CryHop: Baby Cry Classification on Dunstan Dataset using Transfer Learning and Ensembling

Soheil Khooyooz*[1], Loic Ricard*[2], Jiayue Liu[1], Mostafa Haghi[1], Nima TaheriNejad[1]
[1]Heidelberg University, Germany
[2]National Graduate School of Engineering and Research Center Caen, France
Email: {soheil.khooyooz, mostafa.haghi, nima.taherinejad}@ziti.uni-heidelberg.de,
jiayue.liu@uni-heidelberg.de, loic.ricard@ecole.ensicaen.fr

***Abstract*—Identifying the cause of infant crying is essential for new parents, particularly since the task is challenging amid sleep deprivation and disrupted routines. Prior infant-cry studies often report high accuracies without accounting for variable audio durations or biases from audio resizing (cropping, padding, or resampling), which distort temporal–spectral structure and inflate performance. We circumvent these biases and present a reliable baby-cry classification method using Dunstan Baby Language (DBL), which defines five cry types: belly pain, need to burp, discomfort, hunger, and tiredness. To remove length bias while producing fixed-size inputs, we compute Mel-spectrograms with a variable hop length, yielding a constant number of time frames for the classifier. We apply Gradient-weighted Class Activation Mapping (GCAM) heatmaps to show our method avoids biased learning and use ensemble softmax averaging for inference. Lastly, top-2 predictions improve parental guidance, achieving 94.88% top-1 and 99.49% top-2 accuracy on the five Dunstan classes, enabling parents to address infants' needs in at most two tries.**

I. Introduction

Newborns cry frequently, yet because they cannot verbally express their needs, parents often struggle to determine the cause, e.g. hunger, sleep, discomfort, or a health issue [1]. The task can feel overwhelming, especially for new parents coping with sleep deprivation and the many changes in daily life following the baby's arrival. Recent progress in audio and speech Machine Learning (ML) makes it feasible to investigate automatic classification of baby cries as a supportive resource. The Dunstan Baby Language (DBL) theory identifies five typical cry types including *eair, eh, heh, neh*, and *owh* corresponding to the basic needs belly pain, need to burp, discomfort, hunger, and tiredness [2]. The DBL corpus has emerged as a standard reference for recognizing multi-class infant cries [3]–[5]. Regarding acoustic representations, Mel-Frequency Cepstral Coefficients (MFCC) features have been extensively and successfully employed for traditional ML models [6]. In parallel, time-frequency inputs have become popular: spectrograms and Mel-spectrograms combined with Convolutional Neural Network (CNN) establish robust baselines on DBL [2], [5], [7]. Under a five-class classification setting, the following studies have been conducted on the DBL dataset. Using a Mel-spectrogram + ResNet-152V2 setup, Junaidi et al. report 77.3% accuracy [8]. In addition to basic CNN, hybrid temporal models have been investigated: CNN–RNN frameworks, such as the one by Nadia et al. [9], demonstrate 94.97% accuracy. Bhagatpatil et al. [10] propose a K-means–based Vector-Quantization (VQ) model using Linear-Frequency Cepstral Coefficients (LFCC) features and achieve 91.02% accuracy. Anjali et al.'s transfer-learning approach, fine-tuning VGG16 with the spectrogram images, achieves the peak accuracy of 92.00% [11]. Abbaskhah et al.'s best approach includes introducing a CNN-based architecture using MFCC and energy features achieving 92.1% accuracy [12]. Qiu et al. report 92.92% accuracy using their hybrid-feature (MMT: MFCC + Mel-spectrogram + Tonnetz) ResLSTM model under 10-fold cross-validation [13].

Although earlier DBL findings show impressive accuracies even in the range of 90%–94%, some of the reported results lack direct comparability or reproducibility due to experimental designs that -even though unintentionally- create length bias (e.g., fixed 1s segments) or data leakage between splits. We explicitly structure our pipeline and metrics to minimize these issues. We propose CryHop, a transfer-learning approach that fine-tunes VGG16 for infant-cry classification and addresses length bias induced by zero-padded inputs. Specifically, we use a variable hop-length approach to transform each recording into a fixed-width time–frequency representation without trimming or zero-padding. This satisfies VGG16's input dimensionality requirements while preserving the recordings' full temporal extent, encouraging reliance on acoustic content rather than clip duration.

