DeepInfant

Skytells AI Research, Inc.

research@skytells.io

February 2, 2025

Abstract

DeepInfant V2 is an advanced neural network **model** that classifies infant cries to help caregivers respond to a baby's needs more effectively. This model also supports **CoreML** for Apple devices, enabling efficient real-time inference on iOS. This paper provides an in-depth overview of the collection and preprocessing of infant cry data, feature extraction techniques, and the design of the **DeepInfant V2** architecture—an integrated Convolutional-Recurrent Neural Network (CRNN). The system leverages acoustic features from short-time Fourier transforms (STFT) and Mel-scaled spectrograms to achieve robust detection of various infant distress signals. Results from experimental evaluations suggest that **DeepInfant V2** achieves superior performance (up to 89% accuracy on a diversified dataset) compared to conventional methods, offering insights into how real-time cry analysis can optimize baby care.

1 Introduction

Crying is the primary mode of communication for infants, serving as an audible indicator of their needs and well-being. Despite its importance, interpreting baby cries remains challenging, especially for new or untrained caregivers. Recent advancements at **Skytells AI Research, Inc.** have made it possible to develop a deep learning **model** that reliably identifies the reason behind an infant's cry, such as hunger, discomfort, pain, or tiredness. With built-in support for **CoreML**, the **DeepInfant V2** model can be deployed seamlessly on Apple devices for real-time baby cry analysis.

1.1 Motivation

DeepInfant emerged as a solution to the growing need for accurate and accessible baby cry interpretation. **DeepInfant V2**, the latest iteration, combines robust data augmentation techniques, advanced convolutional feature extraction, and sequence modeling via recurrent neural networks to achieve reliable cry classification in a wide range of acoustic conditions.

1.2 Contributions

- **Integrated CNN-LSTM Architecture:** We incorporate convolutional networks for spectral feature extraction and recurrent units for temporal context modeling, enabling more nuanced cry classification.
- **Enhanced Dataset & Augmentation:** We leverage multiple publicly available and in-house infant cry datasets, enriching training samples with pitch shifting, time stretching, and noise addition to increase robustness.
- **Real-Time Deployment (CoreML):** We detail a pipeline for processing live audio streams, demonstrating feasible, real-time inference on mobile and embedded systems. Specifically, the **DeepInfant V2** model supports **CoreML**, facilitating straightforward deployment on iOS devices.

2 Background and Related Work

2.1 Infant Cry Classification

Infant cry classification has been approached via acoustic analysis and pattern recognition techniques. Early studies often relied on handcrafted features such as fundamental frequency, formants, and spectral flux [1–3], combined with traditional machine learning classifiers like SVM or HMM. While moderately successful, these methods typically struggled with large, diverse datasets and significant environmental noise.

2.2 CNN and CRNN Approaches

Convolutional Neural Networks (CNNs) have become a dominant approach for audio classification tasks, particularly for environmental and speech-related sounds, due to their ability to capture local patterns in spectrogram representations [4]. When combined with Recurrent Neural Networks (RNNs) such as LSTM or GRU layers, the resulting CRNN model captures both frequency- and time-domain information, improving classification performance [5].

3 Data Collection and Preprocessing

3.1 Dataset Composition

- **Publicly Available Datasets:** We integrated multiple real-world infant cry datasets reported in the literature [1–3], each containing short labeled cry segments.
- **In-House Recordings:** A series of diverse audio recordings were collected by **Skytells AI Research, Inc.** from volunteers under controlled conditions, ensuring higher-quality data for training.

Overall, the combined dataset includes:

- **Audio Length:** 2–7 seconds (average 4.5 seconds).
- **Sampling Rate:** 16 kHz.
- **Number of Samples:** ∼10,000 labeled clips (cry and non-cry, subdivided by suspected reason).

3.2 Data Augmentation

To improve model generalization, we employed several augmentation strategies:

- **Pitch Shifting:** ±2 semitones to simulate cry variations across infants.
- **Time Stretching:** ±10% speed modifications to reflect different crying tempos.
- **Adding Noise:** Synthetic or recorded background noises resembling domestic environments (e.g., mild chatter, humming appliances).

3.3 Labeling and Tagging

All audio clips were manually labeled with potential cry reasons (e.g., hunger, discomfort, pain, tiredness). Additional metadata such as infant age range and recording environment were also documented where available. Labels were verified by a panel of caregivers, pediatric nurses, and volunteer annotators to increase reliability.

4 Methodology

4.1 Feature Extraction

Raw audio waveforms are transformed using **Short-Time Fourier Transform (STFT)** windows of size 2048 samples with a hop length of 512 samples. We then map the resulting spectrogram to the **Mel scale**, compressing frequency bins for better correlation with human auditory perception. Finally, the **log transformation** is applied to reduce dynamic range and emphasize relevant spectral variations.

$$\text{STFT}(x(t)) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi \frac{fn}{N}}, \quad \text{MelScale} \rightarrow \log(\cdot).$$

4.2 Model Architecture

DeepInfant V2 adopts a CRNN architecture consisting of:

1. **Convolutional Layers:** Multiple CNN blocks to capture local patterns in the spectral domain.
2. **Batch Normalization & Dropout:** Regularization to prevent overfitting.

