Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Speech Signal Processing and Speech Enhancement
Sommersemester 2025

Infant cry classification: a Convolutional Neural Network Approach

Gloria Galasso
3731017

Studiengang: M.Sc. Computational Linguistics

Prüfer: Prof. Dr. Wolfgang Wokurek

Datum: 27.10.2025

1 Introduction

Crying is the first form of communication for newborns. It is an innate reflex used to express their need and their only way to communicate with the world. As babies do not have a formal language, they use their cries to express their physical and emotional needs. Infants cry for different reasons: hunger, pain, need to burp, etc. In recent years, advances in technology have shown that professionals and non-professionals can benefit from automated classification. Cries serve as an audio signal that conveys the baby's state. Hence, researchers have long explored how the fundamental frequencies and acoustic features of the cry correlate with various factors of the baby's state. However, caregivers and parents who are not experts in acoustic research find it challenging to understand their infants' needs and often misinterpret them [1]. Therefore, automatic classification is beneficial to both ordinary individuals and professionals because it provides guidance on better baby care, reduces caregivers' stress, and can inform healthcare and infant care services. Moreover, cry interpretation is a valuable tool for early detection of diseases such as hyperacusis and deafness [2].

Recent advancements in Deep Learning (DL), particularly Convolutional Neural Networks (CNNs), have enabled significant improvements in speech recognition and infant cry analysis. Namely, CNNs are multi-layered neural networks that have been widely applied to audio processing tasks due to their reliable performance. Specifically, Mel-spectrograms and Mel-frequency cepstral coefficients (MFCC) have been employed as input for cry recognition [1]. Consequently, this project aims to apply DL techniques to classify infant cries based on their needs: belly pain, burping, discomfort, hungry, and tired. Nonetheless, the challenge lies in transforming the raw audio into a representation that the CNNs can analyze [3].

The central question of this project is: Given a small, imbalanced dataset, do spectrogram- or MFCC-based feature representations enable effective classification of infant cries when fed into Convolutional Neural Networks?

To respond to the aforementioned question, this study compared two approaches for transforming the audio signal: creating a visual representation (a spectrogram) and extracting a feature tensor through Mel-frequency cepstral coefficients. In this regard, a central theme was pre-processing the imbalanced and small `donatea-corpus` dataset [1] by applying data augmentation (AD) techniques.

The following paragraphs will address prior work on automatic infant cry classification. A subsequent section explains the dataset and its augmentation, which was generated by applying signal-processing techniques. Afterwards, this report will clarify the pre-processing of spectrograms, distinguishing between narrow-band and wide-band spectrograms. It will also elaborate on the transformation of raw audio into a tensor using MFCCs. Once the pre-processing phase has been discussed, there will be an explanation of how CNNs were built, showing the functions of their components, how they extract features and patterns, and how they perform classification. Thereafter, a quantitative and qualitative analysis will be conducted regarding the CNN model results derived from:

1. Narrow-band spectrograms from the original raw dataset,
2. Narrow-band spectrograms from the augmented dataset,
3. Wide-band spectrograms from the augmented dataset,
4. Feature tensors based on MFCCs extracted from the original dataset.

Finally, the conclusion will provide a comprehensive overview of the study's main findings and implications.

2 Related Work

Automatic classification of infant cries in machine learning (ML) and deep learning (DL) is an active area of research, and the literature offers established approaches [2] [4].

[1] The dataset is available at: `https://github.com/gveres/donateacry-corpus/tree/master/donateacry_corpus_cleaned_and_updated_data`

On the one hand, transforming the audio signal into a two-dimensional visual representation is one of the most common techniques in DL. This is achieved by converting the raw data into a Mel-spectrogram, which visually represents sound as humans perceive it. Consequently, the image becomes the input to a CNN model that has the goal of solving an image classification problem [2].

On the other hand, Ozcan & Gungor (2025) [4] support the claim that a popular alternative for extracting acoustic features from audio signals is the Mel-Frequency Cepstral Coefficients. Furthermore, Liang et al. (2022) [5] state that DL algorithms, such as CNNs, take as input features extracted from MFCCs of infant cries. For instance, Ozcan & Gungor (2025) [4] achieved 86% of accuracy through this approach.

