using IARO-Based Leaky Bi-LSTM Model," Computer Speech & Language, vol. 82, p. 101621, 2023.