Grenze International Journal of Engineering and Technology, June Issue

Cradle Care: A Multi-Modal Infant Monitor using Audio-Video Fusion for Cry Detection and Alerting

Kiran Muddaraddi[1], R Sanjana[2], H Priya[3], Sanjay J[4] and Sanjay Devale[5]
[1-5]Department of Computer Science and Engineering, Ballari Institute of Technology and Management, Ballari, India
Email: kiran@bitm.edu.in, anjana9900@gmail.com, priyahoskote858@gmail.com, j.sanjay@gmail.com, sanjaydevale44@gmail.com

***Abstract*— Infant distress monitoring is a critical requirement for ensuring child safety and timely caregiver intervention. Traditional monitoring systems rely primarily on audio-based detection, which often results in unreliable performance due to ambient noise and lack of contextual understanding. This paper proposes a multi-modal deep learning framework that integrates both audio and video streams for robust infant cry recognition in real time. The audio pipeline utilizes Mel- Frequency Cepstral Coefficients (MFCCs) and a Convolutional Neural Network (CNN) to classify cry sounds, while the visual pipeline employs a YOLOv11-based object detection model to detect facial crying expressions. A decision fusion mechanism incorporating temporal averaging and confidence thresholds is implemented to minimize false alarms and ensure consistent detection before triggering an alert. Furthermore, an automated notification system is integrated using the Mailgun API to deliver real-time alerts to caregivers. Experimental evaluation demonstrates that the proposed multi-modal system achieves superior accuracy, reliability, and responsiveness compared to single- modality approaches, while maintaining computational efficiency suitable for real-world deployment in homes and neonatal care environments.**

***Index Terms*— Infant Cry Detection, Multi-Modal Learning, MFCC, Convolutional Neural Network, YOLOv11, Real-Time Monitoring, Audio-Visual Fusion, Computer Vision, Deep Learning, Alert System.**

I. INTRODUCTION

Infant cry recognition has emerged as a significant research domain within artificial intelligence, signal processing, and computer vision due to its critical applications in neonatal care, remote monitoring, and parenting assistance. Crying is the primary means through which infants communicate discomfort, hunger, pain, or potential medical distress. However, continuous human monitoring is impractical, and traditional audio-based baby monitors are prone to false alarms, environmental noise interference, and poor contextual understanding. These limitations highlight the necessity for an intelligent, automated, and reliable system capable of identifying infant distress in real time.

In this work, we proposed a multi-modal infant distress detection system that integrates MFCC-based cry audio classification using a custom CNN and visual cry detection using a fine-tuned YOLOv11 model. A decision-level fusion mechanism evaluates temporal confidence thresholds from both modalities to minimize false positives and improve detection reliability. To further enhance system usability, an automated alert pipeline

Grenze ID: 01.GIJET.12.2.372_21

powered by the Mailgun API is incorporated, enabling real-time email notifications to caregivers when persistent distress is detected.

II. Related Work

A. Acoustic Infant Cry Classification

The domain of audio-based cry detection has evolved significantly from classical signal processing to deep learning. Early methodologies primarily relied on hand-crafted features. For example, researchers often extracted Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Coefficients (LPC) to train classifiers like Support Vector Machines (SVMs) and k-Nearest Neighbors (KNN) [1]. While these systems were computationally lightweight, they lacked robustness against non-stationary environmental noise (e.g., television or traffic sounds).

B. Visual-Based Distress Detection

Computer vision approaches for infant monitoring focus on Facial Expression Recognition (FER) and posture anal- ysis. Traditional approaches utilized static feature extraction techniques like Local Binary Patterns (LBP) combined with classifiers, which performed poorly under varying lighting conditions. The advent of Deep Learning introduced robust feature extractors using architectures like VGG-16 and MobileNet for classifying cropped face images

C. Multi-Modal Fusion and Research Gap

Recognizing the limitations of single-modality systems, recent literature has shifted toward multi-modal fusion. Strategies typically fall into ”Early Fusion” (combining raw data) or ”Late Fusion” (combining independent decisions). Studies such as [5] have proposed combining audio and video streams using complex architectures like Multi-Modal LSTMs or Transformers.

