future internet

MDPI

Article

Evaluating Convolutional Neural Networks and Vision Transformers for Baby Cry Sound Analysis

Samir A. Younis, Dalia Sobhy * and Noha S. Tawfik

Computer Engineering Department, Arab Academy of Science and Technology and Maritime Transport, Alexandria 1029, Egypt; samirelgeheiny@gmail.com (S.A.Y.); noha.abdelsalam@aast.edu (N.S.T.)
* Correspondence: dalia.sobhi@aast.edu

Abstract: Crying is a newborn's main way of communicating. Despite their apparent similarity, newborn cries are physically generated and have distinct characteristics. Experienced medical professionals, nurses, and parents are able to recognize these variations based on their prior interactions. Nonetheless, interpreting a baby's cries can be challenging for carers, first-time parents, and inexperienced paediatricians. This paper uses advanced deep learning techniques to propose a novel approach for *baby cry classification*. This study aims to accurately classify different cry types associated with everyday infant needs, including hunger, discomfort, pain, tiredness, and the need for burping. The proposed model achieves an accuracy of 98.33%, surpassing the performance of existing studies in the field. IoT-enabled sensors are utilized to capture cry signals in real time, ensuring continuous and reliable monitoring of the infant's acoustic environment. This integration of IoT technology with deep learning enhances the system's responsiveness and accuracy. Our study highlights the significance of accurate cry classification in understanding and meeting the needs of infants and its potential impact on improving infant care practices. The methodology, including the dataset, preprocessing techniques, and architecture of the deep learning model, is described. The results demonstrate the performance of the proposed model, and the discussion analyzes the factors contributing to its high accuracy.

Keywords: audio processing; cry sound analysis; deep learning; spectrogram; transformer models; convolutional neural networks

Citation: Younis, S.A.; Sobhy, D.; Tawfik, N.S. Evaluating Convolutional Neural Networks and Vision Transformers for Baby Cry Sound Analysis. *Future Internet* **2024**, *16*, 242. https://doi.org/10.3390/fi16070242

Academic Editors: Domenico Santaniello, Mario Casillo and Marco Lombardi

Received: 21 May 2024
Revised: 24 June 2024
Accepted: 25 June 2024
Published: 7 July 2024

1. Introduction

The Internet of Things (IoT) has revolutionized various aspects of our daily lives, and its influence extends to the care and development of children and infants [1,2]. IoT–enabled devices bring about significant improvements in health monitoring, safety, and education for young ones [3–5]. For example, smart baby monitors equipped with sensors can monitor an infant's vital signs, sleep patterns, and room conditions, providing real-time alerts to parents and caregivers [6]. These devices not only ensure a higher level of safety but also enable prompt interventions when necessary. Moreover, educational toys and tools connected to the IoT can facilitate interactive learning experiences tailored to a child's developmental stage, fostering cognitive and motor skills through engaging activities. The incorporation of IoT in these areas not only enhances children's well-being and safety but also provides parents with actionable insights and data-driven decision making, thereby supporting them in their caregiving roles.

For newborns, crying is the primary means of communication. Even though baby cries seem to be similar to each other, they are physically generated and have distinctive features. Experienced carers such as parents, physicians, and nurses can identify these differences based on their previous interactions [7,8]. However, it can be difficult for carers, new parents, and untrained paediatricians to interpret baby cries. Accordingly, it is imperative to differentiate between cries that have distinct meanings by utilizing associated auditory characteristics [9–11]. This can interpret a baby's needs and provide parents with

appropriate ways to calm their infants. It can reduce stress for parents and other carers by helping them to avoid misinterpreting their infants' cries. Moreover, it can prevent infant mistreatment and neglect. Furthermore, analyzing an infant's screams provides a non-invasive way to evaluate their health without needing to perform invasive tests. In this situation, an automatic infant classification system helps medical personnel keep an eye on the baby's health and helps new parents understand their baby's needs based on their cries.

One innovative application of IoT for neonatal care is the real-time monitoring and analysis of infant cries [12]. This process entails recording and labeling baby cry sounds, typically carried out by parents or trained medical professionals. Because infants and young children have limited control over their vocal tracts, their cry sounds are considered short-term stationary signals. However, the available infant cry datasets are relatively small, with most being self-recorded for research purposes and not publicly accessible due to the sensitive nature of the data collection and limited resources. As a result, some studies have turned to generating synthetic data for testing and training their methods [13].

Infant cry categorization is extremely difficult because of the available datasets' small size and class imbalance [14–16]. Dunstan Baby Language [17], Baby Chillanto Database [18], and donate–a–cry Corpus [19] are a few of the datasets that are accessible in the literature. The cry recordings' duration, the classifications of cries they contain, and their sizes differ among these datasets [20].

The categorization of baby cries is typically a signal processing and machine learning task. In particular, it classifies the reason for crying using spectrogram image processing and machine learning approaches. In this context, research has been conducted [11,21–27]. For instance, Sharma et al.'s study [23] focused on using four primary classes: hunger, attention, discomfort, and stomach problems. All other classes were grouped under the category "Unidentified Reasons". To investigate the recordings' acoustic properties, the authors extracted 20 audio variables, such as median/mean frequency, standard deviation frequency, kurtosis, spectral flatness, and more. Three clustering algorithms were used in their experiments: Gaussian mixture models, hierarchical clustering, and k-means clustering. Gaussian mixture models performed the best of the clustering models, showing the fewest overlapping points.

Jiang et al. [24] similarly concentrated on four classes: fatigued, terrified, hungry, and uncomfortable. Several preprocessing procedures are used for cry signals in their suggested model. They first performed linear filtering to highlight high-frequency regions (i.e., High-order Spectral (HiSpec)). The SAI model was then used, which consists of peak picking, power-law compression, half-wave rectification, and a gamma tone filterbank. The next step was to extract logarithmic energy and linear prediction coefficients (LPCs) to model the envelope and amplitude. Ultimately, the researchers employed the Gaussian Mixture Model–Universal Background Model (GMM-UBM) for classification since it performed better than the support vector machine (SVM) model. In light of this, their research shows that the SVM model struggles with damaged signals but works well for small datasets. The more adaptive UBM, GMM-UBM, performs better at identifying baby cry signals than the SVM model. Under clear cry signals, however, there is a considerable difference in the HiSpec feature between SVM and GMM-UBM. By 14.6% and 13.7%, respectively, their suggested auditory model-based feature increases recognition accuracy. However, additional research into alternative techniques for extracting features from the auditory model remains imperative.

Furthermore, the authors of [25] proposed a deep learning-based ensemble model that uses three different input types: MFCC features, STFT features, and face photos of infants in tears. Applying the donate–a–cry Corpus data [19], the authors validated the sound features retrieved using the suggested STFT and MFCC methods. They used a combination of autoencoders and deep neural network (DNN) algorithms to classify audio features. Ozseven [27] recently investigated various methods for analyzing baby cries, including cutting-edge deep learning models and conventional machine learning models using man-

ually constructed features. Their analysis of the donate–a–cry dataset showed that the most promising outcomes came from separating crying signals into spectrograms and scalograms and then classifying the data using either ResNet-18 or Shuffle-Net. Scalograms with ResNet-18 and Shuffle-Net and spectrograms with ResNet-18 were their best-performing models. The use of pre-trained CNN models limits the usefulness of this work. Similar to this, the authors of [26] studied feature ablation using multiple spectral features and a range of classifiers, including Random Forest, SVM, KNN, and Logistic Regression, on recordings belonging to four categories: burping, hungry, belly pain, and discomfort. The most effective features, when combined with the Random Forest and K-Nearest Neighbor algorithms, were found to be Mel-frequency cepstral coefficients (MFCCs), Gamma Tone Frequency Cepstral Coefficients (GFCCs), and Zero Crossing Rate (ZCR).