II. Dunstan Dataset and the Length Bias Problem

The DBL dataset was originally created to teach humans how to recognize a baby's needs by identifying the Dunstan baby words [14]. It comes in the form of a video, requiring manual extraction to transform it into an audio dataset [12]. Humans are able to generalize well because of their real-world experience, which explains why the dataset is relatively small. The theory behind the baby's words is taught to participants in DBL courses, making learning easier. As a result, only five babies were included in the dataset. By contrast, an Artificial Intelligence (AI) model does not usually generalize as easily, since it lacks human context and relies on mathematical

* Both authors contributed equally to this research.
This work was partially supported by Hector Stiftung.

TABLE I: Composition of the Dunstan dataset

Classes	Number of files	Avg duration (s)
Eair – Belly pain	32	1.567
Eh – Need to burp	46	0.209
Heh – Discomfort	44	0.418
Neh – Hunger	32	0.513
Owh - Tiredness	42	0.979

minimization during training. This can lead to problems when the dataset is prone to biases—often unnoticed by humans. In the case of the DBL dataset, one such bias is length, which can yield deceptively high accuracy if preprocessing is not handled carefully. This paper specifically addresses this length bias.

Based on the aforementioned composition of the dataset, zero-padding the audio files tends to produce high classification accuracy, an effect noted in [9]. However, upon inspection of the Gradient-weighted Class Activation Mapping (GCAM) heatmaps, it becomes evident that the model is relying primarily on the length of the files rather than on the informative spectrogram patterns present in the sounds. Although the file lengths reported in Table I indeed provide a cue regarding the type of cry, if the goal is to develop a robust AI model for real-world deployment, such a reliance on length alone would render the model unsuitable.

One potential solution to address the varying lengths of audio samples, and simultaneously enlarge the dataset, since AI models are typically data-hungry, is to segment the audio samples. However, in this dataset the recordings are already short, and further segmentation would likely confuse the AI model, as similar phonetic elements occur across different cry words (e.g., the phonetic unit "eh" appears in *eh*, *heh*, and *neh* classes). Therefore, we opt to preserve the audio samples in their entirety without segmentation, as the full temporal structure provides important contextual cues that help distinguish one cry type from another, and propose a different solution as described in the following.

III. Methodology

We focus on a preprocessing method that satisfies VGG16's fixed-dimension input and mitigates bias with respect to audio duration, achieved via Mel-spectrograms with a variable hop length.

A. Time-frequency features

As we need a fixed spatial size for the VGG16 model, converting audio to time-frequency images is standard. A Mel-spectrogram, which is a time-frequency representation of the auido files, compresses frequency onto a perceptually driven Mel scale and has proven to be an effective feature in the context of infant cry classification [15]. Given a discrete-time signal $x[n]$ (length N samples), window $w[\cdot]$ of length L, hop length H, target frame count T, and DFT size N_{fft}, we compute the Short Time Fourier Transform (STFT) ($X[k,t]$), the Spectrogram ($S[k,t]$), and the Mel-spectrogram ($S_{\text{Mel}}[m,t]$) as follows [16]–[18]:

$$X[k,t] = \sum_{n=0}^{N-1} x[n]\, w[n-tH]\, e^{-j2\pi \frac{k}{N_{\text{fft}}} n} \tag{1}$$

$$S[k,t] = \left|X[k,t]\right|^2 \tag{2}$$

$$S_{\text{Mel}}[m,t] = \sum_{k=0}^{K} M_{m,k}\, S[k,t], \qquad m = 0, \ldots, n_{\text{Mels}} - 1 \tag{3}$$