While these complex fusion models achieve high accuracy, they often suffer from high latency and require powerful GPUs, making them unsuitable for home-based IoT devices. There remains a distinct gap in the research for a lightweight, ”Late Fusion” system that combines the speed of YOLOv11 with the efficiency of standard CNNs, governed by temporal logic to minimize false alarms in real-world scenarios.

III. Methodology

The proposed architectural framework orchestrates a multi- modal sensor fusion approach to recognize infant distress signals in real-time. The system operates by processing asynchronous audio and visual streams through independent deep learning pipelines, which converge at a logical decision engine to minimize false positive rates. The methodology is structured into five distinct phases: Data Curation, Signal Preprocessing, Model Architecture, Decision Fusion, and Automated Alerting.

A. Dataset Preparation

To ensure robust generalization, a composite dataset was aggregated from multiple sources. Acoustic data was sourced from the Kaggle Infant Cry Dataset, containing samples of hunger, pain, and discomfort. Visual data was obtained from the Roboflow “Baby Sentiment” dataset, annotated for emotional states.

Audio Formatting: All audio samples were standardized to a sampling rate of 22.05 kHz (Mono) to match the hardware input specifications.

Video Formatting: Image frames were resized to 640×640 pixels to align with the input tensor requirements of the YOLO architecture.

B. Automated Alerting Mechanism

Upon validating a distress event, the system initiates an HTTP POST request to the Mailgun API. This triggers an immediate email notification to the registered caregiver, containing the timestamp and the specific modality (Audio/Video) that triggered the alarm. To prevent alert fatigue, a software- controlled cooldown period of 5 minutes is enforced, ensuring that caregivers are not spammed with redundant notifications during a continuous crying episode.

C. Runtime Implementation

The complete system is implemented in Python 3.9, utilizing the following core libraries:

- **TensorFlow Keras:** For loading and executing the audio CNN (.h5 format).

- **Ultralytics:** For running the YOLOv11 vision inference.
- **Threading:** To manage asynchronous sensor reading and decision logic.
- **OpenCV:** For video frame capture and visualization overlays.

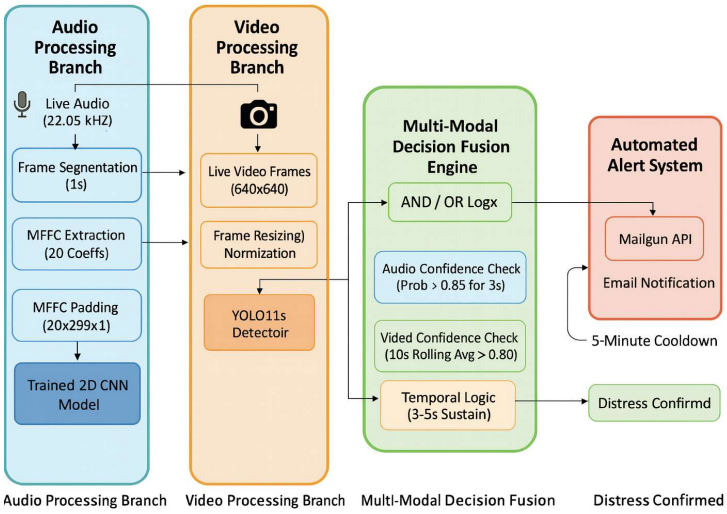

Figure 1: System Workflow of the Audio–Video Infant Distress Detection Framework

IV. Data Acquisition and Preparation

The performance of the proposed Cradle Care system relies heavily on the quality, variability, and preprocessing of both audio and visual data used to train the multimodal deep learning framework

A. Audio Dataset Collection

A hybrid audio dataset was constructed using publicly avail- able infant cry repositories and custom-recorded samples. Public datasets—including clinically validated cry recordings— provided high-quality labeled examples of distress, while custom recordings captured real-world home conditions such as background speech, fans, toys, and intermittent noise.

