Article

A study of deep learning models for audio classification of infant crying in a baby monitoring system

Denisa Maria Herlea 1, Bogdan Iancu 1 and Eugen-Richard Ardelean 1,\*

|  |
| --- |
| Academic Editor: Firstname Lastname  Received: date  Revised: date  Accepted: date  Published: date  **Citation:** To be added by editorial staff during production.  **Copyright:** © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Computer Science Department, Technical University of Cluj-Napoca; [denisa.herlea@student.utcluj.ro](mailto:denisa.herlea@student.utcluj.ro); [bogdan.iancu@cs.utcluj.ro](mailto:bogdan.iancu@cs.utcluj.ro); ardeleaneugenrichard@gmail.com;

**\*** Correspondence: ardeleaneugenrichard@gmail.com;

**Abstract:** This study investigates the ability of well-known deep learning models, such as ResNet and EfficientNet, for audio-based infant cry classification. By comparing the performance of different machine learning algorithms, this study seeks to determine the most effective approach for the detection of infant crying, enhancing the functionality of baby monitoring systems and contributing to a more advanced understanding of audio-based deep learning applications. Understanding and accurately detecting a baby's cries is crucial for ensuring their safety and well-being, a concern shared by new and expecting parents worldwide. Despite advancements in child health, as noted by UNICEF's 2022 report of the lowest ever recorded child mortality rate, there is still room for technological improvement. This paper presents a comprehensive evaluation of deep learning models for infant cry detection, analyzing the performance of various architectures on spectrogram and MFCC feature representations. A key focus is the comparison between pretrained and non-pretrained models, assessing their ability to generalize across diverse audio environments. Through extensive experimentation, ResNet50 and DenseNet trained on spectrograms emerged as the most effective architectures, significantly outperforming other models in classification accuracy. Additionally, the study investigates the impact of feature extraction techniques, dataset augmentation, and model fine-tuning, providing deeper insights into the role of representation learning in audio classification. The findings contribute to the growing field of audio-based deep learning applications, offering a detailed comparative study of model architectures, feature representations, and training strategies for infant cry detection.

**Keywords:** deep learning; convolutional neural network; classification; infant crying; resnet; efficientnet;

1. Introduction

Ensuring a baby’s well-being is a top priority for new and expecting parents, but the need for constant monitoring can be physically and emotionally exhausting. Monitoring children's safety and health both in the first months of life and beyond is not only a crucial necessity, but also a moral obligation that lays the foundations of the broad term 'parenting'. At the beginning of their lives, babies are sensitive and vulnerable to various external factors that can jeopardize their life and health, which is why constant monitoring of their developmental parameters. According to UNICEF, the year 2022 saw the fewest number of deaths among children under five, but nevertheless, the reported 4.9 million number can and should be improved every year.

With advancements in technology, AI-powered baby monitoring systems have become an essential tool in assisting parents with infant care. Mobile applications are now an integral part of daily life, providing solutions for various aspects of parenting, including baby monitoring. Also, the use of IoT (Internet of Things) devices has gained momentum in various fields, with various monitoring devices such as smart cameras, remote communication devices, devices for monitoring breathing or heart rate and many such examples now on the market.

Integrating an AI-based model for recognizing infant cries can alleviate parental stress by providing timely alerts even when they are not in close proximity to their baby. Such a model, integrated with hardware components would be capable of alerting parents in very close to real time to any discomfort their child is experiencing, thus being able to reach their child and solve problems as quickly as possible.

2. Materials and Methods

Data preprocessing

Data preprocessing methods make up one of the most important steps in training an artificial intelligence model because the quality of the preprocessing determines the quality of the predictions made by it.

There are many methods by which data can be modified to match the architecture of a model, among which the most common ones are the subject of this paper:

* *Cleaning*: by replacing missing or null values with other suggestive values close to the real data, such as mean, median or other derivatives, or removing duplicates and irrelevant data for the training process.
* *Normalize* the data by fitting them into a more precise and easier to analyze range.
* *Encode* data by transforming labeled categories into numbers, vectors or other data structures.
* *Data augmentation*, by modifying the dataset to force the model to generalize better on unknown data and to broaden the range of cases recognized by the model.
* *Data resizing*, by adding new data to match a desired length or deleting existing data to bring all the data in the dataset to the same size.

Data representations

For effective classification of infant cries, selecting the appropriate feature representation is crucial, as different methods capture varying aspects of audio signals. This task becomes more complicated as various other sounds occur in the background or the environment is acoustically changing [1]. This study evaluates two widely used representations: spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), analyzing their impact on model performance across different deep learning architectures. In Figure 1, we can see the different options for representing five seconds of infant crying, each of which is important and different, which is why choosing the right option is challenging.

