Kuwait Journal of Science 51 (2024) 100221

Contents lists available at ScienceDirect

Kuwait Journal of Science

journal homepage: www.sciencedirect.com/journal/kuwait-journal-of-science

Full length article

Infant cry classification using an efficient graph structure and attention-based model

Xuesong Qiao [a], Siwen Jiao [b], Han Li [a], Gengyuan Liu [a], Xuan Gao [a], Zhanshan Li [a,*]

[a] College of Computer Science and Technology, Jilin University, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Neural network
Multi-head attention
Infant cry
Audio classification

ABSTRACT

Crying serves as the primary means through which infants communicate, presenting a significant challenge for new parents in understanding its underlying causes. This study aims to classify infant cries to ascertain the reasons behind their distress. In this paper, an efficient graph structure based on multi-dimensional hybrid features is proposed. Firstly, infant cries are processed to extract various speech features, such as spectrogram, mel-scaled spectrogram, MFCC, and others. These speech features are then combined across multiple dimensions to better utilize the information in the cries. Additionally, in order to better classify the efficient graph structure, a local-to-global convolutional neural network (AlgNet) based on convolutional neural networks and attention mechanisms is proposed. The experimental results demonstrate that the use of the efficient graph structure improved the accuracy by an average of 8.01% compared to using standalone speech features, and the AlgNet model achieved an average accuracy improvement of 5.62% compared to traditional deep learning models. Experiments were conducted using the Dunstan baby language, Donate a cry, and baby cry datasets with accuracy rates of 87.78%, 93.83%, and 93.14% respectively.

Introduction

The problem of caring for newborns has recently garnered more attention due to the decline in fertility rates (Alkema et al., 2011) and the increase in population aging. While adults use language to convey various types of information, infants' language abilities are not yet fully developed, and crying is their primary way of expressing needs before they learn to speak. According to Mukhopadhyay et al. (2013), infant cries can be used as a basis for classification. However, even after a short period of training, it is difficult for humans to achieve a high accuracy rate in recognizing cries. With the further development of artificial intelligence technology, using machine learning algorithms to determine the classification of infant needs is a feasible and efficient research direction with practical significance.

In the classification of infant cries, there are generally two directions (Ji et al., 2021). The first is based on the infant's crying to determine whether the infant is choking (Ting et al., 2022b; Saraswathy et al., 2012a), in pain (Felipe et al., 2019), or healthy, in order to assess the infant's physical health. This type of classification task usually involves two categories. The second is to classify the infant's cry to determine the reason for the crying, such as whether the infant is crying due to hunger, pain, or tiredness, in order to help parents better care for the infant. This type of classification task usually involves three or more categories. The method this paper propose focuses on addressing the second issue.

For the classification of infant cry sounds, the process can be divided into two main steps: preprocessing and classification (Fig. 1). Firstly, the data needs to undergo preprocessing, where different methods are applied to process the original audio signals, making them suitable for deep learning or machine learning models. This step is often very important, as demonstrated by Alaie et al. (2016) who processed the signals into mel frequency cepstral coefficients (MFCC), Franti et al. (2018) who processed them into spectrograms, and Ting et al. (2022a) who combined signals obtained from different processing methods. In our approach, the signals were processed into an efficient graph structure composed of spectrogram, mel-scaled spectrogram and MFCC. In the second step, A local-to-global convolutional neural network (AlgNet) based on convolutional neural networks and attention mechanisms was employed. AlgNet is capable of processing local and global features in infant cries through a local feature extraction network based on convolutional neural networks and a global feature extraction network based on attention mechanisms, thus achieving better classification of infant cries.