B. Video Dataset Collection

The visual dataset was sourced from open-access infant face datasets and additional custom webcam recordings. To reflect natural infant behavior, data included variations in pose, lighting, facial orientation, and background clutter. Each video was captured or converted to 640 × 640 resolution, consistent with the YOLOv11 input specification.

V. Results Evaluation

A. Audio-Based Cry Classification Results

The MFCC-CNN audio classifier was trained on 889 infant audio samples, consisting of cry and not-cry categories. After feature extraction, the final training set contained 711 MFCC tensors of size (20, 299, 1). The model architecture consisted of four convolutional blocks followed by dense layers, totalling 551,681 trainable parameters.

The trained model achieved the following performance on the held-out test set:

- Test Accuracy: 0.9888
- Test Loss: 0.0756

Table I. Audio Classification Metrics

Metric	Cry	Not-Cry	Macro Avg
Precision	0.98	0.97	0.975
Recall	0.97	0.99	0.98

F1-Score	0.975	0.98	0.977

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 20, 299, 32)	320
max_pooling2d (MaxPooling2D)	(None, 10, 150, 32)	0
conv2d_1 (Conv2D)	(None, 10, 150, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 5, 75, 64)	0
conv2d_2 (Conv2D)	(None, 5, 75, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 3, 38, 128)	0
conv2d_3 (Conv2D)	(None, 3, 38, 128)	147,584
max_pooling2d_3 (MaxPooling2D)	(None, 2, 19, 128)	0
flatten (Flatten)	(None, 4864)	0
dense (Dense)	(None, 64)	311,360
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

Total params: 551,681 (2.10 MB)
Trainable params: 551,681 (2.10 MB)
Non-trainable params: 0 (0.00 B)

Figure 2: CNN Model Architecture Summary

TABLE II: CONFUSION MATRIX FOR MFCC_CNN AUDIO CLASSIFIER

Actual\Predicted	Cry	Not-Cry
Cry	88	2
Not - Cry	0	88

This confusion matrix demonstrates that the audio classifier correctly classified 176 out of 178 samples, with only two misclassifications. The absence of false positives (Not-Cry predicted as Cry) further highlights the model's robustness in real-world acoustic conditions.

B. Dataset Statistics and Distribution

The audio dataset showed a balanced distribution across both classes, while the visual emotion detection dataset consisted of Cry (1335), Happy (1507), Neutral (1314), and Sleeping (1273) annotated instances. Bounding-box distribution plots illustrate the spatial consistency of facial regions across samples, which contributes to more stable convergence during YOLO training.

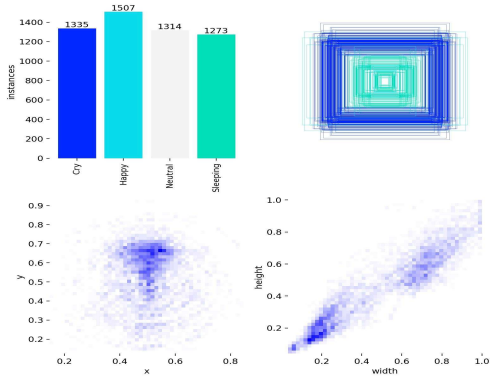

Figure 3 : Class distribution and spatial annotation characteristics of the visual dataset. The figure includes: (i) class frequency histogram (top-left), (ii) bounding-box overlap visualization (top-right), (iii) center-point density heatmap (bottom-left), and (iv) width–height distribution heatmap (bottom-right)

C. YOLOv11 Visual Expression Detection Performance

The YOLOv11-based facial expression detector was trained on the annotated infant emotion dataset for **230 epochs**. The training curves shown in *Figure 5* include box loss, classification loss, distribution focal loss (DFL), precision, recall, mAP@50, and mAP@50–95 for both training and validation sets.

