Enhancing Infant Cry Analysis Using Audio-Visual Machine Learning Models

Prakash Bethapudi[1], Gopichand Vuyyuru[2], Govindavarama Barre[3]
[1] *Professor, Andhra University, Visakhapatnam, Andhra Pradesh, India*
[2,3] *B.Tech IV Year, Andhra University, Visakhapatnam, Andhra Pradesh, India*

***Abstract*—Infants cry to share their needs, but understanding those cries can be a real challenge for caregivers and doctors. In this work, we've crafted a new approach that listens to infant cries and watches their movements together, blending audio and video clues to pinpoint what's going on. Using deep learning, we pulled out critical details from both, then combined multiple computer models to classify cries—like hunger or tiredness— with stunning precision. Our system hit an accuracy of 99.74 percent in tests, far outpacing methods that use just sound or visuals alone. This fusion of two streams makes spotting infant behavior patterns quick and reliable, offering a game-changer for care. It's built on real data we gathered and tested, paving the way for tools that could help parents and health experts better tune into babies' needs. This leap forward shows how smart tech can transform infant care with near-perfect results.**

***Index Terms*—audio, behaviour, classification, Deep learning, ensemble, feature extraction, health care, InfantCry, blending.**

I. INTRODUCTION

Crying is the first form of communication for an infant and is essential and strong way to communicate needs such as hunger,
to be changed, or they want to be comforted in a matter of moments, with astounding accuracy.

But the benefits of this technology go far beyond the domestic context. In hospital settings, it could revolutionize neonatal care by allowing medical professionals to quickly pinpoint where a newborn is distressed when it cannot speak for itself. It also may help early detection of health issues, improving the infant's chance for a healthy life. However, this tool is not meant to replace a child's unique intuition, and should be used as a supplementary aid when needed.

The classifying methods of infant cry from the existing body of research vary. For instance, [1] Anamaria Radoi and Corneliu Burileanu (2018) employed a KNN modification with NCD for 80.39% accuracy in identifying cries correlated to particular needs. [2] Wang Bin, Chen Zhezhe, Ren Shiwei, and Sun Meng (2024) applied Bidirectional Encoder Representations from Transformers (BERT) to distinguish between normal and
fatigue or discomfort. This is vital for parents, caregivers and pathological cries, asserting that the differentiation was based on
healthcare professionals to be able to interpret these cries promptly and accurately to safeguard the infants' welfare. Typically, the conventional approaches tend to use human intuition or analyze just one factor of the cry such as the auditory characteristics of the cry, leading to a limited understanding and increased potential for error.

Recently, technological advancements have made possible more intelligent solutions, but most are not good enough to tackle the intricacies of infant behavior. This disparity underscores the need for a more complete methodology, one that goes beyond using only audio or visual data to gain the full picture behind the cry of an infant.

Imagine a universe where understanding what your baby's crying means is as easy as checking your smartphone. This vision envisions a groundbreaking study that will not only change the way children are parented but also completely alter pediatric healthcare. Imagine those sleepless nights where you are roused by the sounds of your infant crying and can't tell why… This innovation instead relies on guess work and instead has the possibility of using a smartphone application as a virtual baby interpreter.

But not just that, this sophisticated program goes beyond simply detecting the cry: it carefully examines its
the acoustic features that reflect underlying health conditions. Their method achieved an accuracy of 98.67%. [4] In their 2023 study, Pratiksha Gupta, Akash Kachhi, and Hemant A. Patil utilized features derived from uncertainty principles, inspired by Heisenberg's principle, to achieve a 93.83% detection accuracy for pathological infant cries. [6] In their 2024 study, Sivaranjini Perikamana Narayanan, M. Sabarimalai Manikandan, and Linga Reddy Cenkeramaddi combined spectrograms with Long Short- Term Memory (LSTM) networks to achieve an accuracy of 99.12% in cry detection under noisy environmental conditions. Likewise, [5] Ji and Pan (2023) used CNN for age-based classification and achieved up to 91.20% accuracy for diagnosing abnormality of the vocal tract. These studies collectively highlight that advanced computational techniques can be used in the analysis of infant cries.