In which $M_{m,k}$ is the Mel filterbank weight that maps the k-th linear-frequency FFT bin to the m-th Mel band. Then, we apply log compression on Equation (3) to approximate human loudness perception, compress dynamic range (reducing large variations), and make the features more Gaussian-like for modeling as follows:

$$\text{Mel}[m,t] = \log\big(\epsilon + S_{\text{Mel}}[m,t]\big) \tag{4}$$

The relationship between T and H is described as follows:

$$T = 1 + \left\lfloor \frac{N}{H} \right\rfloor \tag{5}$$

$$H \approx \left\lfloor \frac{N}{T-1} \right\rfloor \tag{6}$$

Therefore, using a variable hop length, we target $T = 224$ frames so that, with $n_{\text{Mels}} = 32$ and $N_{\text{fft}} = 128$, each clip yields a Mel-spectrogram of size $n_{\text{Mels}} \times T = (32 \times 224)$ (Mel bins $\times$ frames). To make these features compatible with VGG16's 224×224 input, we resize only the *vertical* (frequency) axis from 32 to 224. The spectrogram magnitudes are then normalized to $[0,1]$ and subsequently scaled to $[0,255]$ for VGG16 input. For VGG16's three-channel requirement, we replicate the single-channel image to form a $3 \times 224 \times 224$ tensor. The hop length (frame step) is chosen per clip to achieve $T = 224$; smaller hops increase overlap between windows, whereas larger hops reduce it [19].

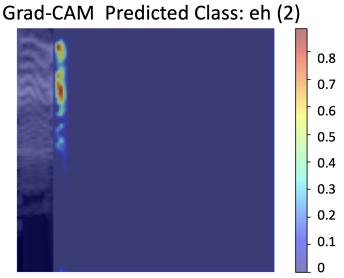


Fig. 1: GCAM heatmap of a Mel-spectrogram using zero-padding

B. Model architecture

In this study, we employ transfer learning by fine-tuning the VGG16 architecture, a very deep CNN originally designed for large-scale image recognition [20]. This approach is inspired by the prior work [11], which applied the same model for infant cry classification. Since CNN models are typically data-hungry, transfer learning constitutes a rational strategy for the relatively small DBL dataset. Accordingly, we added a classification head to the model and trained it for three epochs, followed by fine-tuning the network end-to-end for 30 epochs

TABLE II: Settings and Hyperparameters for the proposed VGG16 fine-tuned model using Mel-spectrograms with variable hop length.

Category	Setting/ Hyperparameters
Audio sampling rate	`SR = 16 kHz`
Base model	Pre-trained **VGG16** (ImageNet), 5 convolutional blocks
Trainable layers	Block 5 (fine-tuning) + Head layers (custom Fully-Connected (FC) layers)
Head layers (custom FC layers)	Pooling on the backbone's output → Dense(512, ReLU) → Dropout(0.5) → Dense(256, ReLU) → Dropout(0.5) → Dense($N_{classes}$, Softmax)
Optimizer	Adam, 2-phase schedule: `lr = 1e-3 (head), lr = 5e-5 (fine-tune)`
Loss	Categorical cross-entropy

after unfreezing the fifth block of the backbone. The architecture and hyperparameters of this model are summarized in Table II.

IV. Evaluation

A. Experimental setup

We perform a 5×5 nested cross-validation for robust evaluation. In each outer fold, we use 80% of the data for model development and 20% for final testing. Within the training portion, a 5-fold inner cross-validation selects the best-performing models (highest accuracy and lowest validation loss). Then, we combine these inner models using an ensemble softmax-averaging (voting) approach to predict the outer test set. Finally, the reported performance corresponds to the average accuracy across the five outer folds. Since end-users (parents/clinicians) frequently operate on several reasonable hypotheses, we additionally provide top-2 accuracy in addition to top-1, representing a more practical decision-support context for mobile applications. We run these experiments on a 64-bit system with an Intel Core i9-13900 CPU, 32 GB RAM, and an NVIDIA GeForce RTX 4060 (8 GB) GPU using PyTorch with CUDA acceleration.