Most studies used convolutional neural network (CNN) architectures, neural network-based classifiers, and conventional machine learning classifiers that used hand-crafted features, according to an overview of the literature on baby cry classification. After reviewing the literature, we discovered that transformers and attention-based categorization techniques still need to be applied in these investigations [28]. Transformer architecture has been applied, adapted, and extended to various areas, including computer vision, audio processing, and more, following its breakthrough in natural language processing (NLP) applications. Transformers have also been used in the audio domain for various tasks, including audio categorization, speech recognition, and music production [29–31].

Our paper contributes to the research literature as follows:

- We carried out an extensive cry sound analysis study involving data preprocessing and the creation of spectrograms from cry audio signals generated from IoT sensors.
- Addressing the challenges of limited data and class imbalance, we implemented data augmentation techniques that proved to be highly effective.
- Our research introduced an improved Vision Transformer architecture, which not only was optimized for small-sized datasets but also outperformed all other published results across the five classes in the donate–a–cry dataset, a significant achievement.

The remainder of this paper is structured as follows. Section 2 describes the methodology. Section 3 presents the experimental evaluation of our method. Section 5 discusses the threats to validity. The paper concludes and provides future work in Section 6.

2. Methodology

2.1. Data

The data used in our research consist of cry audio recordings obtained from the donate–a–cry Corpus [19]. It comprises baby cries recorded by volunteers using their cell phones through a mobile application and made publicly available through a GitHub repository. The dataset initially consisted of 1128 audio recordings, divided into nine classes (hungry, needs burping, belly pain, discomfort, tired, lonely, cold/hot, scared, don't know). However, due to labeling inconsistencies caused by volunteers, the research group performed data cleaning, reduced the dataset in size, and used the Dunstan Baby Language (DBL) [17] categories as a reference. The dataset includes a total of 462 recordings covering five classes of baby cry analysis: belly pain (17 recordings), burping (9 recordings), discomfort (28 recordings), hungry (383 recordings), and tired (25 recordings). Moreover, each record also contains information about the gender (male, female) and age of the baby (0–4 weeks, 4–8 weeks, 2–6 months, 2 months–2 years, or more). The recordings vary in length from 3 to 6 s and are available in wav format. The sampling frequency is 8000 Hz, and the bit rate is 128 kbps.

2.2. Data Preprocessing and Augmentation

Since the data were crowd-collected, a data cleaning step was mandatory to ensure the quality of all records. All the cry recordings were manually audited, and any irrelevant audio segments were removed in the case of empty records or trimmed in the case of records with trailing/opening non-cry sounds. This step ensured that the dataset only contained

cry sounds relevant to the classification task. After the cleaning process, the distribution of cry sound classes was as follows: burping (8 samples), discomfort (24 samples), hungry (252 samples), pain (14 samples), and tired (17 samples).

A significant class imbalance in the dataset presents challenges in training a robust cry classification model. The dataset has an uneven distribution among the classes, with the *hungry* class having a more significant number of samples than the other classes. This imbalance raises concerns about potential bias towards the majority class and the risk of overlooking the minority classes during model training [27].

To mitigate the effects of limited data and class imbalance, four data augmentation techniques were systematically employed. First, we applied non-overlapping random cropping to extract smaller segments of different durations from the cry recordings. Specifically, we generated randomly cropped segments, each with a duration of 3 s, introducing temporal variations and enabling the model to handle cry sounds of different lengths [32].

Next, we apply noise injection, enhancing the model's ability to generalize across different environmental conditions. This technique involved adding Gaussian noise, also known as *white noise*, to the cry recordings. Gaussian noise is a type of random noise that simulates various background noise levels, improving the model's performance in real-world scenarios where cry sounds are often accompanied by ambient noise [33].

Additionally, volume adjustment was also performed to modify the sound intensity of the cry recordings. By adjusting the volume, our model recognizes crying sounds with different intensity levels, improving its robustness to variations in sound amplitude.

Finally, speed perturbation and pitch variation techniques were employed to introduce variations in the tempo and pitch of the cry sounds, respectively. These techniques enabled the model to recognize cry patterns across different speeds and frequencies, enhancing its ability to generalize and classify cry sounds accurately [32].

To ensure the absence of duplicates in the augmented dataset, several key measures were implemented during the data augmentation process. Firstly, unique random seeds were utilized for each augmentation operation, including cropping, noise injection, volume adjustment, speed perturbation, and pitch variation. This ensured that the random variations applied to each audio segment were distinct, reducing the risk of generating duplicates.

Additionally, meticulous records of augmentation choices were maintained for each audio sample, encompassing parameters such as crop duration, noise level, volume adjustment factor, speed factor, and pitch shift amount. These records facilitated subsequent checks for similarity between augmented samples.

To further prevent duplicates, a robust hashing mechanism was employed. Unique hashes were generated for each augmented audio segment based on its content, using perceptual audio hashing techniques. Prior to adding a new augmented sample to the dataset, hashes were compared with those of previously generated samples to identify duplicates or highly similar segments.

In cases where duplicates were detected, a rejection sampling strategy was employed. Duplicate or overly similar augmented samples were promptly discarded, and the augmentation process was reattempted with different parameters to obtain diverse samples. Randomization and shuffling of augmentation order and original audio samples were also employed to introduce additional randomness and minimize the risk of duplicates.

These comprehensive measures were instrumental in significantly reducing the likelihood of duplicates in the augmented dataset, ensuring a diverse and unbiased training dataset for cry sound classification.

As a result of the employed augmentation techniques, the dataset was expanded to 1252 samples. Each class was balanced with 250 samples, except for the hungry class, which had 252 samples. This balancing approach addressed the class imbalance issue and ensured the model received equal exposure to each class during training.

2.3. Feature Extraction

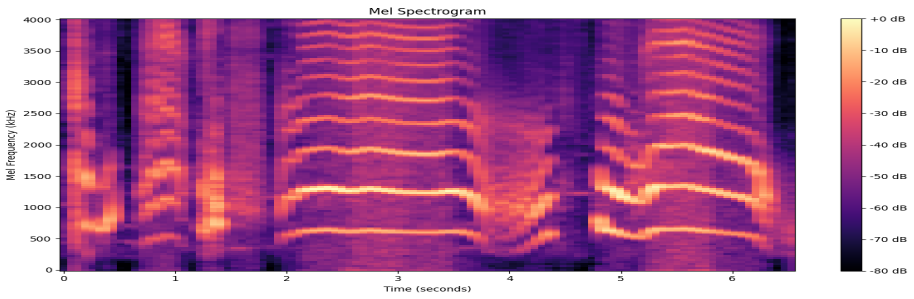
2.3.1. Mel Spectrogram

Mel spectrograms provide a valuable means of representing audio characteristics in diverse different classification tasks [34–36]. During a baby's cry, the acoustic properties of that sound carry information about its physiological and emotional state. The Mel spectrogram is a visual depiction of the frequency spectrum of an audio signal that is transformed onto the Mel scale. This scale is designed to simulate the way the human ear perceives different frequencies. Fundamentally, it serves as a mechanism for representing the energy composition across distinct frequency ranges, providing both temporal and spectral characteristics of an auditory stimulus. In the Mel spectrogram, the temporal aspect is represented by the X-axis, while the Y-axis represents the frequency. The intensity of color corresponds to the amplitude or power of the frequencies. Typically, the process involves a sequential execution of steps, beginning with segmenting the audio signal into short frames and then applying the Fourier Transform to convert the time-domain signal of each frame into the frequency domain. Subsequently, the Mel scale approximates the human ear response and is employed to partition the frequency spectrum into various bands [37]. The Mel spectrogram conversion can be mathematically represented as:

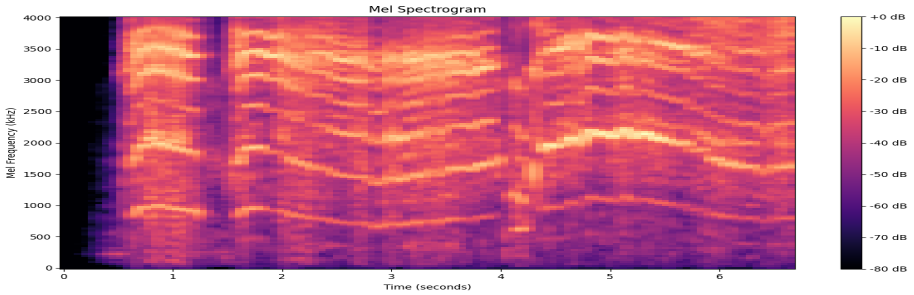
$$STFT(cry\_recording) = \int cry\_recording(t) \cdot e^{-2\pi i f t}\, dt \tag{1}$$

$$Mel(STFT(cry\_recording)) = \log\left(1 + \frac{|S(f)|^2}{N_{MELS}}\right) \tag{2}$$

where $STFT$ represents the Short-Time Fourier Transform, t is the time variable, f is the frequency variable, and $S(f)$ is the STFT magnitude. *Mel* denotes the Mel filterbank transformation, and N_{MELS} is the number of Mel filters used. The resulting Mel spectrogram serves as a more perceptually relevant depiction of sound; samples of cry sounds from the donate–a–cry dataset mapped to the Mel spectrogram scale are illustrated in Figure 1. The process of feature extraction using Mel spectrograms entails converting the preprocessed cry recordings into a more computationally comprehensible format, which enables the use of machine learning algorithms.

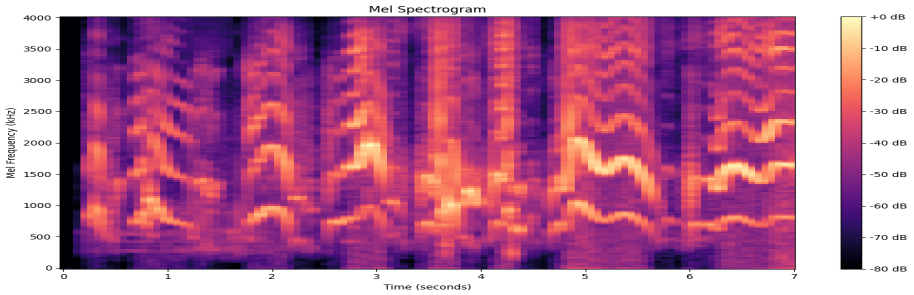


(**a**) Mel spectrogram of a baby crying due to belly pain (Class 1).

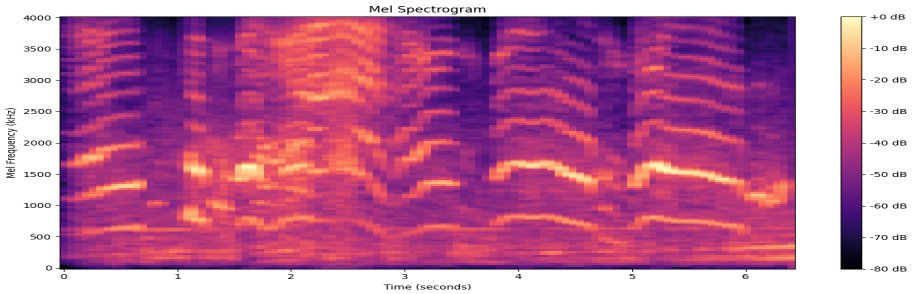


(**b**) Mel spectrogram of a baby crying from burping (Class 2).

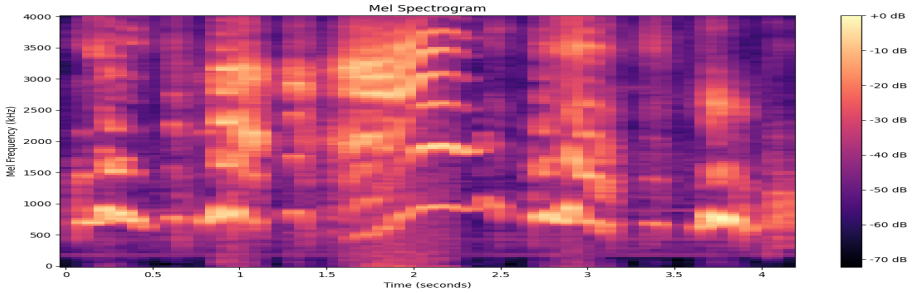
Figure 1. *Cont.*



(**c**) Mel spectrogram of a baby crying due to discomfort (Class 3).



(**d**) Mel spectrogram of a baby crying due to hunger (Class 4).



(**e**) Mel spectrogram of a baby crying when tired (Class 5).

Figure 1. Mel spectrograms of the five classes included in the donate–a–cry dataset.

2.3.2. CQT Spectrogram

The Constant-Q Transform (CQT) spectrogram is a valuable tool for analyzing baby cry sounds. It is particularly useful for capturing audio features at different pitch scales [38,39], which can be essential for distinguishing the nuances in baby cries. The CQT spectrogram is computed using the following equation:

$$CQT(cry\_recording) = |CQ(x)(f)| = \left| \int x(t) \cdot e^{-2\pi i f t} \, dt \right| \tag{3}$$

In the equation, $x(t)$ represents the input audio signal, f is the frequency variable, and $CQ(x)(f)$ is the CQT magnitude at frequency f. The CQT is particularly adept at capturing the varying pitch characteristics present in baby cries, which can convey important information about the baby's needs and condition.

Using the CQT spectrogram for baby cry analysis allows for a more detailed and pitch-sensitive representation of the cries, making it a valuable feature extraction technique for machine learning algorithms aimed at understanding and classifying baby cries.

2.4. Classification

For this paper, we experimented with several models (the data augmentation techniques, spectrograms, and models we experimented with are available at https://github.com/SamirElgehiny/baby-cry-analysis (accessed on 1 June 2024) to develop an effective cry classification system. By exploring these diverse architectures, we aimed to leverage the strengths of each model and identify the most effective approach for cry sound analysis. Each model's training process involved optimizing model parameters and hyperparameters, such as learning rate, batch size, and regularization techniques, to achieve the best cry classification performance.