* Corresponding author.
E-mail address: lizs@jlu.edu.cn (Z. Li).

https://doi.org/10.1016/j.kjs.2024.100221
Received 8 October 2023; Received in revised form 21 February 2024; Accepted 24 March 2024
Available online 26 March 2024

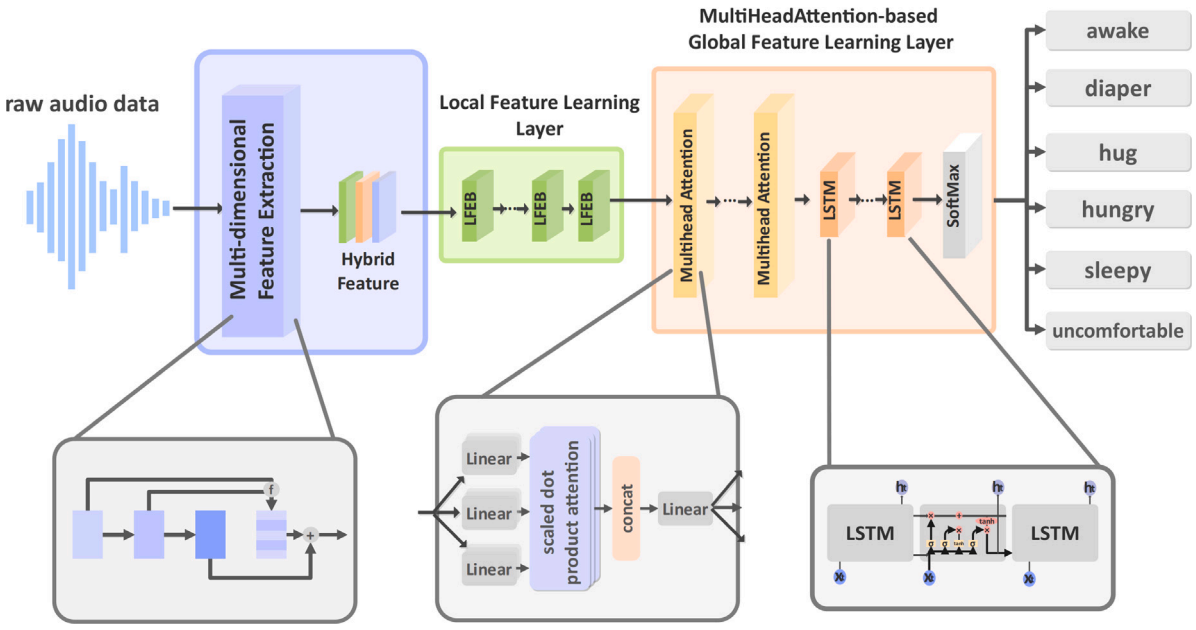

Fig. 1. Classifying infant cries according to their needs.

The contributions of this paper can be summarized as follows. Firstly, it proposes an efficient graph based on multi-dimensional hybrid features. Most existing research typically relies on using single features directly for training or simple combinations of multiple features within the same dimension. The proposed efficient graph structure can more fully utilize the correlations and complementary information among multiple features. Secondly, the paper proposes a local-to-global convolutional neural network (AlgNet). This network can better extract information from the efficient graph structure to achieve higher classification accuracy. Thirdly, the proposed methods are validated through experiments using four experimental groups on three public datasets. The experimental results demonstrate the higher accuracy of the proposed approach. Furthermore, ablation experiments confirm the contributions of the efficient graph structure and AlgNet to the improvement in accuracy.

Related work

The binary classification problem of infant cries into two categories has been relatively well-developed (Mukhopadhyay et al., 2013; Salehian Matikolaie et al., 2022). As early as 2010, Sahak et al. (2010) achieved a binary classification of infant cries into suffocation and non-suffocation using MFCC features and traditional machine learning methods, with an accuracy of over 95%. After that, various machine learning or neural network-based methods such as support vector machine (SVM) (Ozseven, 2023), genetic selection of a fuzzy model (GSFM) (Rosales-Pérez et al., 2015), AlexNet (Moharir et al., 2017), probabilistic neural network (PNN) (Saraswathy et al., 2012b) have been applied to the problem of binary classification of infant cries, including determining whether an infant is in pain, hungry, or choking.

Lahmiri et al. (2022) preprocessed the speech signal into cepstral coefficients and used deep feedforward neural networks (DFFNN), long short-term memory (LSTM), and convolutional neural networks (CNN) models for classification. The infant cries were categorized into healthy and unhealthy classes, achieving an accuracy of 95.31%. Ting et al. (2022a,b) preprocessed the speech signal into a hybrid feature by combining features such as MFCC, chromagram, mel-scaled spectrogram, spectral contrast, and tonnetz. Using DNN and CNN models, the infant cries were classified into choking and non-choking categories with an accuracy of over 99%. This is currently the highest accuracy rate for binary classification problems. However, there is still a gap in the field of problems with more than three classifications.