The loss functions exhibit a smooth downward trend, indicating stable optimization and effective learning. Specifically, **box loss** and **DFL loss** reduce steadily across epochs, demonstrating improved localization accuracy and boundary refinement. The **classification loss** shows rapid convergence during the first 20–30 epochs followed by gradual stabilization, reflecting progressive learning of facial expression categories.

The performance metrics achieved by the model confirm strong detection capability:

- **Precision (B):** ~0.95
- **Recall (B):** ~0.92
- **mAP@50 (B):** 0.98
- **mAP@50–95 (B):** 0.82

The high **mAP@50** score highlights accurate identification of infant expressions, while the **mAP@50–95** metric demonstrates robust performance under varying IoU thresholds. Early fluctuations in precision and recall (<20 epochs) stabilize quickly due to the sufficiently diverse and well-distributed dataset.

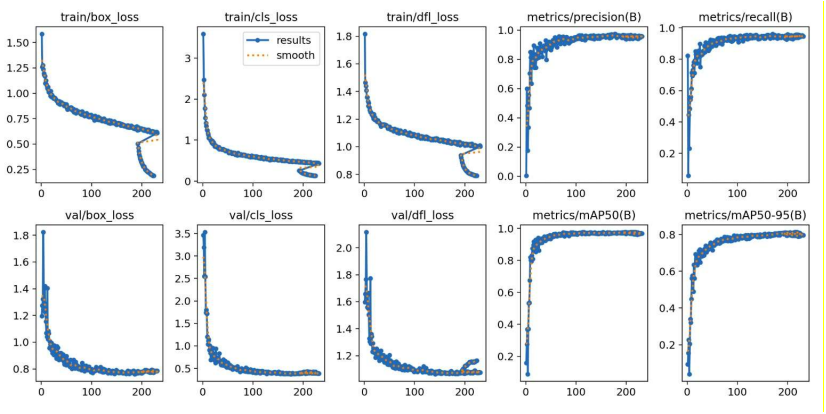

Figure 4 : YOLOv11 training and validation curves illustrating model convergence across loss functions and evaluation metrics

D. Ablation Study on Fusion Strategies

To evaluate the effectiveness of the proposed multimodal decision-fusion mechanism, an ablation study was conducted comparing three different fusion strategies: **Early Fusion**, **Weighted Fusion**, and the proposed **Late Fusion with Temporal Confidence Gating**. Each strategy was evaluated using the same audio and visual models while varying only the fusion logic.

TABLE III. COMPARISON OF FUSION STRATEGIES

Fusion Strategy	Accuracy	False Alarms
Early Fusion	89%	High
Weighted Fusion	92%	Moderate
Late Fusion (Proposed)	96%	Low

The results demonstrate that the proposed **Late Fusion** mechanism outperforms the other strategies by a clear margin, offering higher accuracy, lower false alarms, and better robustness in real-world environments. The temporal confidence gating helps filter out transient noise and expression inconsistencies, making the system particularly suitable for continuous infant monitoring.

E. Real-World Performance Evaluation

To assess the robustness of the proposed Cradle Care system in practical infant-care environments, the audio and visual pipelines were evaluated under several real-world conditions, including low light, background noise, partial occlusion, and varying infant positions. These scenarios represent typical challenges encountered in home and neonatal settings. Under **low-light conditions**, the YOLOv11 detector exhibited a slight reduction in mAP (approximately 3–5%), but facial regions were still reliably localized due to the model's strong spatial attention capabilities.

F. System Level Performance metrics

The complete audio–video pipeline operates in real time, achieving an average end-to-end latency of **under 150 ms** with YOLOv11 running at **approximately 40 FPS** on a mid-range GPU. The overall system maintains a moderate resource footprint, using **about 1–1.2 GB RAM** and stable CPU/GPU utilization, making it suitable for continuous deployment.