2.4.1. Convolutional Neural Network with Mel Spectrograms (CNN)

We employed a CNN architecture specifically designed for cry sound analysis using *Mel spectrograms* [40]. The model consists of several convolutional layers, activation functions, and pooling layers, followed by fully connected layers and a softmax output layer. The convolutional layers in the model serve as feature extractors, capturing relevant patterns and acoustic features from the Mel spectrograms. The activation functions introduce non-linearity, enabling the model to learn complex relationships within the cry sounds. The pooling layers help reduce the spatial dimensions and retain the most salient features. The output of the convolutional layers is flattened and passed through fully connected layers to further refine the representations. The final softmax activation function produces the predicted probabilities for each cry class.

1. *Convolutional layers:* Three convolutional layers are used with increasing filter sizes (32, 64, 128). This helps in learning hierarchical features from Mel spectrograms. The Rectified Linear Unit (ReLU) activation function is applied after each convolutional layer to introduce non-linearity and enhance feature representation.
2. *Pooling layers:* Max-pooling layers with a pool size of (2, 2) follow each convolutional layer. This reduces spatial dimensions and extracts essential features, contributing to computational efficiency.
3. *Flatten layer:* Positioned after the convolutional layers, the flatten layer transforms the spatially structured output into a flat vector. This facilitates the integration of the model with subsequent dense layers for classification.
4. *Dense layers:* Two dense layers with 256 and 128 neurons are incorporated for high-level feature learning and classification. Dropout layers with a rate of 0.5 are introduced after dense layers to prevent overfitting. Dropout randomly deactivates connections during training, promoting a more robust model.
5. *Output layer:* The output layer has five neurons that work together with a softmax activation function. This is because the classification task needs five classes corresponding to a different baby cry pattern.

The described architecture is designed to process and sort audio signals efficiently by focusing on obtaining hierarchical features from spectrograms and stopping overfitting using dropout layers. The overall design aligns with established practices in audio signal analysis and classification.

2.4.2. Multiple-Input CNN with CQT and Mel Spectrograms (Multi-CNN)

The multiple-input CNN model was designed to take advantage of both Constant-Q Transform (CQT) and Mel spectrograms for cry sound analysis. The architecture consists of parallel convolutional neural networks, with each network processing one type of spectrogram. The CQT network and Mel network are connected at the feature fusion layer, where the extracted features from both networks are combined. This fusion enables the model to capture complementary information from both spectrogram types. Each network consists of several convolutional layers, followed by activation functions and pooling layers to extract hierarchical features from the input spectrograms. The number of filters, kernel sizes, and pooling sizes are carefully selected to balance the model's capacity and

complexity. Batch normalization and dropout layers are also included to regularize the model and prevent overfitting. The output of the convolutional layers is flattened and fed into fully connected layers, which gradually reduce the dimensionality and extract more abstract representations. Finally, a softmax activation function is applied to generate the class probabilities, indicating the predicted cry class.

1. *Convolutional layers for Mel spectrograms:* The Mel spectrogram branch consists of three convolutional layers (32, 64, and 128) with ReLU activation, followed by max-pooling layers for spatial dimension reduction. The architecture aims to capture hierarchical features from Mel spectrograms.
2. *Convolutional layers for CQT spectrograms:* The CQT spectrogram branch mirrors the structure of the Mel branch, featuring three convolutional layers (32, 64, and 128) with ReLU activation and max-pooling layers. This design choice ensures a consistent approach for feature extraction from both types of spectrograms.
3. *Merging of branches:* The outputs of the Mel and CQT branches are concatenated, creating a unified feature representation that combines the distinctive information from both types of spectrograms. This merged tensor serves as input to the subsequent dense layers.
4. *Dense layers for classification:* The merged tensor undergoes dense layers with 256 and 128 neurons, each followed by dropout layers with a rate of 0.5. These layers facilitate high-level feature learning and classification. The softmax activation function in the output layer ensures the model's ability to classify baby cries into five distinct classes.

This innovative neural network architecture strategically leverages the complementary information embedded in Mel and CQT spectrograms. By synergistically combining these spectrogram types, the model is primed to discern intricate patterns within baby cries, enhancing its ability to capture nuanced acoustic features. This sophisticated fusion ensures a more comprehensive representation of infant vocalizations and signifies a pivotal step towards refining baby cry analysis for advanced well-being monitoring systems.

2.4.3. Vision Transformer (ViT)

The Vision Transformer model, originally proposed for image classification [41,42], was adapted for cry sound analysis. The architecture comprises a stack of transformer layers, which consist of self-attention mechanisms and feed-forward neural networks. In the cry sound analysis context, the input cry spectrograms are reshaped into a sequence of patches, denoted by:

$$P \in \mathbb{R}^{N \times D}. \tag{4}$$

where $N = \frac{H \times W}{D}$ and D is the patch size. Each patch p_i is then embedded using an embedding layer, resulting in an embedded patch. The transformer layers enable the model to attend to different patches and capture the relationships between them. Given an embedded patch sequence $E = (e_1, e_2, \ldots, e_N)$, the self-attention mechanism calculates the attention scores which are then passed through a feed-forward neural network (FFN) layer. The transformer layers are stacked to attend to different patches and capture their relationships. The output of the last transformer layer is passed through a classification head, typically consisting of one or more fully connected layers and a softmax activation function, to produce the predicted cry class probabilities.

2.4.4. Vision Transformer with Shifted Patch Tokenization and Locality Self-Attention (ViT-SPT-LSA)

Building upon the Vision Transformer, we apply modifications to enhance its performance for cry sound classification. These modifications include shifted patch tokenization and locality self-attention following the approach proposed in [43]. Shifted patch tokenization involves shifting the patch positions during tokenization, allowing the model to capture more fine-grained details within the spectrograms. Let P' be the sequence of shifted patches and $E' = (e'_1, e'_2, \ldots, e'_N)$ be the embedded sequence of shifted patches. To consider

shifted patches and capture temporal information, positional encodings are added to the embedded patches. For the locality self-attention modification, the attention mechanism is modified to attend to local regions of the cry spectrograms. The attention matrix $A \in \mathbb{R}^{N \times N}$ is calculated as:

$$A_{ij} = Softmax\left(\frac{(E'_i \cdot (E'_j)^T) \odot Mask_{ij}}{\sqrt{E'}}\right) \quad (5)$$

The Mask_{ij} enforces the local attention constraint by setting non-local positions to zero, allowing the model to capture local dependencies and patterns within the cry sounds. The remaining aspects of the architecture and training process for *ViT-SPT-LSA* generally follow the same principles as the Vision Transformer discussed in Section 2.4.3.

Figure 2 illustrates the shifted patch tokenization process employed in our modified Vision Transformer (ViT) model, and Figure 3 demonstrates the architecture of the final modified model. As illustrated, the original input Mel spectrogram is augmented through techniques such as horizontal flipping, random rotation, and random zoom. The SPT component divides augmented images into smaller patches. Each patch represents a portion of the image and is treated as a token for further processing. These patches are shown as individual blocks. Next, the Patch Encoder processes these tokenized patches, preparing them for further transformation within the model. Finally, a series of transformer blocks are repeated eight times within the model. Each transformer block captures hierarchical features within the image. These blocks typically consist of components like Multi-Head Self-Attention, Layer Normalization, Multi-Layer Perceptron (MLP), and Skip Connection layers. These components collectively enable the model to understand and process the input image in a hierarchical manner. Within each transformer block, Multi-Head Self-Attention is a core component. It allows the model to capture dependencies between patches by assigning different attention weights. It is worth noting that Layer Normalization is applied before each attention mechanism and MLP, enhancing training stability.