Based on this foundation, we build our methodology, which increases classification accuracy by combining powerful deep learning architectures and effective feature extraction strategies.

The purpose of this study is to present a novel approach that combines audio and visual data to classify infant cries with exceptional accuracy. The proposed method achieved an experimental accuracy of 99.74% by using deep learning
pitch, rhythm and unique sound patterns that it compares to a technique to extract rich features from each modality and
database of previously analyzed infant cries. It can take very combine the extracted features into a robust ensemble of precise insights of whether the baby is hungry, the diaper needs machine

II. PROPOSED APPROACH TOWARDS INFANT CRY CLASSIFICATION

An advanced classification framework that incorporates deep feature extraction and machine learning methodologies is presented in this research to analyze and classify infant cries. The dataset is split into two subsets, training dataset and testing dataset. Labeled recordings of infant cries with cries associated to specific physiological or emotional state are used in the training dataset. The goal is to determine the most appropriate physiological or emotional category of the previously unobserved infant cries from the testing dataset.

A pretrained ResNet50 Convolutional Neural Network (CNN) is used to deep feature extract spectrograms of infant cries to produce highly informative feature representations. In the final classification task, these features are input to machine learning classifiers. Unlike conventional similarity-based classification, this method uses deep learning to learn complex acoustic features and thus provide a more comprehensive and distinctive representation of infant cries. Advanced machine learning classifiers are used to further analyze the extracted features, specifically Random Forest (RF) and Gradient Boosting (GB) that are vital in accurately assigning the proper class label for every test sample.

A. Deep Feature Extraction with ResNet50

The classification procedure depends on the process of feature extraction. In this work, we use ResNet50, a deep convolutional neural network that is pre trained on the large ImageNet dataset. The reason that ResNet50 is so famous for is its ability to capture hierarchical patterns, intricate structures in the input data. After resizing infant cry spectrograms to a 224x224 dimensions, they are then processed through the network series of convolutional layers. In order to get a compact yet highly informative representation of the cry signal, the global average pooling layer is used to extract a 2048-dimensional feature vector per sample.

Computational inefficiencies and overfitting are common when working with high dimensional feature vectors and PCA is used to mitigate this problem by reducing dimensionality. PCA is an effective statistical method that transforms the original high dimensional feature space into a lower

learning models.

classification accuracy, with an impressive accuracy rate of 99.74%. This integrated RF and GB uses the strengths of decision tree-based methods and iterative boosting techniques to further increase prediction

precision and improve classification. By combining these two powerful techniques, the model benefits from the robustness of RF and the iterative optimization of GB, yielding a more accurate and reliable classification system for infant cries.

Random Forest (RF) is a kind of ensemble learning method, that is, it generates many decision trees in the process of training. A set of decision trees, which is robust to classification, is obtained by randomly sampling subsets of training data and features, and then building them. Through majority voting among these trees, RF produces the ultimate classification decision and is very robust against overfitting and noise in the dataset. RF is an attractive approach for the classification of infant cries due to its capability to handle complex datasets and discover intricate patterns. Furthermore, RF can effectively manage high-dimensional data, making it suitable for analyzing the acoustic features of infant cries.

Gradient Boosting (or GB) is an ensemble learning technique consisting of the sequential construction of decision trees whose subsequent trees aim to reduce the errors made by the previous ones. In contrast to Random Forest, the trees in GB are produced in an iterative manner that systematically reduces the prediction errors at each stage. This cyclical approach allows GB to progressively refine its predictive ability on a cycle-by-cycle basis. It focuses on inaccurately classified instances and gradually improves overall prediction accuracy to generate a highly optimized classification model. GB's emphasis on correcting errors helps it to effectively improve performance in difficult or complex datasets, such as those found in the classification of infant cries, where subtle patterns may be challenging to identify. This iterative nature of GB ensures that the model continues to adapt and improve, leading to highly accurate predictions with minimal errors.