One of the most commonly used representations is the spectrogram, a visual representation of frequency over time, obtained by applying the Short-Time Fourier Transform (STFT) to the audio signal. Spectrograms retain a high-resolution depiction of time-frequency relationships, making them particularly effective for distinguishing crying sounds from background noise such as household chatter, television sounds, or environmental disturbances. However, one limitation of spectrograms is that overlapping noises of similar frequencies may not always be clearly distinguishable. To address this, the Mel-filter log representation provides an alternative that maps frequencies onto the Mel scale, better aligning with human auditory perception. This transformation can help distinguish between similar-sounding noise sources, such as adult speech mixed with infant cries. A widely used method in speech and audio recognition is the Mel-Frequency Cepstral Coefficients (MFCCs), which are derived from the spectrogram but emphasize the spectral envelope. While MFCCs are beneficial for speech-based classification tasks due to their ability to capture vocal tract characteristics, they may not always be optimal for non-speech sounds like infant cries. In particular, MFCCs tend to be more robust against constant background noise (e.g., air conditioning hums, white noise) but may struggle with transient noises or overlapping acoustic events.

In practice, it is more common to use the discrete Fourier transform (DFT) [2]. The DFT was chosen for this study as it provides a full representation of spectral characteristics, ensuring that both low- and high-frequency components of infant cries are captured. Given that infant cries exhibit relatively stable frequency characteristics over short time windows, STFT-based spectrograms and MFCCs derived using the DFT provide an effective balance between feature richness and computational efficiency. The calculation of the DFT is performed according to the formula:

The spectrogram and Short-Time Fourier Transform (STFT) are two very closely related concepts, the difference in their applicability being in the way the frequency of the data changes (stationary or non-stationary). Spectrogram is more suitable for stationary data, it displays only the magnitude of the power spectrum , whereas STFT displays both magnitude and phase. The calculation formula for the Short-Term Fourier Transform is as follows:

Waveforms are another way of representing audio data for the purpose of preprocessing and extracting relevant features for machine learning models. These, however, are not usually used in this form for training models due to their complexity. This graphical representation depends on the signal amplitude, as described in [3], which can change if the pressure around the microphone changes.

The waveform is generally used for extracting other key features needed by the machine learning domain. For example, one can infer from it the spectrogram mentioned above, but also other data such as Mel Frequency cepstral Coefficients (MFCC). The differences between the two, displayed graphically in *Figure 1*, are in the time-frequency representation, as the spectrogram contains details of the entire audio spectrum, whereas the MFCC uses a logarithmic scale based on human perception of sound frequencies. Therefore, their dimensions differ, with MFCC being more reduced by synthesizing spectral information into cepstral coefficients.

A screenshot of a computer screen

Description automatically generated

*Figure 1 – A sample of a crying sample with its signal, its MFCC and its Fourier spectrogram.*

Data augmentation

Dataset augmentations play a role in the robustness and generalization abilities of deep learning models, especially when a high number of samples is lacking. For audio data, augmentations such as time stretching, pitch shifting, addition of noise, and changes in volume introduce variations that can occur when testing such a model in real environments. This aids in not overfitting to the training data and therefore assures the model's performance over unseen data segments. Augmentation can also help to balance datasets, especially when classes are imbalanced, by artificially increasing the diversity and quantity of training samples. These techniques thereby help improve pattern recognition and feature extraction within a model that has invariance to some distortions, which has shown better performance in tasks such as speech recognition, music genre classification, and environmental sound detection.

* Pitch shifting seeks to alter the pitch of the audio without changing its duration. This is used in applications such as music genre classification or speech recognition, in which many pitches would help refer to various speakers or musical instruments. By training on pitch-shifted audio, a model has a greater likelihood of becoming invariant to these variations.
* Adding noise involves overlaying background noise onto the original audio, usually done through the addition between the original signal and a noise signal that is multiplied with a noise factor. This simulates real-world environments where background noise can be present and helps in making models more robust. Different types of noise can be used, such as white noise, pink noise, brown noise, or realistic environmental sounds like traffic or crowd noise.
* Time shifting moves the audio signal forward or backward in time by a certain amount. This augmentation can help a model become invariant to slight timing differences, which can be especially useful in tasks like speech recognition or event detection where exact timing can vary. This is especially useful for longer audio signals that only have certain segments with sounds, such that the model does not correlate the position in time with the classification; thus, making it more robust.
* Volume change, also known as gain adjustment, involves altering the amplitude of the audio signal to make it louder or quieter. This is achieved by multiplying the audio waveform by a constant factor greater than 1 for amplification or between 0 and 1 for attenuation. Volume change is a simple yet effective augmentation that helps in simulating real-world scenarios where the audio might be recorded at different levels of loudness due to varying distances from the microphone or different recording equipment.