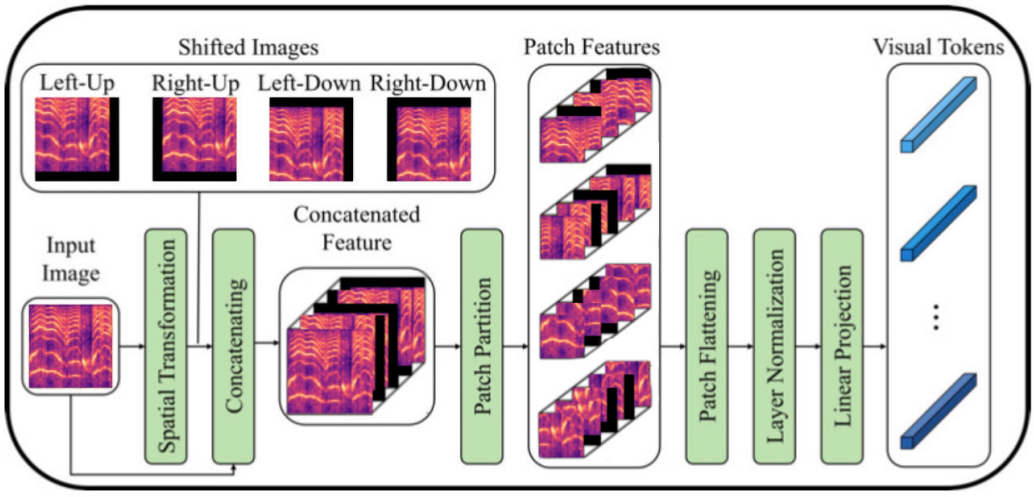


Figure 2. Shifted patch tokenization process [43].

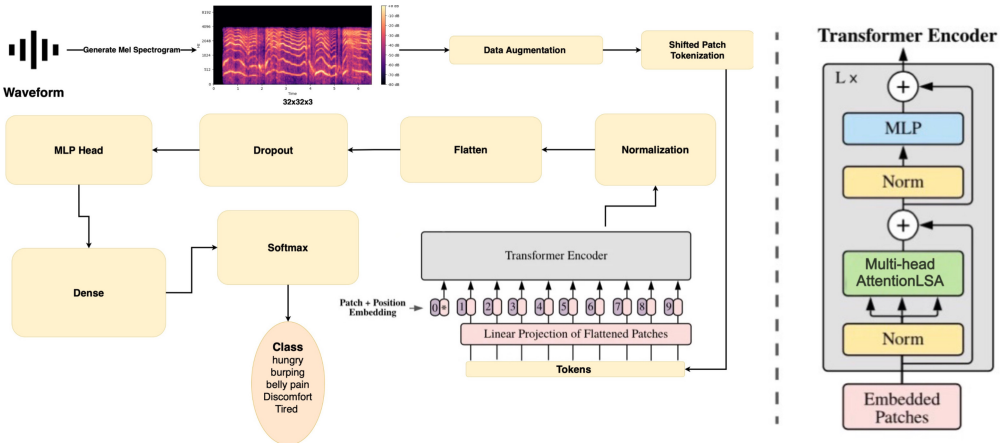


Figure 3. The architecture of ViT-SPT-LSA.

3. Results

In this work, we conduct a comparative analysis of using CNN and transformer-based architectures and their ability to classify cry sounds accurately. In this context, we experimented with the four models mentioned in Section 2.4: *CNN*, *Multiple-input CNN*, *Vision Transformer*, and *Vision Transformer with Shifted Patch Tokenization and Locality Self Attention modification*. We then provided a comparative analysis against the state-of-the-art studies discussed in Section 1. We now discuss the findings of our experiments by evaluating each model using performance metrics, including class-wise accuracy, precision, recall, and F1, in addition to overall accuracy. Cross-validation (5-fold) was applied during the training process to prevent models from overfitting. For all models evaluated, the hyperparameters used are defined as follows: The learning rate is set to 0.001, which controls the step size during gradient descent. The weight decay, a regularization parameter, is assigned a value of 0.0001 to help prevent overfitting by penalizing large weights. The batch size, determining the number of samples processed before the model is updated, is set to 256.

Table 1 summarizes the results obtained from each model across the five different classes: *belly pain*, *burping*, *discomfort*, *hungry*, and *tired*. It also indicates the overall accuracy of each model. The high accuracy achieved by the four models signifies their potential to support infant care and improve parent–infant communication. By accurately identifying the specific reasons behind a baby's cry, parents and caregivers can promptly address the infant's needs and provide appropriate care and comfort.

Overall, ViT-SPT-LSA outperformed the other methods by achieving a relatively high accuracy across all classes, with particularly notable performance in three classes: discomfort, hungry, and tired (95.4%) and Class 3 (97.9%). The second best is the Multi-CNN model which achieved a total of 96.3% with 100% accuracy for Classes 1, 2, and 3. The CNN and the Vanilla ViT models achieved 95.9% and 94%.

More in-depth analysis shows that ViT-SPT-LSA has a precision of 1 and a recall of 0.98 for both the belly pain and burping classes and an F1 score of 0.99. The F1 scores for the discomfort, hungry, and tired classes are 0.97, 0.99, and 0.98, respectively, as shown in Table 2. It is worth noting that the Multi-CNN architecture also achieved a 96.3% accuracy, outperforming almost all previous research.

To further investigate the effectiveness of the augmentation pipeline, we evaluated the best performing model, ViT-SPT-LSA, using a minimally augmented version of the donate–a–cry dataset, which relied solely on random crops of the existing original data. This new subset comprises 53 instances for burping, 144 for discomfort, 102 for tiredness, 84 for belly pain, and 252 for hunger. Our findings reveal a decrease in performance, as evidenced by our best model achieving only 72.1% accuracy.

Additionally, the ViT-SPT-LSA model exceeded the accuracy and F1 scores published in previous research on the same dataset, as shown in Table 3. We attribute such performance

to the effectiveness of both the classification and feature extraction phases. Not only is the Mel spectrogram feature extraction simple to compute and interpret, but it may also be more advantageous than MFCCs for cry signal processing since it can record complicated frequency variations and maintain the complete frequency spectrum. This is crucial for examining acoustic cues in infant cries. Secondly, Mel spectrograms provide a comprehensive visual representation of the signal's frequency content. This allows the ViT's self-attention mechanism to efficiently capture long-range dependencies within the spectrogram images.

Table 1. The accuracy of 5 classes and their overall accuracy for the four models.

	CNN	Multi-CNN	ViT	ViT-SPT-LSA
Class 1—belly pain	96%	**100%**	84%	97.8%
Class 2—burping	**100%**	**100%**	**100%**	95.4%
Class 3—discomfort	90%	87.5%	93.7%	**97.9%**
Class 4—hungry	97%	94%	97.4%	**100%**
Class 5—tired	**100%**	**100%**	83.3%	**100%**
Overall	95.9%	96.3%	94%	**98.33%**

Table 2. Classification report for five classes for ViT-SPT-LSA.