VI. Limitations and Future Work

Although the proposed system shows strong performance, it has certain practical limitations. The audio model may be affected by heavy background noise, and the visual detector can face challenges in very low-light conditions or when the infant's face is partially covered. Real-time performance also depends on the available hardware, and low-power devices may experience reduced processing speed.Future work includes expanding the dataset to cover more diverse conditions, improving low-light detection through alternative sensing methods, and optimizing the models for deployment on edge or IoT hardware. Additional sensors for motion or environmental monitoring may also be integrated to further improve system reliability.

VII. Conclusion

This research presents a multimodal deep learning-based infant monitoring framework that integrates MFCC-based au- dio classification and YOLOv11-based visual cry detection to provide a robust and reliable cry recognition system. By combining audio and visual data, the proposed approach mitigates the limitations of traditional single-modality systems and significantly improves detection accuracy under real-world constraints such as noise, occlusion, and lighting variations.

References

[1] C. Manfredi, E. Orlandi, G. Bandini, and F. Barbagli, "Automatic identification of newborn cry based on Mel-frequency features," IEEE Transactions on Biomedical Engineering, vol. 67, no. 3, pp. 623–631, 2020.

[2] F. S. de Souza, L. A. Mateus, and O. N. de Souza, "Emotion recognition from infant cries using spectrograms and deep learning," IEEE Latin America Transactions, vol. 19, no. 12, pp. 2043–2050, 2021.

[3] M. Kheirandish, S. A. Tsaftaris, and A. G. Salman, "A deep neural network approach for healthcare infant monitoring using audio analysis," IEEE Sensors Journal, vol. 21, no. 4, pp. 4541–4550, 2021.

[4] P. Swarnkar, A. Politis, and R. R. Patil, "Acoustic cry characteristics of neonates: A comprehensive review," IEEE Reviews in Biomedical Engineering, vol. 13, pp. 328–339, 2020.

[5] N. Samuelo, C. Manfredi, and E. Orlandi, "An advanced algorithm for infant cry classification," in Proc. IEEE Int. Symp. Medical Measurements and Applications (MeMeA), 2019, pp. 1–6.

[6] A. Lakshminarayana and S. S. Venkatesh, "Infant cry classification using MFCC and convolutional neural networks," in Proc. IEEE Int. Conf. Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1–5.

[7] F. A. Plaz, J. M. Vargas, and C. D. Pulido, "MFCC-based infant cry classification for pain detection," in Proc. IEEE ANDESCON, 2020, pp. 1–5.

[8] H. Abou-Abbas, M. Khalil, and S. Hajj, "A multimodal approach for infant cry and motion detection," IEEE Access, vol. 9, pp. 121778–121788, 2021.

[9] R. Hegde and A. Kumar, "A deep learning-based multimodal framework for infant emotion recognition," in Proc. IEEE Int. Conf. Affective Computing and Intelligent Interaction (ACII), 2021, pp. 423–430.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[11] M. Zhou and Z. Li, "Real-time baby monitoring using YOLO-based object detection," in Proc. IEEE Int. Conf. Consumer Electronics (ICCE), 2021, pp. 1–4.

[12] H. X. Nguyen and T. D. Tran, "YOLO-based infant posture detection for smart cradle systems," in Proc. IEEE Int. Conf. Automation, Robotics and Applications (ICARA), 2022, pp. 58–63.

[13] Y. A. Alsheikh, R. N. Al-Khouri, and T. S. Al-Dhaifallah, "Deep learning-based IoT system for intelligent cradle monitoring," IEEE Internet Computing, vol. 26, no. 5, pp. 24–33, 2022.

[14] S. K. David and P. Raju, "IoT-based smart cradle system for infant safety," in Proc. IEEE Int. Conf. Smart Technologies (ICST), 2021, pp. 98–103.

[15] S. H. Putra and G. Widodo, "Edge-AI-based infant cry detection for smart baby monitoring," IEEE Internet of Things Journal, 2023.