B. Results

The experimental results for the 5 outer folds are summarized in Table III. Our pipeline achieves **94.88%** top-1 accuracy, **95.00%** F1-score, and **99.49%** top-2 accuracy, representing an improvement of around 2% over the existing literature as shown in Figure 2. The confusion matrix in Figure 4 indicates consistently high class-wise performance, with only a small misclassification rate for the *owh* class. Further details on training/validation accuracy and loss, averaged over the 25 folds are provided in Figure 3.

C. GCAM visualisations

To ensure that the model focuses on meaningful spectral features rather than artifacts caused by signal length, we generate GCAM heatmaps and average them across the five available classes, as shown in Figure 5.

TABLE III: Ensemble performance on outer test folds

Outer Fold	Top-1 (ACC)	Top-1 (F1-score)	Top-2 (ACC)
Outer 1	0.9750	0.9787	1.0000
Outer 2	1.0000	1.0000	1.0000
Outer 3	0.8974	0.8992	1.0000
Outer 4	0.9231	0.9271	0.9744
Outer 5	0.9487	0.9450	1.0000
Mean ± SD	0.9488 ± 0.0363	0.9500 ± 0.0368	0.9949 ± 0.0102

V. Discussion

These GCAM visualizations highlight the spectral regions attended to by the model during inference.

In Figure 5 (*eair*), the model attends to the number of harmonics produced by the baby; compared to the other Mel-spectrograms, the harmonics appear almost doubled, which to a human listener corresponds to the baby experiencing intense pain. In Figure 5 (*eh*), the model focuses primarily on the low-frequency region and the horizontal structure of the lines, reflecting the fact that this short word exhibits very little temporal variation. In Figure 5 (*heh*), although the model does not focus directly on it, it highlights the low-frequency components produced by the exhaled sound of the initial "h" in *heh*, visible in the upper-left region of the spectrogram. In Figure 5 (*neh*), the "n" part of the word *neh* which signifies hunger, appears as the initial vertical line across all frequencies. Although the model does not focus exclusively on this feature, it attends to the surrounding region near this line and uses it to support its decision. Finally, In Figure 5 (*owh*), the model pays attention to the gradual frequency changes over time.

As a result, the GCAM visualizations reveal that the model focuses on different and significant spectral regions for each type of cry. This suggests that its decisions are based on genuine acoustic cues rather than on a systematic bias toward a single feature, such as the length of the signal.

We find variable hop length method to be a suitable method for solving the length bias problem. Looping, zero-padding, noise-padding, stretching, speeding up, segmentation and realistic cries were part of our failed attempts to correct this issue without success. Looping, zero-padding and noise-padding revealed heavy length bias through the GCAM heat maps. For instance, the study in [9] was prone to length bias; therefore, we exclude it from the benchmark comparisons reported in Figure 2. We observed approximately a 10% drop in accuracy when reimplementing their pipeline using the trimming/padding approach compared to our variable-hop method.

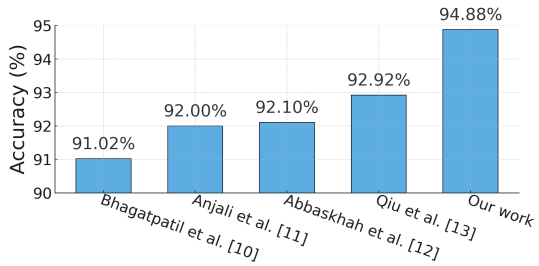
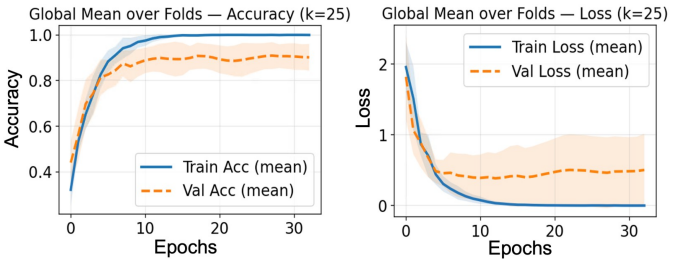


Fig. 2: Quantitative comparison of prior studies versus ours



(a) Accuracy (b) Loss

Fig. 3: Average training curves across inner folds for (a) accuracy and (b) loss using VGG16 fine-tuning with variable-hop Mel-spectrogram features over 30 epochs