Currently, the mainstream datasets for infant cry classification include the Baby Chillanto database (Reyes-Galaviz et al., 2008), Donate a Cry Corpus (Veres, 2023), Dunstan Baby Language (Dunstan, 2012; Anon, 0000), COPE/iCOPE databases (Brahnam et al., 2006). The Baby Chillanto database and the COPE/iCOPE databasesmainly focus on a specific type of data, making them commonly used for binary classification of infant cries. On the other hand, the Donate a Cry Corpus and Dunstan Baby Language datasets are more balanced (Özseven, 2022), making them suitable for classification of three or more types of infant cries. Therefore, there are currently many studies that have achieved good results by using these two datasets for classification of three or more types of infant cries. In 2018, Frant attempted a five-class classification of infant cries using a CNN model with mel spectrogram features as the standard. The results were promising, achieving an accuracy of 89%. However, due to the use of binary cross-entropy as the loss function in the research process, it was considered unfair for a multi-class task. Therefore, based on the research and replication by Maghfira, Frant's method was re-evaluated using categorical cross-entropy as the loss function, resulting in an accuracy of 81.56%. Additionally, Maghfira replaced Frant's CNN model with a CRNN model, achieving an accuracy of 86.03%. This improvement of 5% in accuracy surpassed Frant's research and currently stands as the highest accuracy on the Dunstan Baby Language dataset. In 2019, Sharma conducted experiments using the Donate a Cry Corpus database. After organizing the original dataset, experiments were performed on five classification categories, resulting in an accuracy of 81.2%.

In Table 1, relevant studies that achieved good results using the Dunstan dataset and the donate dataset are listed. Additionally, some researchers in certain studies augmented the original public datasets with a substantial amount of their own collected data for training. In order to ensure fair comparison of the results, these studies were not included in the table. Apart from these datasets, due to the privacy of infant cry data, a large portion of research in this field is based on infant cry datasets recorded by researchers themselves, which cannot be publicly disclosed. As a result, fair comparisons with these studies cannot be made. Nevertheless, they still constitute extremely important work in this field, and the ideas and methods proposed in these studies have made significant contributions to the development of the field, representing milestones in its progress. Therefore, these works have still been included in Table 1.

Table 1
Summary of the works on baby cry classification with more than three categories.

First author	Dataset	Features	Classifiers	Best performance
Maghfira et al. (2020)	Dunstan Baby Language	Spectrogram	CNN-RNN	86.03
Franti et al. (2018)	Dunstan Baby Language	Spectrogram	CNN	81.56
Bano and RaviKumar (2015)	Dunstan Baby Language	MFCC, Pitch, ENG	k-NN	85.00
Bănică et al. (2016)	Dunstan Baby Language	MFCC	GMM	81.80
Sharma et al. (2019)	Donate a Cry	Frequency, Entropy, Spectral	GMM	81.27
Rosen et al. (2021)	Donate a Cry	MFCC	SVM	88.00
Messaoud and Tadj (2010)	Self-Recorded	MFCC	PNN	71.40
Abou-Abbas et al. (2015)	Self-Recorded	MFCC	HMM	83.79
Alaie et al. (2016)	Self-Recorded	MFCC	GMM	81.27
Ashwini et al. (2021)	Self-Recorded	Spectrogram	CNN+SVM	88.89
Liang et al. (2022)	Self-Recorded	MFCC	ANN, CNN, LSTM	95
Satapathy et al. (2021)	Self-Recorded	peak, pitch, MFCCs, ΔMFCCs, LPCC	GSVM	91
Alaie et al. (2016)	Self-Recorded	Spectrogram	MHSE	93.70

Materials and methods

Extraction of the efficient graph structure

Unlike image, speech data often cannot be directly input into a model for training. It usually has few feature changes in the long time domain. Additionally, at different sampling rates, the amount of data is often excessive and redundant. To address this, handcrafted features like MFCC and LPCC are often used, but these features may discard some information during extraction. According to Ting et al. (2022b)'s research, simply splicing these features can yield better results, but it does not take into account their similarities and complementarity. To overcome this, a novel efficient graph structure based on multidimensional spectrograms is proposed. This graph structure effectively eliminates redundant information that exists among multiple spectrograms while preserving their complementarity. The approximate process of producing such graph structures is illustrated in the Fig. 2.