However, a recurring challenge is the limited data availability, as many studies use their proprietary data [5]. One of the most widely utilized publicly available datasets is the `donateacry-corpus` dataset. Nonetheless, this dataset is characterized by a high imbalance between the classes (`hungry, burping`, etc.) and is small in size [2]. Ozcan & Gungor (2025) [4] confirm that the models' unsatisfactory performance is mainly due to their poor performance on minority classes. To address this issue, data augmentation techniques have been applied to improve model generalization and reduce bias towards the majority classes.

Acting on this knowledge, this project conducted a comparative analysis of spectrograms and MFCC features as inputs to the CNN model. Specifically, it also investigated the impact of spectrogram parameters (narrow-band vs wide-band). Finally, the role of data augmentation contributed to the understanding of best practices.

3 Material and Methods

3.1 Data and Data Augmentation

As shown in Table 1, the donateacry-corpus dataset contains a total of 457 .wav files divided into five different classes: belly pain, burping, discomfort, hungry, and tired. The dataset was collected through the Donate-a-cry campaign, in which parents used a mobile application to record their infant's cry and provide information on gender, age, and the presumed reason for crying. The files were converted to .wav format with a uniform bit rate of 128 kbps and a sampling rate of 8kHz. Furthermore, non-cries and other data, such as white noise, baby chat, etc., were manually removed [4]. In Table 1, under the column "Original" dataset, it is evident that the number of files per cry category is uneven. The class with the most files is `hungry`, with 382, and the class with the fewest, only eight audio files, is `burping`. Therefore, this project was based on a small and imbalanced dataset. To further analyze the audio classification, the data augmentation technique was applied to equalize the number of samples across all categories. As the column "Augmented" in Table 1 displays, the categories of `belly pain, burping, discomfort, and tired` were augmented to reach a total of 382, like in the `hungry` category. This approach was adopted because data augmentation increases dataset diversity and prevents models from becoming biased toward `hungry` [4]. In particular, three signal processing techniques were applied to enhance the size of audio files [2]:

1. Noise Injection: the audio waveform was added with low-amplitude Gaussian noise. In this way, the model was trained to account for imperfect and real-world sounds, including background static or electronic hiss from a recording device.

2. Time Stretching: `librosa.effects.time_stretch` was employed to change the duration and therefore the speed of the audio signals without changing the pitch. The audio signal was made 1.25 times faster by default.

3. Band-Pass Filtering: the audio files were modified by keeping frequencies between 300 Hz and 3000 Hz and filtering out frequencies outside those ranges.

Moreover, to increase dataset variability, original .wav files from the four categories were randomly selected and augmented using one of the techniques cited above. This resulted in a varied dataset with a broader range of acoustic conditions.

Table 1: Overview of the donateacry-corpus dataset before and after data augmentation

Class Label	**Original**	**Augmented**
Belly pain	16	382
Burping	8	382
Discomfort	27	382
Hungry	382	–
Tired	24	382
Total	**457**	**1910**

3.2 Pre-processing: spectrograms

Following the example of Le et al. (2019) [3], in the pre-processing stage, .wav audio files were converted to Mel spectrograms. The Mel spectrogram is a two-dimensional image where the x-axis represents time and the y-axis represents frequency. The spectrogram image is modeled as a nonlinear representation of human sound perception, mapping the frequency axis to the Mel scale. This scale shows a linear distribution below 1000 Hz and an increasing logarithmic scale above 1000 Hz. For these reasons, Melspectrogram is used for audio classification tasks, as it is susceptible to low-frequency sounds.

In a first stage, narrow-band spectrograms were generated with a long analysis window of 2048 samples (`n_fft`) and a hop length of 512 samples.

In a second stage, a wide-band spectrogram was employed with a shorter analysis window of `(n_fft=512)` and a smaller hop length of `(hop_length=128)`.

Specifically, the narrow-band spectrogram (Figure 1) provides high frequency resolution and low temporal resolution. It means it is suitable for capturing the pitch and harmonic details, but not as good at capturing rapid changes. Whereas, the wide-band spectrogram has better temporal resolution for capturing short transients, onset timing, and rapid intensity changes, but poorer frequency resolution. As a consequence, using both types of spectrogram for infant cry analysis is relevant for collecting complementary acoustic information [1].