Speeding audio files up with pitch compensation caused too much data loss because 1.5 second files were sped up to match the shortest files post noise removal which happens to be approximately 0.1 seconds long. Stretching audio files with pitch compensation caused time dilation artifacts to appear in the Mel-spectrograms and the models could also identify those and be biased because of the audio length, indirectly in this case. The time dilation was heavy since 0.1 seconds files were stretched to last 1.5 seconds. Segmentation led to poor performances given that the cues to identify the Dunstan baby words are not present throughout the entire audio recording, for example *eh, heh*, and *neh* all have "eh" as a basis and some classes add prefixes such as "n" and "h" which happen to be the only cue to differentiate the 3 classes amongst themselves. Finally, an attempt to create realistic cries from the Dunstan dataset also resulted in length bias in the heat maps. The method was to turn one original Dunstan audio file and loop it with slight randomly chosen time dilation with pauses of various length in between cries to mimic repeating cries from a baby that would be heard in a real recording.

Our pipeline, variable hop Mel-spectrograms to remove length bias, nested cross-validation for model selection, and transfer learning with a fine-tuned VGG16, outperforms prior work relying on hand-crafted features and non-pretrained models. Notably, although inner-fold validation accuracies were comparable to (or slightly lower than) outer-fold results, ensembling the best inner models (softmax averaging) improved and stabilized test accuracy. This gain stems from variance reduction across folds and the suitability of pretrained features

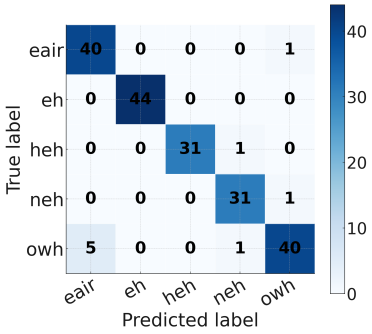


Fig. 4: Summed confusion matrix of the 5-class cry classifier

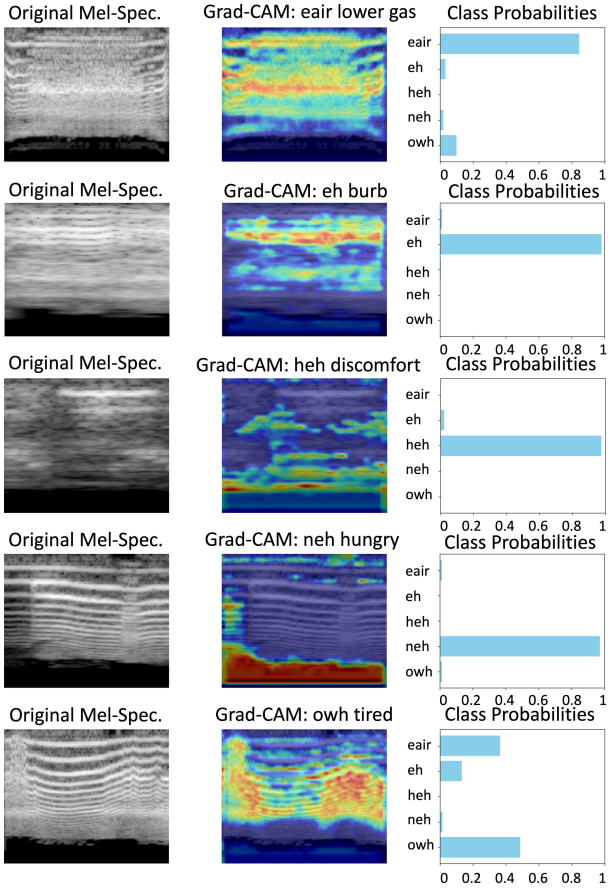


Fig. 5: GCAM heatmaps for the five Dunstan words—eair, eh, heh, neh, and owh—ordered from top to bottom. The frequency axes are inverted, with low frequencies displayed at the top and high frequencies at the bottom. The probability of each class is also shown on the right side of the images.

for the small DBL dataset, enabling better generalization with limited data.