Machine learning models

Support Vector Machines (SVM) are a simple yet widely used classification technique that has shown high accuracy in various pattern recognition tasks. The SVM operates by finding the optimal hyperplane that separates data points of different classes in a high-dimensional space. This hyperplane is determined such that the margin, which is the distance between the hyperplane and the nearest data points of each class, is maximized. By leveraging the kernel trick, SVMs can efficiently perform non-linear classification by implicitly mapping input features into higher-dimensional spaces. This ability to handle complex boundaries and avoid overfitting makes SVMs particularly suitable for tasks involving intricate data structures, such as audio classification.

Random Forest (RF) is an ensemble learning method that uses multiple decision trees during training and outputs the majority vote of the classes for classification. Random Forest reduce the overfitting that individual trees are prone to, through the use of this array of decision trees and the majority vote to decide. Each tree in the forest is trained on a random subset of the data and features, which enhances the model's generalizability and resilience to noise.

Deep learning models

Neural networks are a subset of machine learning that have seen extensive application and development in various domains, including speech and sound recognition, which are pertinent to the baby monitoring system proposed in this study. This section identifies relevant contributions toward progress in neural networks, especially in domains such as audio recognition and health monitoring systems. Neural networks have revolutionized the field of audio recognition, making significant developments in tasks such as speech and sound classification.

DeepSVM [4] is an ensemble learning approach based on SVM to create a decision-making system for the classification of infant crying using windowing and the Fourier transform for feature extraction and selection. In this paper, they managed to classify the reason for which the infant was crying with high accuracy.

In [5], convolutional and recurrent neural networks were integrated for the classification of infant crying using various features extraction methods such as mel frequency, bark frequency and linear prediction cepstral coefficients. They have shown that their model is capable of achieving higher performance than previous models.

In another study [6], a pretrained convolutional neural network, specifically VGG16, was used in conjunction with Gammatone frequency cepstral coefficients and spectrograms in order to diagnose infant crying. They have showed that the best performance was obtained through the fusion of various audio features and the applicability of transfer learning.

Performance evaluation

Accuracy is the most common metric used in classification. However, it can be misleading for imbalanced datasets. Nevertheless, this effect can be mitigated through augmentation for the balancing of class samples. In simple words, accuracy represents the probability with which the classifier predicts correctly and it can be expressed mathematically as:

F1-score is used as an alternative to accuracy, which better reflects the classifier quality. It is often used when optimising recall or precision results in decreased model performance. Mathematically, F1 is the harmonic mean of precision and recall, but it can also be defined using the confusion matrix.

Proposed Method

Overall system architecture

In this paper, we propose a baby monitoring system that uses deep learning for the detection of infant crying but for it to work properly the integration of hardware and software is necessary. The core elements of its architecture and how they interact are shown in Figure 2. These include:

* The user, who is the main actor using the system.
* The Raspberry Pi board, the main component and brain of the entire hardware device.
* Camera, for recording video frames and sending them to the mobile application.
* The microphone and speaker component that plays a dual role in the hardware: sending the data for further processing to the baby crying detection model and communicating with the mobile application both by recording the sound and playing it back through the hardware system speaker, and in the reverse direction, by recording it through the microphone and playing it back to the application.
* Baby cry detection model, which sends notifications to the app via the Flask server.
* The application itself, which communicates with the server via the HTTP communication protocol for audio and video transmission, and via a websocket connection for receiving notifications sent by the server to it.
* SQLite database, for storing user data and using it for further analysis.

A diagram of a computer network

Description automatically generated

*Figure 2 – The general system architecture*

The main advantage of such a system is its high potential for extensibility. Existing functionalities can be enhanced to include additional sensors, integration with other smart devices in the home, or the development of advanced analytics and predictive algorithms. The data collected and analyzed can provide useful information about babies' sleep, feeding and growth patterns, thus helping to identify potential health problems early and improve care strategies.

The central component for the functionality of the baby monitoring system is the deep learning-based baby cry detection model. The accurate detection of infant crying is critical for the system's efficacy. The deep learning model processes audio data in real-time, distinguishing between crying and other sounds. Practically, the deep learning model should provide a binary classification of samples recorded by the microphone, to announce whether a cry was detected. Below, we describe the components and processes involved in the machine learning detection of infant crying.

Data manipulation

To obtain a robust model for such a system, a copious amount of data is required and as such, data was extracted from several sources:

* CryCeleb [7], segmented cry sounds from 786 infants. Only contains examples of crying sounds and were labelled as such in the dataset used.
* Donateacry [8], a corpus of various instances of infant crying that have been classified by the reason for crying. The purpose of the system and the deep learning model is to distinguish between infant crying and any other sound that may be present in the environment. As such, these examples have all been labelled as crying, ignoring the reason.
* LibriSpeech [9], a dataset of read english speech from audiobooks. This dataset was used to enable the model to distinguish between human sounds, specifically to not detect adult speech as infant crying. All these examples have been labelled as ‘not crying’.
* Infant’s Cry Sound [10], a dataset containing 3 classes of crying (due to being hungry, tired or discomfort). These examples have all been labelled as crying, ignoring the reason.
* SilentBabyMonitor [11], containing various classes of sounds, such as noise, silence, crying and laughing. This dataset contains examples of baby crying and laughing. All other sounds of a baby’s repertoire have been labelled as ‘not crying’, in order to increase the robustness of the evaluated models.