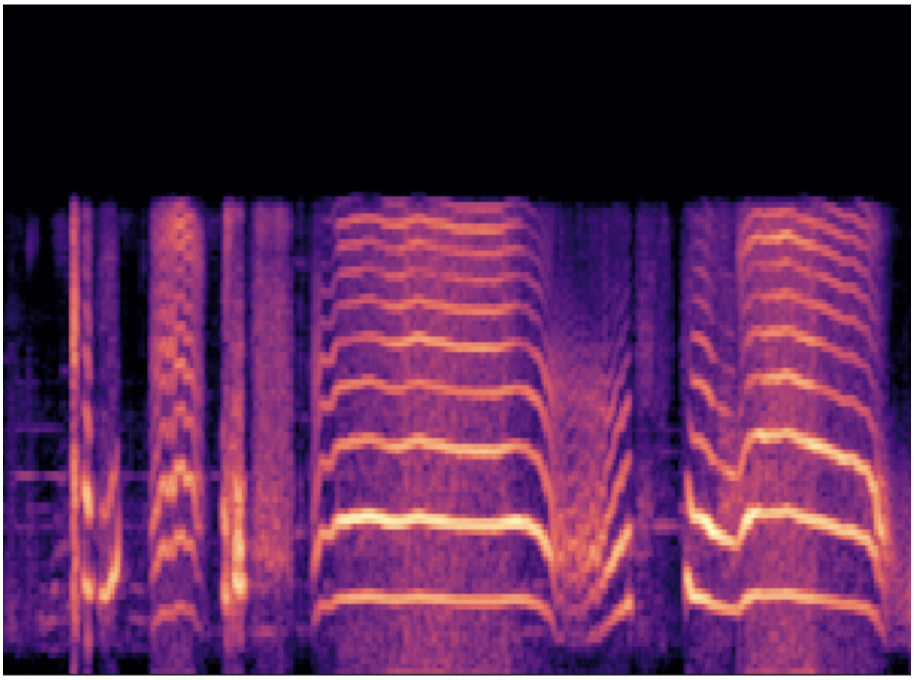

Figure 1: A narrow-band spectrogram from the category of belly pain, showing clear harmonic structures (the horizontal lines).

3.3 Pre-processing: Mel-Frequency Cepstral Coefficients

An alternative approach to create input for the model is not to create spectrograms from audio files, but to extract Mel-Frequency Cepstral Coefficients. This is a further method of the visual analysis. MFCCs capture spectral characteristics of an audio signal in a similar way to how humans would perceive them and provide a robust representation of the timbre [4].

All the audio files were set to 5 seconds. Then, a sequence of 40 MFCCs was extracted for each short time frame. Next, .wav files were converted into a matrix that represents how the timbre changes over time.

In the end, the output of the pre-processing is a standardized (same length and the same number of extracted features) numerical tensor (multi-dimensional array) for each cry.

3.4 Model implementation

The CNN was employed to automatically recognize patterns in infant cries and classify them based on the infants' needs. As input, the model used spectrogram images for a part of the experiment and, for another part, tensors derived from MFCCs [5].

The CNN architecture consisted of two components: a feature extraction block and a classification block. In the case of spectrograms, a two-dimensional Convolutional Layer (Conv2D) was used to learn spatial patterns that represented acoustic events , while for the MFCC tensors, a one-dimensional Convolutional Layer (Conv1D) network learned temporal patterns.

In the case of images, in the feature extraction block, the first convolutional layer detects recurrent patterns in the input at the stage of low-level features. Basic elements such as horizontal lines (stable pitch) or vertical lines that represent short transient bursts or impulsive sounds. Instead, the second layer receives a map of the simple features and detects and combines them into more complex patterns. Then, the Max Pooling layer creates a more compact representation of the features by retaining the most relevant information from each local region. In the next stage, the summary of the key patterns is passed to the classification block. The Dense layer weights evidence from the feature list. For instance, it learns that high-frequency harmonic stacks are an indicator of belly pain. The Softmax function turns scores into probabilities for each cry type. It outputs five numbers that sum to 1, which are the probabilities of belonging to each class.

In a parallel experiment, the multidimensional MFCC tensors were passed to a 1D Convolutional Neural Network (CNN). As the signal passes through the deep layers, the network recognizes complex patterns. A `GlobalAveragePooling` operation represents all the characteristics into a single vector. This vector is passed to the Dense layers and a Softmax function, which produces probabilities for each of the cry classes.