Class	Precision	Recall	F1 Score	Support
Class 1—belly pain	**1.00**	0.98	**0.99**	47
Class 2—burping	**1.00**	0.98	**0.99**	44
Class 3—discomfort	0.98	0.96	0.97	49
Class 4—hungry	0.98	**1.00**	**0.99**	**51**
Class 5—tired	0.96	**1.00**	0.98	48

Table 3. Comparison with state-of-the-art models on donate–a–cry dataset.

	Accuracy	F1 Score	Feature	Classifier	# of Classes
Sharma et al. (2019) [23]	81.7%	0.93	Audio features	GMM	5
Jiang et al. (2021) [24]	-	0.93	SAI modeling	GMM-UBM	4
Cha and Bae (2022) [25]	85.59%	-	STFT	DNN	4
Kulkarni et al. (2021) [26]	84%	0.80	GTCC	Random Forest	4
Ozseven (2022) [27]	95.2%	0.94	Scalogram	ResNet-18	5
ViT-SPT-LSA	**98.33%**	**0.98**	**Mel spectrogram**	**ViT**	5

4. Discussion

The findings agree with our assumptions and the previous findings of other research [43] that the modified architecture for the ViT model allows a more accurate pattern recognition for such a relatively modest-size dataset like donate–a–cry, even after applying the augmentation techniques. As reported in the results, the original Vision Transformer architecture performed worse than both models based on convolutional neural networks when applied to donate–a–cry datasets. Employing shift patch tokenization (SPT) and locality self-attention (LSA) techniques significantly enhanced the performance of ViT. SPT effectively incorporates spatial information into visual tokens through multiple transformations, while the LSA mechanism increases the local attention through softmax with learnable parameters. These changes allowed the model to capture finer-grained details within the spectrograms and focus on local dependencies. They separately augment the locality's inductive bias of ViTs to acquire knowledge from limited-scale datasets without any prior training.

The confusion matrix in Figure 4 and Table 2 illustrates the performance of model ViT-SPT-LSA in classifying cry sounds across different classes during the last fold. Each row represents the actual cry class, while each column represents the predicted cry class. The numbers in the matrix indicate the count of cry instances classified into each class. It is important to note that while the models demonstrated high accuracy, there may be limitations and challenges associated with certain classes. The ViT-SPT-LSA model struggles with the recall of the discomfort class with a recall value of 0.96, which decreases the accuracy of the specific class compared to the other and the rest of the models in the same class. On the other hand, the model exhibits robust performance, particularly excelling in the belly pain and burping classes with strong precision, recall, and F1 scores ranging from 0.98 to 1.

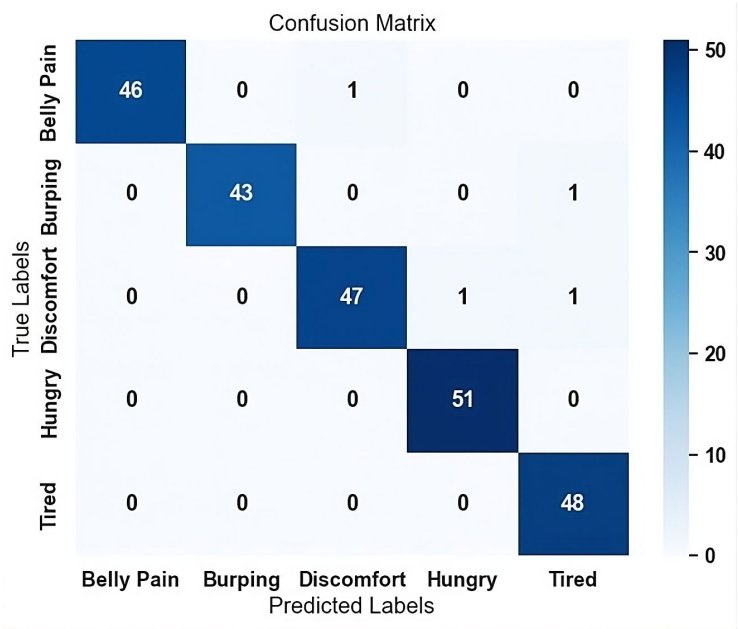


Figure 4. Confusion matrix for the 5th fold.

Additionally, Figures 5 and 6 provide insights into the performance, learning progress, and generalization ability of model *ViT-SPT-LSA* in accurately classifying cry sounds. Figure 5 displays the training and testing accuracy curves for model *ViT-SPT-LSA* during the training process. The *x*-axis represents the number of training iterations or epochs, while the *y*-axis represents the accuracy values. The curve illustrates the accuracy improvement over the training iterations and the model's ability to generalize well to unseen test data. Furthermore, it depicts the training and testing loss curves for model *ViT-SPT-LSA* during the training process. The *x*-axis represents the number of training iterations or epochs, while the *y*-axis represents the loss values. The curve demonstrates the decrease in loss over the training iterations, indicating the model's learning progress and its ability to minimize errors during training.

We further investigated the effectiveness of our augmentation pipeline by conducting an experiment on a reduced dataset of only 635 samples. This subset results from solely applying non-overlapping random cropping to the original data. In this subset, we preserved the original class imbalance present in the dataset, ensuring that the distribution of instances among the classes remained the same as in the original data to evaluate the model's performance when certain classes are underrepresented. Table 4 presents the performance of the enhanced ViT model across all classes. As seen in Table 4, we find a high recall but low precision in instances of belly pain, burping, discomfort, and hunger. This indicates that the model is good at identifying positive instances but has a high rate of false positives. In this experiment, the model struggles significantly with the tired class as it has both low precision (0.17) and low recall (0.43), which indicates that the model

often misclassifies instances of this class. On the other hand, the burping class shows the best overall performance. It has the highest F1 score among the classes, reflecting a good balance between precision and recall.

Table 4. Classification report for five classes for ViT-SPT-LSA on donate–a–cry dataset without augmentations.

Class	Accuracy	Precision	Recall	F1 Score
Class 1—belly pain	0.78	0.37	**1.0**	0.54
Class 2—burping	**0.92**	**0.5**	0.98	**0.67**
Class 3—discomfort	0.56	0.33	0.96	0.50
Class 4—hungry	0.50	0.44	0.94	0.60
Class 5—tired	0.57	0.17	0.43	0.25

Figure 5. Training and testing accuracy.

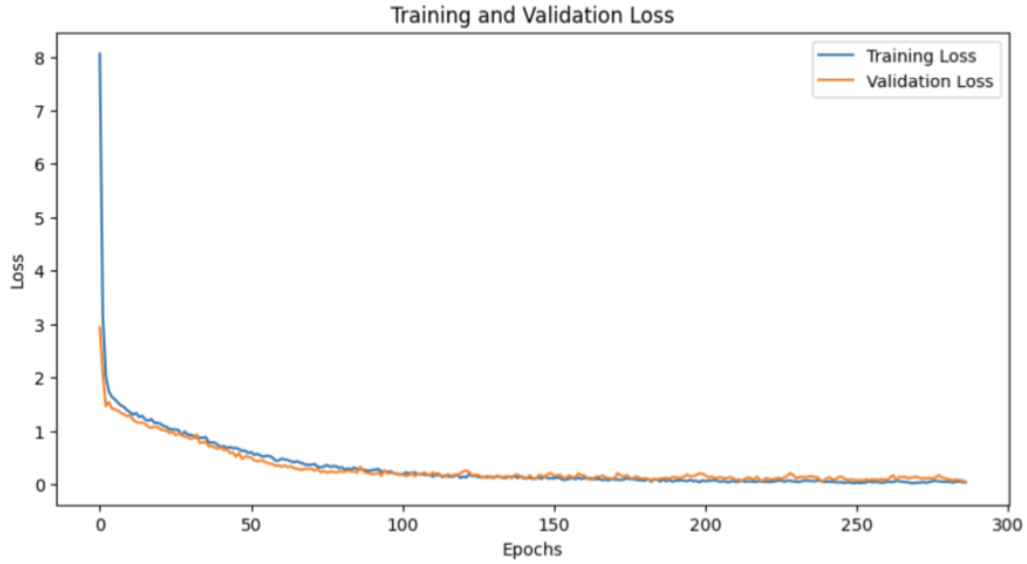


Figure 6. Training and testing loss.