Spectrogram preparation by short-time fourier transform

Spectrogram is a visual representation of the spectrum of audio frequencies over time, showing how the frequency content of a given audio signal changes over time. It is obtained by performing a short-time fourier transform (STFT) to convert the original time-domain signal into the frequency-domain signal. The spectrogram provides a display of the spectral distribution of the sound signal. The x-axis of the spectrogram represents time, the y-axis represents frequency, and the color or grayscale level of the image represents the magnitude of the spectral energy density. Typically, low-frequency sounds are displayed at the bottom of the spectrogram, while high-frequency sounds are displayed at the top. Compared to the original audio information, the spectrogram is more suitable for direct input into deep learning networks, as it preserves both the time and frequency information of the sound signal. First, the signal is divided into short-time frames through frame segmentation. This is because speech signals are non-stationary, and performing Fourier transform directly on the entire signal would be meaningless as it would lose the frequency contour of the signal over time. However, the signal can be considered relatively stationary within each frame by dividing it into short-time frames. During this process, a window function is applied to constrain the temporal length of the signal, suppressing edge effects and preventing the time boundaries from affecting the signal, thus improving signal smoothness. In this paper, the Hann window was utilized for windowing. Its formula is as follows:

$$w(n) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi n}{N-1}\right)$$

Where w(n) represents the nth value of the window function, n denotes the index of the nth value of the window function, N represents the length of the window function, and π is the mathematical constant pi. The Hann window is a window function with gradual tapering, which can suppress edge effects and improve computational accuracy. The Hann window is commonly used in signal filtering and signal analysis. After performing frame segmentation and windowing, the short-time Fourier transform STFT is applied to each frame, transforming the signal from the time domain to the frequency domain. The formula for the STFT is as follows:

$$\mathrm{STFT}\,(x\,(t))\,(\tau\omega) = \int x\,(t)\cdot w\,(t-\tau)\cdot e^{-j\omega t}\mathrm{d}t$$

where x(t) is the input signal, w(t) is the window function, τ is the time offset (translation parameter), ω is the frequency (angular frequency), j is the imaginary unit.

Preparation of mel spectrogram by mel scale

Mel-spectrogram is a spectral signal obtained by further transforming the spectrogram onto the mel scale. After performing short-time fourier transform (STFT), mel filters are applied to the signal. This is because the human cochlea perceives sound frequency in a nonlinear manner, and humans are generally less sensitive to the high-frequency components of speech. The mel filterbank, derived from psychoacoustic experiments, simulates the cochlear response to different frequency bands, thereby filtering the signal. Consequently, the resulting features represent a set of auditory characteristics that mimic the human ear's perception, known as the mel spectrogram. The mel scale is a frequency scale for sound signals that represents frequency based on the characteristics of the human auditory system. The fundamental idea behind the mel scale is that human perception of different frequencies varies, with lower frequencies being more easily perceived by the human auditory system compared to higher frequencies. The formula for the mel scale is as follows, where MEL(f) represents the mel scale value of frequency f, and f represents frequency.

$$MEL(f) = 2595log\left[10(1+\frac{f}{700})\right]$$

Preparation of mel-frequency cepstral coefficients by DCT

Mel-frequency cepstral coefficients (MFCC) is another widely used feature in the field of speech (Rosales-Pérez et al., 2015), which is based on the auditory characteristics of the human ear. This feature is extensively applied in various domains of speech recognition and has also achieved remarkable results in the field of infant cry analysis. After obtaining the mel-spectrogram, MFCC calculates the discrete cosine transform (DCT) of the spectrogram. This process extracts the main envelope information from the speech signal and discards the remaining information. The formula for DCT is as follows, where C(k) represents the kth coefficient after DCT transformation, f(n) represents the nth element of the original sequence, and N represents the sequence length.

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \quad (k = 0, 1, \ldots, N-1)$$

In this equation, x_n represents the logarithmic magnitude of the mel spectrogram, X_k represents the MFCC coefficients, and N represents the number of mel filter banks.