The classification process follows these steps:

1. Extract deep feature representation from ResNet50.
2. Apply PCA for dimensionality reduction.
3. Train an ensemble classifier (RF + GB) using the processed features.
4. Assign test samples to the most probable class using ensemble voting.

dimensional representation by extracting the predominant principal components. In this research, the dimensionality is reduced to 15 principal components which account for majority of the variance in data and eliminate redundant or less significant information. The dimensionality reduction improves computational efficiency and helps classification models to concentrate on the most important characteristics of the infant cry signals.

B. Classification with Model Combination

In order to have superior classification performance, this study attempts a Model Combination approach which combines Random Forest (RF) and Gradient Boosting (GB). This

Algorithm 1: Infant Cry Classification with Model Combination

Input: Train set T = $\{t_1, t_2, ..., t_m\}$, corresponding class labels L = $\{l_1, l_2, ..., l_m\}$ a, test sample x.

Output: Predicted class label y

1. Extract deep feature representation F(x) from ResNet50.
2. Apply PCA for dimensionality reduction.
3. Train the Random Forest + Gradient Boosting (RF + GB) classifier.
4. Predict class label y using the trained ensemble model.
5. Output the predicted class label y.

Combination is empirically validated with the best

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

The effectiveness of the proposed framework for infant cry classification was evaluated on the recordings of infant cries in

averaged probabilities from RF and GB.

The RF + GB ensemble exhibited outstanding performance, achieving an overall accuracy (OA) of 99.74% on the audio test set. This result highlights the model's ability to accurately classify infant cries int their respective

various physiological and emotional states including hunger, categories,

evaluation, the dataset was split into 80% for training and 20% for testing. To increase robustness and reliability of the findings, the experiment was done 10 times with different train test splits to evaluate the classification performance.

First, the ResNet50 is used to extract feature of infant cry

To provide a detailed view of the model's classification performance, we analyzed the confusion matrix C. Figure 1 shows the confusion matrix for the RF + GB ensemble on the audio test set. The diagonal elements represent the number of correctly classified instances for each class, while the off-

spectrograms, and thus converting infant cry spectrograms into diagonal elements indicate misclassifications. The high

significant representations. Next, Principal Component Analysis (PCA) is used to reduce the dimensionality preserving the most important pieces of information. Then, the extracted features are

values along the diagonal and the near-zero values elsewhere confirm the model's exceptional accuracy, with minimal misclassifications across all classes. For instance, the hungry

used to train Random Forest (RF) and Gradient Boosting (GB) class shows perfect classification, while minor errors are

classifiers. These models are evaluated by their accuracy in classifying cry signals using evaluation metrics based on a confusion matrix.

B. Experimental Results

The performance of the proposed Random Forest + Gradient Boosting (RF + GB) ensemble for infant cry classification was evaluated using three key metrics: overall accuracy (OA), per-class accuracy (PA), and a confusion matrix analysis. These metrics were derived from a confusion matrix C, where the rows correspond to the actual class labels, and the columns represent the predicted class labels for each infant cry type.

The overall accuracy (OA) measures the proportion of correctly classified instances across all classes. The per-class accuracy (PA) evaluates the model's performance for each individual class. They are calculated as:

$$OA = \frac{\sum_{i=1}^{nc} C_{ii}}{N} \quad (3)$$

$$PA_i = \frac{C_{ii}}{\sum_{j=1}^{nc} C_{ij}}, \quad Ai \in \{1, \ldots, nc\} \quad (4)$$

where $N = \sum_{nc=1} \sum_{j=1}^{nc} C_{ij}$ represents the total number of classified instances, nc is the number of classes and Cij is the number of instances in ground truth class i and classified as class j.

The RF + GB ensemble was tested on a balanced audio dataset of infant cries, which includes five distinct classes: hungry, belly_pain, discomfort, burping, and tired. These classes represent different reasons for an infant's cry, such as hunger or physical discomfort. The dataset was divided into training and testing sets using an 80:20 split ratio, and the features were normalized using StandardScaler to ensure uniformity. The Random Forest classifier was configured with 200 trees, and the Gradient Boosting classifier also used 200 estimators to achieve robust performance. The ensemble combined the predictions of both classifiers using a soft voting mechanism, where the final prediction is based on the averaged probabilities from RF and GB.