All of these dataset have been aggregated to obtain the dataset used in this study. These datasets have been converted to the “.wav” format from their original formats and have either been segmented or padded to a 5 second sample of audio.

Such a system should be capable of detecting crying with a high accuracy. But more than that, it should be able to precisely detect the crying with as few false positives as possible and at the same it should not miss the actual crying of the infant. Thus, such a model requires both high precision and high recall. To increase the robustness of any model trained using this curated dataset, it contains samples of noise, silence, animal sounds, ambiental sounds and adult speech.

Furthermore, these samples have then been augmented to increase the ability of the model to generalize. The data augmentations applied were: pitch shifting, noise addition (white, pink, brown), time shifting and volume change. Through this augmentation of the data, balanced classes can be obtained. The augmented dataset was set to a size of 10000 samples where a random audio sample was chosen from the crying or non-crying samples to obtain a balanced dataset. Each augmentation had a probability of 20% to be applied to the sample. Finally, the dataset was partitioned into training, validation and testing of 70%, 15% and 15% percentages, respectively. The spectrograms and MFCCs of each sample were extracted to be used as the features for training.

Machine learning algorithms

We utilized SVM with a linear kernel for the classification of flattened spectrograms from audio signals. The choice of a linear kernel is motivated by its simplicity and effectiveness in high-dimensional feature spaces. We set the regularization parameter *C* to 1.0 to balance the trade-off between a low training error and a high margin. Features were standardized before training to ensure equal contributions. The dataset was divided into training and testing sets with an 80-20.

We utilized the RF classifier to perform the classification of flattened spectrograms from audio signals. The model is configured to use 100 decision trees. The dataset was divided into training and testing sets with an 80-20 and they were standardized before training.

Deep learning models

Several architectures of neural networks were developed for the evaluation of their ability to classify infant crying. All of these models were trained using the Adam optimizer, with sparse crossentropy loss and they had ReLU activation functions for the hidden layers and softmax for the output layer. The architectures are as follows:

* *CNN1*: Convolutional neural network (1x convolution + 1x max pooling) with 3 dense layers
* *CNN3*: Convolutional neural network (3x convolution + 3x max pooling) with 5 dense layers
* *VGG16* [12]: 16 layers, including 13 convolutional layers and 3 fully connected layers. It is widely used due to its balance of depth and computational efficiency.
* *ResNet50* [13]: introduces residual blocks, which help train deeper networks by avoiding the vanishing gradient problem. This makes it a good choice for large datasets with complex patterns.
* *InceptionV3* [14]: uses parallel convolutions with different kernel sizes, allowing the network to capture a variety of features at different scales.
* *Inception-ResNetV2* [15]: combines the inception architecture with residual connections for enhanced learning capabilities.
* *Xception* (Extreme Inception) [16]: deep learning model that improves upon Inception by replacing the standard Inception modules with depthwise separable convolutions.
* *MobileNet* [17]: lightweight and efficient, making them ideal for mobile and edge devices. They use depthwise separable convolutions to reduce computational cost.
* *DenseNet* [18]*:* connects each layer to every other layer in a feed-forward fashion. This enhances the flow of information and gradients through the network, making it highly efficient.
* *EfficientNet* [19]*:* scales up the network width, depth, and resolution in a compound manner, providing a more balanced approach to scaling compared to other models.

The last 8 architectures are used as backbones (encoders) and they continue with a flattening and a series of 5 dense layers to obtain the classification output. These have been pretrained on the ImageNet dataset. The models are presented in the Results section in two variants, frozen and non-frozen. This choice is applied to the architectures, not the last layers of the model.

Although not presented here, simple neural networks containing only dense layers have been evaluated as well, but they were unable to properly learn as their accuracy remained close to that of random. The same conclusions was reached for using only the signal.

3. Results

For the analysis of this array of neural network architectures, we present the loss (Table 1) and accuracy values (Table 2) for the training, validation and testing of these models. Moreover, the F1-score (Table 4) is presented for the testing data. Furthermore, we have trained each model using the spectrogram or the MFCCs in order to determine which of the two types of features are more appropriate for training of such a model in the detection of infant cries. This same procedure was applied to the machine learning models (SVM and RF), however no loss values are present for these models in Table I and the validation subset of the dataset was moved to the testing subset.

Another analysis on the pretrained models was the option of freezing the encoding layers of the pretrained architectures used. The effect of this choice on accuracy can be observed through the comparison of Table 2 and Table 3. Furthermore, in Table 4, the effect on the F1-score for the testing data is presented.