4 Results and Discussion

4.1 Narrow-band spectrograms from the original raw dataset

In the first experiment, the feature-extraction technique was applied to the original data to produce narrow-band spectrograms. After the initial pre-processing phase, the images were split into a training set (80%) and a test set (20%). Furthermore, the model was trained for 50 epochs, and training and validation performance were monitored using accuracy curves. Figure 2 shows that the validation curve stops improving and diverges from the training curve, which reaches 95% in accuracy. The gap between the two lines suggests overfitting, which means that the model learned the characteristics of the training set but failed to generalize. Therefore, the model could not recognize or classify the data characteristics in the test set.

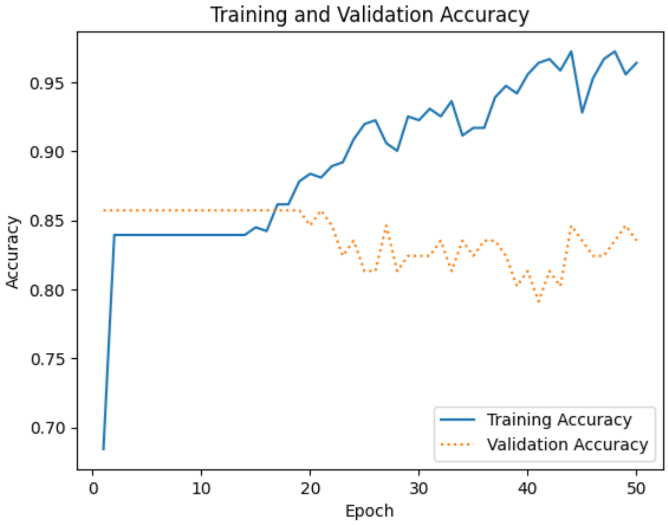

Figure 2: Accuracy in narrow-band spectrogram from original dataset.

In particular, the majority of audio files from underrepresented categories were misclassified as `hungry`, which is the class with the most significant number of audio files. Thus, this misclassification indicates a bias towards the class `hungry`. Additionally, the confusion matrix in Figure 3 shows 76 correct predictions for the `hungry` category. It confirms that the model relied on the dominant characteristics of the majority class and struggled to generalize to minority classes present in the dataset.

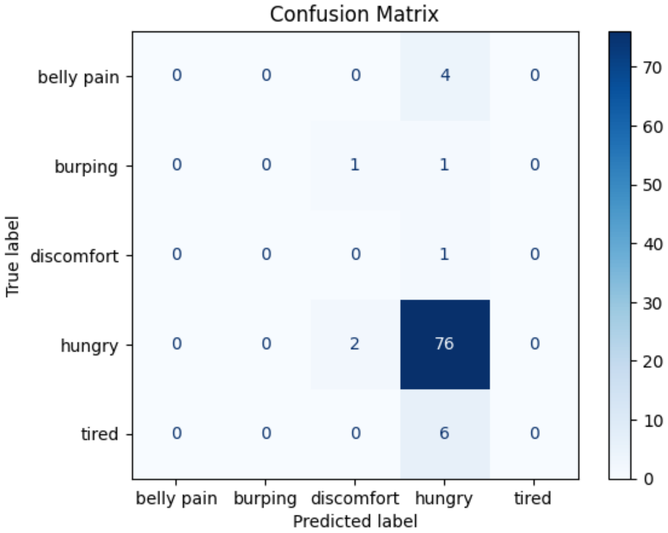

Figure 3: Confusion Matrix of Narrow-band spectrograms from the original raw dataset

4.2 Narrow-band spectrograms from augmented data

As in the previous experiment, the model suffered from class imbalance and overfitting, with a clear bias towards `hungry`. In this second stage, the dataset was enlarged using AD techniques (noise injection, time stretching, and band-pass filtering). In this way, the model achieved an accuracy of around 98% on the validation set. The training and validation accuracy curves converged, indicating consistent performance, as Figure 4 shows.

The confusion matrix in Figure 5 illustrates almost perfect classification of the audio files into their categories. Despite the impressive metrics, the interpretation of these outcomes is necessary. The augmented versions are a slight variant of the data. Thus, the model was able to identify the unique identity of the original audio clip rather than generalizing to completely unseen examples. Therefore, the synthetic audio files made the classification task artificially easier.