5. Threats to Validity

In this section, we discuss potential threats to the validity of our research findings and the possible measures we took to reduce risks, whenever applicable. One threat to validity is sampling bias; our study used the donate–a–cry Corpus [19] as the primary dataset for cry sound analysis. While it provides a crowd-sourced collection of cry recordings, it may only partially represent the diversity of cry sounds encountered in real-world scenarios. Sampling bias in the dataset could impact the generalizability of the used models. In this context, we plan to experiment with other cry datasets mentioned in Section 1.

A key challenge in our study was using a limited crowd-sourced cry sound dataset. However, we systematically mitigated this limitation through data cleaning, augmentation, and robust preprocessing. These strategies allowed us to harness the full potential of our dataset, resulting in reliable and meaningful cry sound classification models. Despite the dataset's constraints, our study's findings are promising, providing valuable insights into the ability of transformer architecture to classify cry sounds. When generalizing the findings, it is important to acknowledge these limitations, and further research with larger datasets could improve the accuracy of our models and confirm our findings.

There is a threat to validity since the experiments performed for cry classification in previously published research on the donate–a–cry dataset do not use the original data directly. Therefore, the results of different research projects related to cry classification using the donate–a–cry dataset cannot be generalized. Various research works attempt to apply different data augmentation techniques to increase the number of samples to enhance prediction in the donate–a–cry dataset [19]. In Sharma et al.'s study [23], for the training process, the authors included all data from the donate–a–cry dataset [19], and an additional 150 baby crying audio samples were recorded and labeled for fine-tuning and testing purposes. Further, various versions of the donate–a–cry dataset exist due to the different extraction methods used by different authors. For example, for his analysis, Ozseven [27] selected only recordings of infants aged between 0 and 6 months free from environmental noise. Each of these recordings was segmented into one-second intervals with a 50% overlap. This process yielded 1987 samples and then was divided into training and testing sets. These versions are independently compiled; we contacted the original creators to verify their availability but, unfortunately, had no luck. To address these constraints and ensure research reproducibility, we have made our code publicly available (https://github.com/SamirElgehiny/baby-cry-analysis (accessed on 1 June 2024)). This includes steps for data cleaning, augmentation, and the generation of the 1252 samples utilized in our study, enabling others to reproduce our experiments and results.

Additionally, each cry dataset contains different types of cry classes. Focusing on the donate–a–cry dataset [19], some of the research in the literature uses only a subset of four cry classes from the dataset. In comparison, others use the five main cry classes defined in the dataset, as summarized in Table 3.

6. Conclusions and Future Plans

By integrating IoT-enabled sensors and devices with advanced audio processing algorithms, healthcare providers can continuously monitor the acoustic environment of an infant. These sensors capture cry sounds, which are then analyzed using machine learning models to detect patterns and identify potential distress signals. This work aimed to distinguish between different types of generated infant cry sounds and determine the underlying causes of each cry. To this end, we proposed a novel cry classification approach. We carried out a study on cry sound analysis using various models and data preprocessing methods. Our auditing process involved removing unnecessary audio clips from the cry recordings, such that the dataset only included cry sounds essential to our categorization objective. By employing augmentation techniques, we added more diversity in time, noise, volume, speed, pitch, and other aspects of the dataset. Our classification algorithms are well-founded, mainly due to these preprocessing steps that effectively tackle the problems of insufficient data and class imbalance.

Four approaches were implemented, each with its own architecture and features. Our findings highlighted each method's advantages and disadvantages and the accuracy range with which they could classify cry sounds. In contrast to the CNN, Multi-CNN, and ViT methods, our proposed method (*ViT-SPT-LSA*), *Vanilla Vision Transformer with enhancements*, achieved the most outstanding overall accuracy of **98.33%**. These results reveal valuable insights into the various classification tasks for which transformer-based systems can be used.

Our plans involve testing the models on various datasets to address sampling bias and offer more insights into the behavior of the models. Examining the use of transfer learning strategies, such as pre-trained models on sizable audio datasets, could promote model convergence and classification performance and be another enhancement in the future. Combining multi-modal data for cry categorization could be the subject of additional research. Apart from audio, photos that depict babies' facial expressions throughout the crying phase can be supplied to classification models to enhance their performance and accuracy even more.

Author Contributions: Conceptualization, D.S.; Methodology, S.A.Y., D.S. and N.S.T.; Software, S.A.Y.; Validation, N.S.T.; Formal analysis, D.S.; Investigation, N.S.T.; Writing – original draft, S.A.Y., D.S. and N.S.T.; Supervision, D.S. and N.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: For the cry classification task, the authors did not collect new data but used the existing publicly collected dataset donate–a–cry (https://github.com/gveres/donateacry-corpus (accessed on 1 June 2024)). The code used for augmenting, training, testing, and preprocessing-related scripts can be accessed at the following Github repository (https://github.com/SamirElgehiny/baby-cry-analysis (accessed on 1 June 2024)). Any additional information required or questions regarding the data or code are available from the lead contact upon request.

Conflicts of Interest: The authors assure that there are no conflicts of interest.