**Table 1.** Loss values across architectures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **Model** | **Train** | **Validation** | **Test** |
| Spectrogram | SVM | - | - | - |
| Spectrogram | RF | - | - | - |
| Spectrogram | CNN1 | 0.926 | 0.928 | 0.958 |
| Spectrogram | CNN3 | **0.222** | **0.153** | **0.230** |
| Spectrogram | VGG16 | 0.467 | 0.401 | 0.420 |
| Spectrogram | ResNet50 | 0.476 | 0.358 | 0.352 |
| Spectrogram | InceptionV3 | 0.569 | 0.469 | 0.467 |
| Spectrogram | Inception-ResNetV2 | 0.562 | 0.441 | 0.441 |
| Spectrogram | Xception | 0.530 | 0.409 | 0.410 |
| Spectrogram | DenseNet | 0.492 | 0.402 | 0.399 |
| Spectrogram | MobileNet | 0.588 | 0.518 | 0.522 |
| Spectrogram | EfficientNet | 0.554 | 0.467 | 0.470 |
| MFCC | SVM | - | - | - |
| MFCC | RF | - | - | - |
| MFCC | CNN1 | 0.541 | 0.512 | 0.507 |
| MFCC | CNN3 | 0.542 | 0.442 | **0.443** |
| MFCC | VGG16 | **0.488** | 0.679 | 3.582 |
| MFCC | ResNet50 | **0.489** | 0.594 | 1.146 |
| MFCC | InceptionV3 | 0.641 | 0.741 | 1.077 |
| MFCC | Inception-ResNetV2 | 0.617 | 0.676 | 0.994 |
| MFCC | Xception | 0.542 | 0.537 | 0.555 |
| MFCC | DenseNet | 0.501 | **0.407** | 0.614 |
| MFCC | MobileNet | 0.578 | 0.527 | **0.783** |
| MFCC | EfficientNet | 0.607 | 0.992 | 0.986 |

The results show a clear trend favoring spectrogram-based feature representations over MFCCs, particularly for deep learning models. From Table 2, ResNet50 trained on spectrograms achieved the highest test accuracy of 99.6%, outperforming other architectures, including VGG16, InceptionV3, and EfficientNet. This suggests that ResNet’s residual connections facilitate better feature propagation, enabling the model to effectively capture frequency variations characteristic of infant cries.

In contrast, MFCC-based models generally performed worse than their spectrogram-based counterparts. For instance, VGG16 trained on MFCCs exhibited a dramatic drop in accuracy (50.0%) compared to its spectrogram-based equivalent (97.3%), indicating that MFCCs might lack the necessary spectral detail for robust cry classification. However, MobileNet and DenseNet performed relatively well on both spectrogram and MFCC features, suggesting that certain architectures are more adaptable to different feature representations.

**Table 2.** Accuracy values across architectures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **Model** | **Train** | **Validation** | **Test** |
| Spectrogram | SVM | **100** | - | 91.73 |
| Spectrogram | RF | **100** | - | 96.83 |
| Spectrogram | CNN1 | 96.34 | 96.53 | 96.33 |
| Spectrogram | CNN3 | 95.13 | 98.27 | 97.80 |
| Spectrogram | VGG16 | 93.09 | 98.53 | 97.33 |
| Spectrogram | ResNet50 | 95.73 | **99.27** | **99.60** |
| Spectrogram | InceptionV3 | 92.67 | 98.67 | 99.00 |
| Spectrogram | Inception-ResNetV2 | 92.77 | 99.07 | 99.06 |
| Spectrogram | Xception | 93.41 | **99.27** | 99.20 |
| Spectrogram | DenseNet | 93.64 | 99.13 | 99.26 |
| Spectrogram | MobileNet | 93.27 | 99.00 | 98.46 |
| Spectrogram | EfficientNet | 92.47 | 96.73 | 96.40 |
| MFCC | SVM | **100** | - | 92.39 |
| MFCC | RF | **100** | - | 96.26 |
| MFCC | CNN1 | 97.46 | **99.07** | **98.73** |
| MFCC | CNN3 | 97.11 | 98.47 | 98.20 |
| MFCC | VGG16 | 91.13 | 73.07 | 50.00 |
| MFCC | ResNet50 | 95.11 | 86.60 | 50.00 |
| MFCC | InceptionV3 | 91.29 | 85.00 | 51.06 |
| MFCC | Inception-ResNetV2 | 92.16 | 92.07 | 72.80 |
| MFCC | Xception | 93.14 | 93.20 | 55.53 |
| MFCC | DenseNet | 92.37 | 97.73 | 91.42 |
| MFCC | MobileNet | 92.91 | 97.20 | 85.13 |
| MFCC | EfficientNet | 90.71 | 51.00 | 51.26 |

Another observation that can be extracted from these analyses is the difference in performance between frozen (Table 3) and non-frozen (Table 2) pretrained models. While fine-tuned models consistently outperformed their frozen counterparts, the gap was particularly evident for EfficientNet, which struggled significantly when its encoder was frozen, dropping to 50% accuracy. This suggests that EfficientNet relies heavily on fine-tuning rather than feature extraction alone, making it less effective in a frozen state for cry detection tasks.