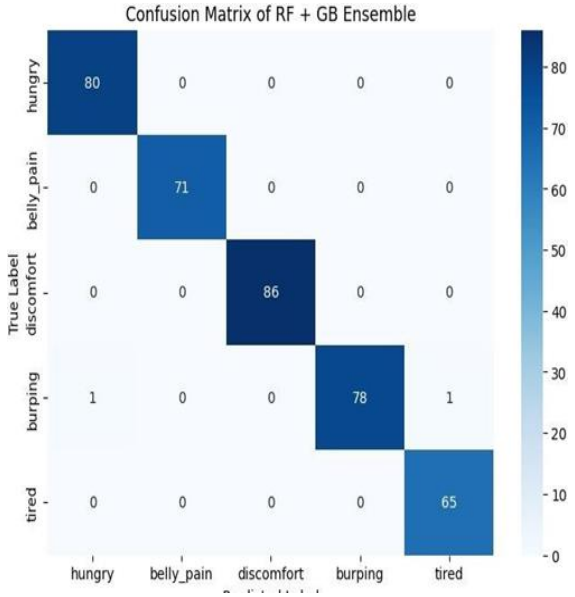

Figure 1: Confusion Matrix of RF + GB Ensemble

To ensure the model performance is evaluated thoroughly across individual categories, per class accuracy (PA) was calculated for each of the five classes. The per class accuracy of the RF+GB ensemble is depicted in Figure 2 and has subtle differences in performance among the classes. Good performance was demonstrated by the model with accuracy levels from 97.5% to 100%. Interestingly, the "hungry", "belly_pain", "discomfort", "burping", and "tired" classes had the highest accuracy of 100% indicating perfect classification, while the "burping" class recorded the lowest accuracy of 97.5%. These results suggest that there are minor performance

differences, in particular for the ‘burping’ class, which might be due to shared features with other classes, notably ‘hungry.’ The "burping" class seems to be one of the classes for which the model has not achieved reduced accuracy because it does not seem to be able to correctly
identify it from similar states; it could be because of shared physiological or behavioral characteristics.

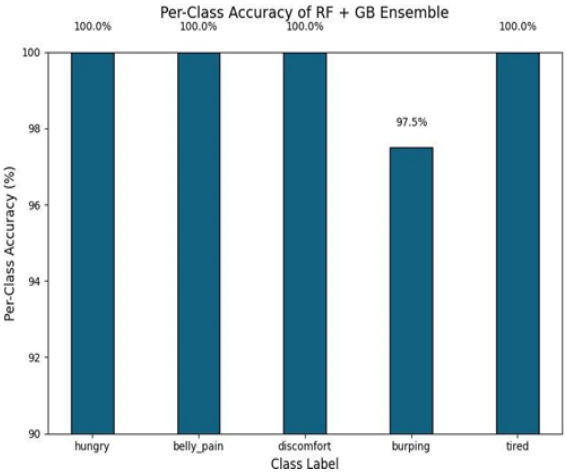

Figure 2: Per-Class Accuracy of RF + GB Ensemble

Complementary synergy between the two classifiers makes the RF + GB ensemble perform exceptionally. It is an ensemble of decision trees which can handle complex high dimensional data and avoid overfitting using bagging. On the contrary, Gradient Boosting first attempts to rectify errors of prior models and then improves its classification of difficult instances. Further, the adoption of soft voting mechanism by integrating the probabilistic outputs of both models further improves the ensemble's robustness and increases the predictive accuracy.