References

1. Islam, S.R.; Kwak, D.; Kabir, M.H.; Hossain, M.; Kwak, K.S. The internet of things for health care: A comprehensive survey. *IEEE Access* **2015**, *3*, 678–708. [CrossRef]
2. Rock, L.Y.; Tajudeen, F.P.; Chung, Y.W. Usage and impact of the internet-of-things-based smart home technology: A quality-of-life perspective. *Univers. Access Inf. Soc.* **2024**, *23*, 345–364. [CrossRef] [PubMed]
3. Kamruzzaman, M.; Alanazi, S.; Alruwaili, M.; Alshammari, N.; Elaiwat, S.; Abu-Zanona, M.; Innab, N.; Mohammad Elzaghmouri, B.; Ahmed Alanazi, B. AI-and IoT-assisted sustainable education systems during pandemics, such as COVID-19, for smart cities. *Sustainability* **2023**, *15*, 8354. [CrossRef]
4. Perez, A.J.; Siddiqui, F.; Zeadally, S.; Lane, D. A review of IoT systems to enable independence for the elderly and disabled individuals. *Internet Things* **2023**, *21*, 100653. [CrossRef]
5. Subhan, F.; Mirza, A.; Su'ud, M.B.M.; Alam, M.M.; Nisar, S.; Habib, U.; Iqbal, M.Z. AI-enabled wearable medical internet of things in healthcare system: A survey. *Appl. Sci.* **2023**, *13*, 1394. [CrossRef]
6. Ruiz-Zafra, A.; Precioso, D.; Salvador, B.; Lubián-López, S.P.; Jiménez, J.; Benavente-Fernández, I.; Pigueiras, J.; Gómez-Ullate, D.; Gontard, L.C. NeoCam: An edge-cloud platform for non-invasive real-time monitoring in neonatal intensive care units. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2614–2624. [CrossRef] [PubMed]
7. Saraswathy, J.; Hariharan, M.; Yaacob, S.; Khairunizam, W. Automatic classification of infant cry: A review. In Proceedings of the 2012 International Conference on Biomedical Engineering (ICoBE), Penang, Malaysia, 27–28 February 2012; pp. 543–548.
8. Jeyaraman, S.; Muthusamy, H.; Khairunizam, W.; Jeyaraman, S.; Nadarajaw, T.; Yaacob, S.; Nisha, S. A review: Survey on automatic infant cry analysis and classification. *Health Technol.* **2018**, *8*, 391–404. [CrossRef]
9. Green, J.A.; Whitney, P.G.; Potegal, M. Screaming, yelling, whining, and crying: Categorical and intensity differences in vocal expressions of anger and sadness in children's tantrums. *Emotion* **2011**, *11*, 1124. [CrossRef] [PubMed]
10. Parga, J.J.; Lewin, S.; Lewis, J.; Montoya-Williams, D.; Alwan, A.; Shaul, B.; Han, C.; Bookheimer, S.Y.; Eyer, S.; Dapretto, M.; et al. Defining and distinguishing infant behavioral states using acoustic cry analysis: Is colic painful? *Pediatr. Res.* **2020**, *87*, 576–580. [CrossRef]
11. Ashwini, K.; Vincent, P.D.R.; Srinivasan, K.; Chang, C.Y. Deep learning assisted neonatal cry classification via support vector machine models. *Front. Public Health* **2021**, *9*, 670352.

12. Sujatha, K.; Nalinashini, G.; Ganesan, A.; Kalaivani, A.; Sethil, K.; Hari, R.; Bronson, F.A.X.; Bhaskar, K. Internet of medical things for abnormality detection in infants using mobile phone app with cry signal analysis. In *Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 169–191.
13. Ferretti, D.; Severini, M.; Principi, E.; Cenci, A.; Squartini, S. Infant cry detection in adverse acoustic environments by using deep neural networks. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 992–996.
14. Zamzmi, G.; Kasturi, R.; Goldgof, D.; Zhi, R.; Ashmeade, T.; Sun, Y. A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Rev. Biomed. Eng.* **2017**, *11*, 77–96. [CrossRef]
15. Dixit, A.A.; Dharwadkar, N.V. A Survey on detection of reasons behind infant cry using speech processing. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 190–194.
16. Xie, J.; Long, X.; Otte, R.A.; Shan, C. Convolutional neural networks for audio-based continuous infant cry monitoring at home. *IEEE Sens. J.* **2021**, *21*, 27710–27717. [CrossRef]
17. Dunstan, P. Calm the crying: Using the Dunstan baby language. *Avery* **2012**, *240*. Kindle Edition.
18. Reyes-Galaviz, O.F.; Cano-Ortiz, S.D.; Reyes-García, C.A. Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence, Atizapan De Zaragoza, Mexico, 27–31 October 2008; pp. 330–335.
19. DonateACry Corpus. Available online: https://github.com/gveres/donateacry-corpus (accessed on 25 July 2023).
20. Ji, C.; Mudiyanselage, T.B.; Gao, Y.; Pan, Y. A review of infant cry analysis and classification. *EURASIP J. Audio Speech Music. Process.* **2021**, *2021*, 1–17. [CrossRef]
21. Liu, L.; Li, W.; Wu, X.; Zhou, B.X. Infant cry language analysis and recognition: An experimental approach. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 778–788. [CrossRef]
22. Dewi, S.P.; Prasasti, A.L.; Irawan, B. The study of baby crying analysis using MFCC and LFCC in different classification methods. In Proceedings of the 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 16–18 July 2019; pp. 18–23.
23. Sharma, K.; Gupta, C.; Gupta, S. Infant weeping calls decoder using statistical feature extraction and gaussian mixture models. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–6.
24. Jiang, L.; Yi, Y.; Chen, D.; Tan, P.; Liu, X. A novel infant cry recognition system using auditory model-based robust feature and GMM-UBM. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e5405. [CrossRef]
25. Cha, J.; Bae, G. Deep learning based infant cry analysis utilizing computer vision. *Int. J. Appl. Eng. Res.* **2022**, *17*, 30–35.
26. Kulkarni, P.; Umarani, S.; Diwan, V.; Korde, V.; Rege, P.P. Child cry classification-an analysis of features and models. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2–4 April 2021; pp. 1–7.
27. Ozseven, T. Infant cry classification by using different deep neural network models and hand-crafted features. *Biomed. Signal Process. Control* **2023**, *83*, 104648. [CrossRef]
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: https://arxiv.org/abs/1706.03762 (accessed on 20 May 2024).
29. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.
30. Chen, X.; Wu, Y.; Wang, Z.; Liu, S.; Li, J. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 5904–5908.
31. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
32. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; pp. 3586–3589. [CrossRef]
33. He, Z.; Rakin, A.S.; Fan, D. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 588–597. [CrossRef]
34. Suhas, B.; Mallela, J.; Illa, A.; Yamini, B.; Atchayaram, N.; Yadav, R.; Gope, D.; Ghosh, P.K. Speech task based automatic classification of ALS and Parkinson's Disease and their severity using log Mel spectrograms. In Proceedings of the 2020 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 19–24 July 2020; pp. 1–5.
35. Zhang, T.; Feng, G.; Liang, J.; An, T. Acoustic scene classification based on Mel spectrogram decomposition and model merging. *Appl. Acoust.* **2021**, *182*, 108258. [CrossRef]
36. Nguyen, M.T.; Lin, W.W.; Huang, J.H. Heart Sound Classification Using Deep Learning Techniques Based on Log-mel Spectrogram. *Circuits Syst. Signal Process.* **2023**, *42*, 344–360. [CrossRef]
37. Dörfler, M.; Bammer, R.; Grill, T. Inside the spectrogram: Convolutional Neural Networks in audio processing. In Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA), Tallin, Estonia, 3–7 July 2017; pp. 152–155. [CrossRef]

38. Khoria, K.; Patil, A.T.; Patil, H.A. Significance of Constant-Q transform for voice liveness detection. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 126–130.
39. Khoria, K.; Patil, A.T.; Patil, H.A. On significance of constant-Q transform for pop noise detection. *Comput. Speech Lang.* **2023**, *77*, 101421. [CrossRef]
40. Leitner, B.Z.J.; Thornton, S. *Audio Recognition using Mel Spectrograms and Convolution Neural Networks*; Noiselab University of California: San Diego, CA, USA, 2019.
41. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding robustness of transformers for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10231–10241.
42. Zhang, Q.; Xu, Y.; Zhang, J.; Tao, D. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *Int. J. Comput. Vis.* **2023**, *131*, 1141–1162. [CrossRef]
43. Lee, S.H.; Lee, S.; Song, B.C. Vision Transformer for Small-Size Datasets. *arXiv* **2021**, arXiv:2112.13492.