The only models that maintained strong accuracy regardless of whether their pretrained layers were frozen or not were MobileNet and DenseNet. This indicates that these architectures may have a more flexible feature extraction capability, making them suitable for low-computational scenarios where fine-tuning is not feasible.

**Table 3.** Accuracy values across architectures with a frozen pretrained encoder.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **Model** | **Train** | **Validation** | **Test** |
| Spectrogram | VGG16 | 88.30 | 96.67 | 97.66 |
| Spectrogram | ResNet50 | 83.37 | 94.40 | 95.26 |
| Spectrogram | InceptionV3 | 85.13 | 97.07 | 97.40 |
| Spectrogram | Inception-ResNetV2 | 86.69 | 94.53 | 94.40 |
| Spectrogram | Xception | 91.21 | 97.20 | 96.26 |
| Spectrogram | DenseNet | 90.89 | 97.00 | 97.46 |
| Spectrogram | MobileNet | **91.59** | **98.07** | **98.66** |
| Spectrogram | EfficientNet | 50.74 | 50.00 | 50.00 |
| MFCC | VGG16 | 81.81 | 88.80 | 87.60 |
| MFCC | ResNet50 | 81.41 | 85.67 | 86.13 |
| MFCC | InceptionV3 | 80.79 | 88/13 | 89.27 |
| MFCC | Inception-ResNetV2 | 86.69 | 94.53 | 94.40 |
| MFCC | Xception | 87.73 | 93.27 | 92.40 |
| MFCC | DenseNet | 85.43 | 90.00 | 89.53 |
| MFCC | MobileNet | **83.69** | **95.73** | **94.47** |
| MFCC | EfficientNet | 50.41 | 50.00 | 50.00 |

While accuracy provides a general measure of classification performance, it can be misleading in cases of class imbalance, making the F1-score a more reliable metric for evaluating model effectiveness. As seen in Table 4, the F1-scores of the best-performing spectrogram-based models—ResNet50 (99.59%), DenseNet (99.26%), and Xception (99.20%)—closely align with their high accuracy values, confirming that these models maintain a strong balance between precision and recall. This is particularly important for infant cry detection, where a high recall ensures that actual cries are not missed, while high precision minimizes false alarms. As expected, the trend seen in MFCC-based models persisted with lower F1-scores. This further reinforces the observation that MFCC features may not be as effective as spectrograms for this task. Additionally, EfficientNet, which had already shown poor accuracy when frozen, also recorded an F1-score of 50%, confirming its struggle to learn meaningful cry-related features without fine-tuning.

**Table 4.** F1-Score values for the testing data across architectures.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **Spectrogram** | **Spectrogram** | **MFCC** | **MFCC** |
| **Models** | **Non-frozen** | **Frozen** | **Non-frozen** | **Frozen** |
| SVM | 91.96 | - | 92.39 | - |
| RF | 96.80 | - | 96.19 | - |
| CNN1 | 96.30 | - | 98.72 | - |
| CNN3 | 96.30 | - | **98.18** | **-** |
| VGG16 | 97.40 | 97.64 | 66.66 | 85.93 |
| ResNet50 | **99.59** | 95.26 | 66.66 | 85.10 |
| InceptionV3 | 99.00 | 97.30 | 66.66 | 90.22 |
| Inception-ResNetV2 | 99.06 | 94.64 | 71.90 | 94.43 |
| Xception | 99.20 | 96.16 | 68.10 | 91.85 |
| DenseNet | 99.26 | 97.50 | 91.89 | 88.41 |
| MobileNet | 98.48 | 98.67 | 86.03 | 94.57 |
| EfficientNet | 96.44 | 00.00 | 66.81 | 00.00 |

While deep learning models achieved the highest accuracy overall, traditional machine learning approaches such as Random Forest (RF) and Support Vector Machines (SVM) performed surprisingly well. RF achieved 96.8% accuracy on spectrograms on the testing data, which is comparable to some deep learning models. This suggests that classical methods can still be viable, particularly for applications requiring lower computational costs. SVM performed slightly worse than RF with an accuracy of 92.3% on the testing dataset, indicating that non-linear kernel methods may not be as effective as ensemble-based approaches for this task.