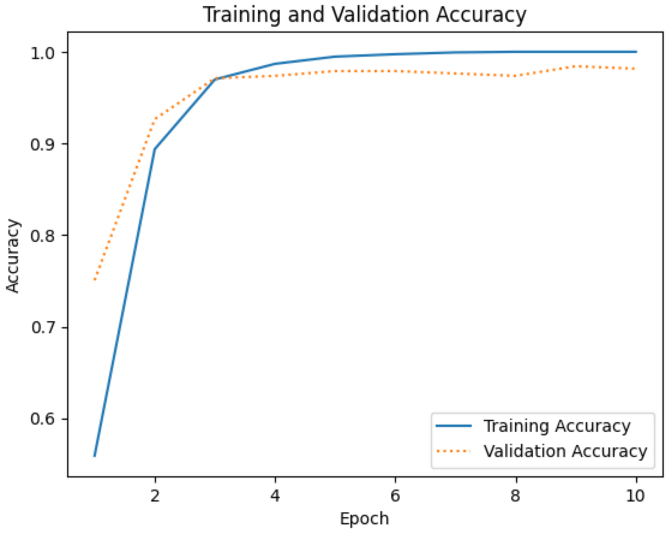

Figure 4: Accuracy in narrow-band spectrogram from augmented dataset.

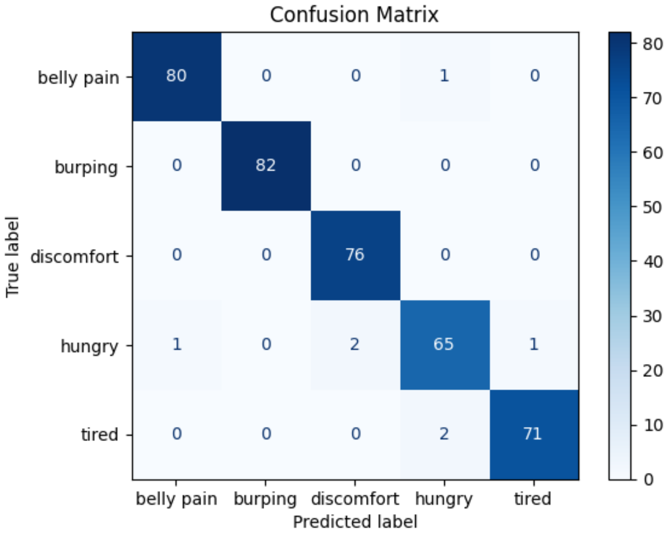

Figure 5: Confusion Matrix of Narrow-band spectrograms from the augmented dataset

Even if the above lines mention the limitations of the results, this project helps illustrate that a CNN can recognize meaningful features in infant cries when the dataset is balanced. The real test would be trying to generalize to a totally different and independent dataset with cries recorded in different environments and different devices/microphones.

4.3 Wide-band spectrogram from augmented data

The approach using wide-band spectrograms has a similar behavior to that described in the previous paragraph. Figure 6 shows high validation accuracy of around 98% and an almost-perfect confusion matrix (Figure 7). Apparently, the model avoided overfitting. However, as in the experiment with narrow-band spectrograms from augmented data, the high performance must be evaluated with the understanding that the data used is synthetic. This was a simplified task, as the model had similar features to recognize in both the training and test sets, since the synthetic audio could have its parent in either set. Again, CNN demonstrated how to detect infant cries and their acoustic identity as known samples.