3. **Recurrent Layers (LSTM):** Two stacked LSTM layers to capture temporal dependencies across audio frames.

4. **Fully Connected Layers:** Provide final classification with a Softmax layer outputting probability distributions over possible cry reasons.

Figure 1: Diagram of the DeepInfant V2 CRNN Architecture

4.3 Training Procedure

We implemented the model using **PyTorch** and **fastai**. The dataset was split into training (80%), validation (10%), and test (10%) sets. We used the Adam optimizer with a cyclical learning rate ranging from 1×10^{-4} to 1×10^{-2}. The mini-batch size was 64 for training. Each training run typically converged within 25 epochs.

5 Solving Key Challenges

During the development of **DeepInfant V2**, several challenges arose:

- **Data Variability:** Infants cry for a multitude of reasons, each potentially exhibiting distinct acoustic patterns. We addressed this by combining multiple datasets and introducing targeted data augmentations (pitch shifting, time stretching).

- **Real-Time Requirements:** Caregivers need quick feedback. We optimized the CRNN architecture and implemented CoreML exports, ensuring low-latency inference on mobile devices.

- **Robustness to Noise:** Household environments are rarely silent. Incorporating noise injections into the training pipeline allowed the model to handle real-world, noisy conditions effectively.

By systematically tackling these issues, **DeepInfant V2** delivers reliable and efficient infant cry classification in diverse scenarios.

6 Experimental Results

6.1 Evaluation Metrics

We report:

- **Accuracy**

- **Precision, Recall, F1-score** for each class

- **Confusion Matrix** to illustrate class-level performance

6.2 Quantitative Performance

Table 1 compares **DeepInfant V2** with two baseline models: `DeepInfant_VGGish` and `DeepInfant_AFP`, highlighting the superior accuracy of **DeepInfant V2**.

Table 1: Overall performance comparison on a combined dataset.

Model	Accuracy	Precision	Recall	F1
`DeepInfant_VGGish`	75.0%	72.5%	74.1%	73.3%
`DeepInfant_AFP`	78.0%	76.0%	77.5%	76.7%
DeepInfant V2	**89.0%**	**88.3%**	**88.9%**	**88.6%**

6.3 Ablation Study

An ablation study revealed that removing data augmentations decreased the system's accuracy by about 5%, illustrating the importance of pitch shifting, time stretching, and noise addition. Similarly, replacing the LSTM with a simpler GRU slightly reduced performance, confirming that the chosen architecture is optimal for this specific task.

7 Discussion

7.1 Real-World Deployment

Our experiments show that **DeepInfant V2** can operate in near-real-time on mobile devices when deployed via optimized ONNX or **CoreML** formats. This cross-platform compatibility makes the model practical for in-home baby monitors, smartphone applications, and Apple devices with low-latency requirements.

7.2 Limitations and Future Work

- **Generalization to Different Languages/Cultures:** While infant cry acoustics share universal traits, cultural and environmental factors may influence how often or how intensely babies cry.
- **Edge Cases:** Extremely noisy settings or older children's vocalizations remain challenging.
- **Additional Cry Types:** Integration of more nuanced states (e.g., colic, fear, teething) could improve system comprehensiveness.

Future research includes exploring Transformer-based architectures for more efficient temporal modeling and extending data collection efforts to reflect broader demographic variances.

8 Conclusion

This paper presents **DeepInfant V2**, a robust cry detection and classification **model** leveraging deep learning to assist caregivers in understanding infants' needs. Evaluations highlight an 89% accuracy on a combined public and private dataset, demonstrating **DeepInfant V2**'s practicality for real-time applications. By continuing to refine data augmentation, model architecture, and real-time inference pipelines—including **CoreML** support—we envision a future where AI-driven baby care technologies become integral tools for parents and pediatric professionals.

Acknowledgments

The authors express gratitude to the dedicated team at **Skytells AI Research, Inc.** for their efforts in data collection, labeling, and model validation.

References

[1] A. de la Torre and B. García-Zapirain, "Analysis of crying baby with pathologies based on instantaneous frequency and statistical features," *Mathematical Problems in Engineering*, vol. 2016, Article ID 2580917, 9 pages, 2016. `https://doi.org/10.1155/2016/2580917`

[2] C. Celikel, N. Chedid, and Q. Huang, "Infant cry signal classification using pattern recognition techniques," in *Proceedings of the 1st International Conference on Biomedical Signal and Image Processing (ICBIP)*, 2016.

[3] S. Alghowinem, J. Epps, and R. Goecke, "Towards Automatic Classification of Infant Crying in NICU," in *Proceedings of Interspeech*, 2016.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, 2015, pp. 1–6. `https://doi.org/10.1109/MLSP.2015.7324337`

[5] S. Adavanne, G. Parascandolo, P. Pertilä, T. Virtanen, "Sound event detection in multichannel audio using spatial and spectro-temporal features," in *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016, pp. 1–5. `https://ieeexplore.ieee.org/document/7760360`

GitHub Repository

The official **DeepInfant** code repository can be found at: `https://github.com/skytells-research/DeepInfant`