These results indicate that ResNet50 (99.6% accuracy, 99.6% F1-score, non-frozen, spectrograms), DenseNet/Xception (99.2% accuracy, 99.2% F1-Score, non-frozen, spectrograms), InceptionV3/Inception-ResNetV2 (99.0% accuracy, 99.0% F1-Score, non-frozen, spectrograms), and MobileNet (98.6% accuracy, 98.5% F1-Score, non-frozen, spectrograms) are the most viable options for real-time infant cry detection. While ResNet50 provides the highest accuracy, MobileNet is more computationally efficient, making it better suited for edge devices such as IoT-based baby monitors. Nevertheless, simpler models such as CNN1 (98.7% accuracy, 98.7% F1-Score, MFCC) or CNN3 (98.2% accuracy, 98.2% F1-Score, MFCC) obtained a surprisingly high performance, offering a more efficient option with only a slight decrease in performance. Additionally, traditional machine learning models like Random Forest could serve as an alternative in resource-constrained environments.

4. Discussion

The development of an effective baby monitoring system such as is described in this paper depends on the integration of both hardware and software components. The key feature of the system is the baby cry detection model, which must achieve a high accuracy with minimal false positives for a robust performance in real-world conditions. In this work, we present an analysis of various machine learning models (including deep learning models) to determine the best option for such as system. In this analysis, we also include two feature spaces created from the recorded signals based on signal processing and we analyze the impact of training the encoding part of deep learning models.

The creation of a baby cry detection model necessitates a comprehensive dataset, which was compiled from various sources, ensuring a diverse range of cry sounds and other audio inputs for robustness. The extended dataset created through augmentation, enhances the model's ability to generalize in real-world environments.

The evaluation of different machine learning models, including neural network architectures, has provided insights into their performance in classifying infant cries and have demonstrated the potential to achieve high precision and recall rates as indicated by the F1-scores obtained on the previously unseen test data. This dual emphasis on precision and recall is essential to minimize false positives while ensuring that actual cries are not missed.

A critical observation from the different options for data features indicate that the pretrained more complex models are more capable of learning from spectrograms than MFCC features, while previously untrained (and simpler) convolutional neural networks do not have this preference. Normally, it is expected that spectrogram-based feature extraction better preserves the **time-frequency characteristics of the data**, allowing deep learning architectures to distinguish between crying and background noises more effectively, while MFCCs, commonly used in speech processing, may lose essential frequency-related information when applied to this specific task. However, our results suggest that more complex models require more complex data (as accuracy is lower in both trainining and testing for the MFCC data), this is especially visible in the test accuracy and F1-score obtained by the more complex models when applied on the MFCC data compared to the spectrogram data.

For the non-frozen models, we observed that their performance is satisfactory only for the spectrogram features, while for the MFCC features, they are often unable to learn with the validation and test accuracy being close to that of random labelling. This is not the case for the frozen architectures, where the models are capable of learning from both the spectrograms and the MFCC features; however, the performance is slightly lower than that of non-frozen models combined with spectrogram features. The exception is the EfficientNet architecture which is able to learn only on the non-frozen setting combined with spectrogram features. Two models escape this pattern, specifically DenseNet and MobileNet, which are capable of learning from both spectrogram and MFCC features regardless whether their layers are frozen or not.

Overall, from the results obtained, the best option is the non-frozen setting with spectrogram features, where the best performance is obtained by the ResNet50 architecture. However, a simpler architecture such as CNN3 with no pretraining is capable of learning from both sets of features with satisfactory performance (~3% decrease in testing F1-score and ~2% in testing accuracy) with a considerable lower amount of training time. For a more robust architecture, our results indicate that DenseNet and MobileNet are the best options as they are capable of high accuracy on any combinations tested.

While deep learning architectures achieved superior accuracy, it is worth noting that traditional machine learning models such as Random Forest (RF) and Support Vector Machines (SVM) performed surprisingly well for both training and testing, showing their ability to generalise. RF achieved an accuracy of 96.8% on spectrograms, making it competitive with several deep learning models. SVM performed well (92.3%) but was slightly less effective than RF and CNN-based architectures. This suggests that in low-computational environments or situations where deep learning is impractical, RF can serve as a viable alternative for infant cry detection. However, deep learning models retain an advantage in their ability to generalize across more complex and diverse datasets.

In summary, the baby monitoring system proposed in this study is based on and necessitates an automatic cry detection model. We have analysed several architectures with various feature types on a comprehensive dataset to determine the best approach for the detection of infant cries in various real-world conditions and environments. Our findings suggest that deep learning-based baby cry detection is a viable and highly accurate solution for real-world monitoring systems. Among the models tested, ResNet50, DenseNet, and MobileNet performed the best, particularly when trained on spectrogram features and fine-tuned. While traditional machine learning models like Random Forest provide a competitive alternative, deep learning remains the superior approach for real-time, high-accuracy detection. Moving forward, improving model efficiency, reducing false positives, and integrating multimodal data sources will be key in making AI-driven baby monitoring systems more practical and accessible. Future work could also encompass a more comprehensive dataset with various sounds from a baby’s repertoire in order to classify more precisely the current state of the infant.