VI. Conclusion

This study introduces a transfer learning–based VGG16 pipeline for infant cry classification on the DBL dataset, incorporating fine-tuning to enhance performance while mitigating the length-bias observed in prior work. By adapting the variable hop length, the length bias is eliminated while preserving content, which we confirmed by visualizing the attention patterns across Mel-spectrogram regions of each class using GCAM heatmaps. The proposed approach achieves an average top-1 accuracy of 94.88% ± 3.63%, F1-score of 95.0% ± 3.68%, and top-2 accuracy of 99.49% ± 1.02% for five different infant needs. The introduction of top-2 accuracy has an important practical consequence; even in the rare cases of misclassification (≈5% for top-1), the parents can quickly use the second most-likely prediction (top-2) to address their infants need. These results demonstrate the robustness and generalization capability of the proposed model, making it a strong candidate for integration into mobile applications aimed at real-world infant monitoring and early distress detection.

REFERENCES

[1] M. Mekhfioui *et al.* Development of a baby cry identification system using a raspberry pi-based embedded system and machine learning. *Technologies*, 13(4):130, 2025.

[2] E. Franti *et al.* Testing the universal baby language hypothesis-automatic infant speech recognition with cnns. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pp. 1–4. IEEE, 2018.

[3] I.-A. Bănică *et al.* Automatic methods for infant cry classification. In *2016 International Conference on Communications (COMM)*, pp. 51–54, Bucharest, Romania, June 2016. IEEE.

[4] R. I. Tuduce *et al.* Why is my baby crying? an in-depth analysis of paralinguistic features and classical machine learning algorithms for baby cry classification. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pp. 1–4, 2018.

[5] X. Qiao *et al.* Infant cry classification using an efficient graph structure and attention-based model. *Kuwait Journal of Science*, 51:100221, 03 2024.

[6] W. S. Limantoro *et al.* Application development for recognizing type of infant's cry sound. In *2016 International Conference on Information Communication Technology and Systems (ICTS)*, pp. 157–161, 2016.

[7] C. A. Bratan *et al.* Dunstan baby language classification with cnn. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 167–171, 2021.

[8] R. Junaidi *et al.* Baby cry sound detection: A comparison of mel spectrogram image on convolutional neural network models. *Journal of Electronics Electromedical Engineering and Medical Informatics*, 6:355–369, 10 2024.

[9] T. Maghfira *et al.* Infant cry classification using cnn – rnn. *Journal of Physics: Conference Series*, 1528:012019, 04 2020.

[10] M. V. Bhagatpatil and V. Sardar. An automatic infant's cry detection using linear frequency cepstrum coefficients (lfcc). *Int. J. Sci. Eng. Res*, 5(12):1379–1383, 2014.

[11] G. Anjali *et al.* Infant cry classification using transfer learning. In *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*, pp. 1–7. IEEE, 2022.

[12] A. Abbaskhah *et al.* Infant cry classification by mfcc feature extraction with mlp and cnn structures. *Biomedical Signal Processing and Control*, 86:105261, 2023.

[13] Y. Qiu *et al.* Classification of infant cry based on hybrid audio features and reslstm. *Journal of Voice*, 2024.

[14] P. Dunstan. *Calm the Crying: Using the Dunstan Baby Language*. Penguin Books Ltd., London, 2012.

[15] A. Mukhamediya *et al.* On the effect of log-mel spectrogram parameter tuning for deep learning-based speech emotion recognition. *IEEE Access*, 11:61950–61957, 2023.

[16] J. B. Allen and L. R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 2005.

[17] J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE transactions on acoustics, speech, and signal processing*, 25(3):235–238, 2003.

[18] P. A. Babu *et al.* Speech emotion recognition system with librosa. In *2021 10th IEEE international conference on communication systems and network technologies (CSNT)*, pp. 421–424. IEEE, 2021.

[19] A. Zhao *et al.* Optimizing short-time fourier transform parameters via gradient descent. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2021.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.