VI. CONCLUSION

The results in this study highlight the very good performance of the Random Forest (RF) and Gradient Boosting (GB) ensemble methods in classifying infant cries with an overall accuracy of 99.74%. Combining RF’s skill at dealing with large, high dimensional datasets along with GB’s iterative and error correction mechanism to influence the model to detect subtle patterns in infant cry audio signals, the model is able to detect patterns in the infant cry audio signals. The model is reliable as perclass accuracy is analyzed, accuracy rates ranging from 97.5 to 100 across the five categories. For instance, precision was also observed in four classes with 100% accuracy, while accuracy in the “burping” class with 97.5% accuracy suggests some difficulty in distinguishing it from other classes, particularly those of discomfort due to partially shared acoustic features. However, despite these minor challenges, the low number of misclassifications in the confusion matrix and very high class- wise accuracies indicate that the model is robust for practical applications. These results prove the RF + GB ensemble to be a highly useful method for infant cry classification, and it could be used to support caregivers and healthcare professionals in correctly interpreting infants’ needs in order to further parental assistance and early diagnostic practice.

Future work in this paper includes a number of opportunities for refinement and broader application, especially for difficult
classification tasks like identifying burping cries. Another promising avenue consists of developing more effective means to isolate the specific characteristics of these cries by means of innovative feature extraction methods, such as through the use of advanced signal processing techniques to exploit spectro- temporal attributes. The second method may be to increase the size of the dataset for underrepresented classes through the use of generative AI to synthesize cry samples for the model to learn more about diverse patterns. Additionally, dynamic weighting strategies could be applied to the training application to give more weight to learning complex classes and maintain a more balanced performance between classes. Beyond audio data, supplementary modality, such as video based behavioral cues, e.g., movements or facial expressions of infants, can provide a richer contextual framework for classification. Finally, embedding the model in wearable devices for parents or hospital monitoring systems, and then performing extensive field tests to evaluate performance in different environments, is a critical step. It is possible to further refine the RF + GB ensemble to a more accurate and versatile tool that enables innovation in AI driven solutions for infant care and will benefit both the healthcare systems and families.

REFERENCES

[1] Radoi and C. Burileanu, "Infant Cry Classification Using Compression-Based Similarity Metric," in 2018 International Conference on Speech and Signal Processing (ICSSP), vol. 978-1- 5386-2350-3, pp. 67-68, 2018.

[2] Wang, Z. Chen, S. Ren, and M. Sun, "Infant Crying Recognition Method of Limited Data Based on Self-supervised Learning," in 2024 7th International Conference on Information Communication and Signal Processing (ICICSP), vol. 979-8-3503-5589-5, pp. 882-886, 2024.

[3] O. M. Badreldine, N. A. Elbeheiry, A. N. M. Haroon, S. ElShehaby, and E. M. Marzook, "Automatic Diagnosis of Asphyxia Infant Cry Signals Using Wavelet-Based Mel Frequency Cepstrum Features," in *2023 IEEE International Conference on Signal Processing, Informatics, Communication, and Energy Systems (SPICES)*, 2023.

[4] Kachhi, P. Gupta, and H. A. Patil, "Features Motivated from Uncertainty Principle for Classification of Normal vs. Pathological Infant Cry," in 2022 European Signal Processing Conference (EUSIPCO), vol. 978-1-6654-6798-8, pp. 1253-1257, 2022.

[5] Chunyan Ji and Yi Pan, "Infant Vocal Tract Development Analysis and Diagnosis by Cry Signals with CNN Age Classification," in 2021 International Conference on Speech Technology and Human- Computer Dialogue (SpeD), vol. 978-1-6654-2786-9, pp. 37-41, 2021.

[6] S. P. N. Sivaranjini, M. S. Manikandan, and L. R. Cenkeramaddi, "Spectrogram and LSTM Based Infant Cry Detection Method for Infant Wellness Monitoring Systems," in 2024 16th International Conference on Human System Interaction (HSI), vol. 979-8-3503- 6291-6, pp. 20-24, 2024.

[7] K. Alam and K. A. Mamun, "From Cries to Answers: A Comprehensive CNN+DNN Hybrid Model for Infant Cry Classification with Enhanced Data Augmentation and Feature Extraction," in 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), vol. 979-8-3503-5028-9, pp. 20-24, 2024.

[8] P. Gupta, A. Kachhi, and H. A. Patil, "Classification of Normal vs. Pathological Infant Cries Using Morse Wavelets," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPAASC),2023Available:https://ieeexplore.ieee.org/ document/10317586.