**Author Contributions:**

Conceptualization, E.R.A., B.I. and D.M.H.; methodology, E.R.A., B.I. and D.M.H.; software, E.R.A. and D.M.H.; validation, E.R.A., B.I. and D.M.H.; formal analysis, E.R.A., B.I. and D.M.H.; investigation, E.R.A., B.I. and D.M.H.; resources, -; data curation, E.R.A. and D.M.H.; writing—original draft preparation, E.R.A. and D.M.H.; writing—review and editing, E.R.A., B.I. and D.M.H.; visualization, E.R.A., B.I. and D.M.H.; supervision, E.R.A. and B.I.; project administration, E.R.A. and B.I.; funding acquisition, -. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used in this study is openly available: CryCeleb [7], Donateacry [8], LibriSpeech [9], Infant’s Cry Sound [10], SilentBabyMonitor [11].

**Conflicts of Interest:** The authors declare no conflicts of interest.

References

[1] R. Cohen, D. Ruinskiy, J. Zickfeld, H. IJzerman, and Y. Lavner, “Baby Cry Detection: Deep Learning and Classical Approaches,” 2020, pp. 171–196. doi: 10.1007/978-3-030-31764-5\_7.

[2] J. Benesty, M. M. Sondhi, and Y. (Arden) Huang, “Introduction to Speech Processing,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., Berlin, Heidelberg: Springer, 2008, pp. 1–4. doi: 10.1007/978-3-540-49127-9\_1.

[3] J. K. Das, A. Ghosh, A. K. Pal, S. Dutta, and A. Chakrabarty, “Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features,” in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Oct. 2020, pp. 1–9. doi: 10.1109/ICDS50568.2020.9268723.

[4] K. Rezaee, H. Ghayoumi zadeh, L. Qi, H. Rabiee, and M. Khosravi, “Can You Understand Why I Am Crying? A Decision-making System for Classifying Infants’ Cry Languages Based on DeepSVM Model,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, Jan. 2024, doi: 10.1145/3579032.

[5] S. Rajagopal, P. Poonkodi, M. Kavitha, and K. Subburathinam, “Premature Infant Cry Classification via Deep Convolutional Recurrent Neural Network Based on Multi-class Features,” *Circuits Syst. Signal Process.*, vol. 42, pp. 1–20, Aug. 2023, doi: 10.1007/s00034-023-02457-5.

[6] Y. Zayed, A. Hasasneh, and C. Tadj, “Infant Cry Signal Diagnostic System Using Deep Learning and Fused Features,” *Diagnostics*, vol. 13, p. 2107, Jun. 2023, doi: 10.3390/diagnostics13122107.

[7] D. Budaghyan, C. C. Onu, A. Gorin, C. Subakan, and D. Precup, “CryCeleb: A Speaker Verification Dataset Based on Infant Cry Sounds,” Mar. 21, 2024, *arXiv*: arXiv:2305.00969. doi: 10.48550/arXiv.2305.00969.

[8] G. Veres, *gveres/donateacry-corpus*. (Jul. 02, 2024). Accessed: Jul. 11, 2024. [Online]. Available: https://github.com/gveres/donateacry-corpus

[9] “Librispeech: An ASR corpus based on public domain audio books | IEEE Conference Publication | IEEE Xplore.” Accessed: Jul. 11, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/7178964

[10] Y. D. Rosita and H. Junaedi, “Infant’s cry sound classification using Mel-Frequency Cepstrum Coefficients feature extraction and Backpropagation Neural Network,” in *2016 2nd International Conference on Science and Technology-Computer (ICST)*, Oct. 2016, pp. 160–166. doi: 10.1109/ICSTC.2016.7877367.

[11] A. E. Jang, *eunbeejang/SilentBabyMonitor*. (Jul. 10, 2024). Python. Accessed: Jul. 11, 2024. [Online]. Available: https://github.com/eunbeejang/SilentBabyMonitor

[12] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *ArXiv151203385 Cs*, Dec. 2015, Accessed: Nov. 15, 2021. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Dec. 11, 2015, *arXiv*: arXiv:1512.00567. doi: 10.48550/arXiv.1512.00567.

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” Aug. 23, 2016, *arXiv*: arXiv:1602.07261. doi: 10.48550/arXiv.1602.07261.

[16] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Apr. 04, 2017, *arXiv*: arXiv:1610.02357. doi: 10.48550/arXiv.1610.02357.

[17] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 16, 2017, *arXiv*: arXiv:1704.04861. doi: 10.48550/arXiv.1704.04861.

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” Jan. 28, 2018, *arXiv*: arXiv:1608.06993. doi: 10.48550/arXiv.1608.06993.

[19] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *ArXiv190511946 Cs Stat*, Sep. 2020, Accessed: Nov. 15, 2021. [Online]. Available: http://arxiv.org/abs/1905.11946