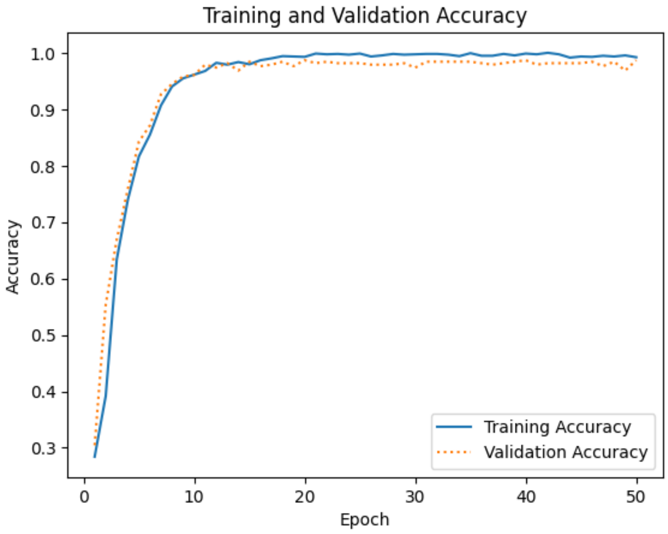

Figure 6: Accuracy in Wide-band spectrograms from the augmented dataset

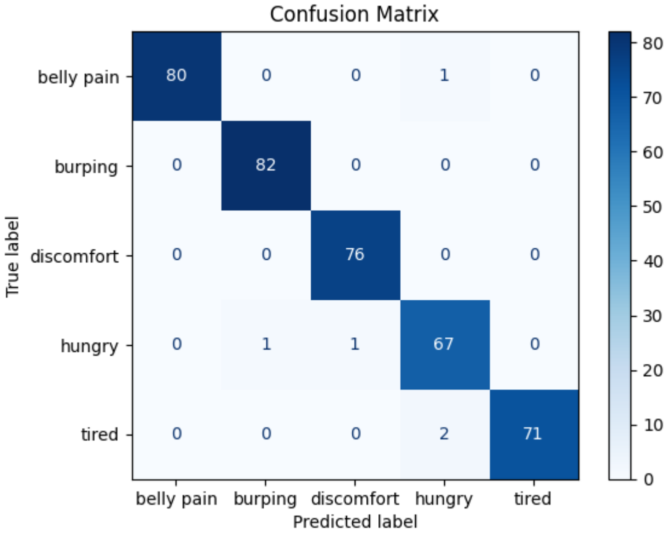

Figure 7: Confusion Matrix of Wide-band spectrograms from the augmented dataset

4.4 Feature tensors based on MFCCs extracted from the original dataset

In this final experiment with MFCC tensor data, the data were divided into a training and a test set at 80% and 20%, respectively. The test accuracy reached 62%, which indicates poor generalization. In fact, many samples were misclassified, showing that the model was unstable.

This issue is due to a mismatch between the model's complexity and the limited diversity and size of the dataset. Therefore, the model memorized non-essential artifacts, such as background noise and microphone differences, rather than learning actual acoustic patterns. To sum up, the model learned the noise instead of the essential signals.

To conclude, this outcome contrasts with the strong performance of 2D CNN models trained on spectrograms with augmentation. The 1D CNN could not capture richer structural information (harmonics, frequency contours, intensity changes) from MFCC sequences alone.

5 Conclusion

This infant cry classification project aimed to illustrate different applications of Convolutional Neural Networks through the use of Mel spectrograms and Mel-Frequency Cepstral Coefficients. The outcomes show that with data augmentation and spectrograms, high accuracy was achieved. However, the good performance was reached because the augmented audio files differed only slightly from one another, making the classification task easier for the model. On the other hand, the MFFC-based approach on the original data did not generalize well due to the model's complexity and the dataset's small size and imbalance. In conclusion, it should be acknowledged that the `donateacry-corpus` dataset was created by parents who classified their babies' cries. Hence, they are not experts in recognizing their cries and needs. For this reason, in the future, it would be adequate to apply the same approaches on a larger, more reliable dataset.

References

[1] Yuta Shinya, Taiji Ueno, Masahiko Kawai, Fusako Niwa, Seiichi Tomotaki, and Masako Myowa-Yamakoshi. Listening deeper: Neural networks unravel acoustic features in preterm infant crying. *Research Square*, Aug 2024. Preprint.

[2] M. Hammoud, M. N. Getahun, A. Baldycheva, and A. Somov. Machine learning-based infant crying interpretation. *Frontiers in Artificial Intelligence*, 7:1337356, Feb 2024.

[3] Lillian Le, Abu Kabir, Chunyan Ji, Sunitha Basodi, and Yi Pan. Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images. In *2019 IEEE 16th International Conference on Mobile Ad hoc and Smart Systems Workshops (MASSW)*, pages 106–110, 2019.

[4] Tayyip Ozcan and Hafize Gungor. Baby cry classification using structure-tuned artificial neural networks with data augmentation and mfcc features. *Applied Sciences*, 15(5):2648, 2025.

[5] Yun-Chia Liang, Iven Wijaya, Ming-Tao Yang, Josue Juarez, and Hou-Tai Chang. Deep learning for infant cry recognition. *International Journal of Environmental Research and Public Health*, 19:6311, May 2022.