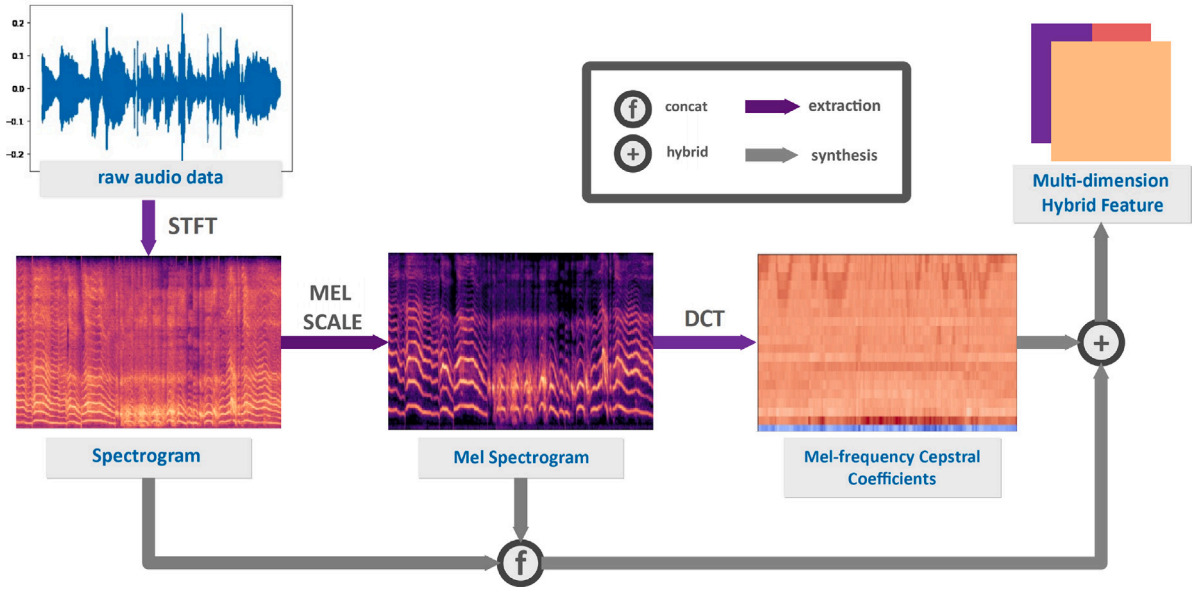

Fig. 2. Extraction of the efficient graph structure.

Creating multi-dimensional hybrid features

Finally, the redundant parts of similarity between the spectrogram and the mel spectrogram were removed, and they were merged in the same dimensions. Then, they were combined with MFCC in different dimensions to create a multi-dimensional hybrid feature.

During the experiments, various features of infant cry were extracted, such as spectrogram, mel spectrogram, MFCC, linear frequency cepstral coefficients (LFCC), bark frequency cepstral coefficients (BFCC), cochleagram, gammatone frequency cepstral coefficients (GFCC), using different methods to extracted (Shaikh et al., 2023) and combined them. Based on previous research (Sharan et al., 2021; Trapanotto et al., 2022) results and our own experimental results, it has been observed in experiments that for infant cry, the combined features often perform better than single features, and when using a single feature, MFCC is usually one of the best features. Therefore, among the features extracted in our experiments, using mel-frequency scaling for infant cry signals often yields the best results. According to experimental evidence, when MFCC features are combined with different features, the experimental results often improve compared to using MFCC alone. This may be because some judgment information is inevitably lost during the extraction of MFCC features.

In summary, this efficient graph structure is proposed, which consists of spectrogram, mel spectrogram, and MFCC. Both the spectrogram and the mel spectrogram are two intermediate features generated during the production of MFCC. Combining these two with MFCC not only retains the excellent effect of mel-frequency scaling for infant cry classification but also complements MFCC. In contrast to simple feature concatenation, combinations of different dimensions were utilized. This approach can supplement MFCC while ensuring that MFCC still plays a dominant role in classification.

Local feature learning network

The present study designed a new local feature extraction block (LFEB) to further improve the traditional CNN structure (Fig. 3). This module employs the ReLU activation function and consists of a convolutional layer, a Dropout layer, a BN layer, and a pooling layer. The network is composed of four such modules.

The convolutional layer is the core layer of this local feature learning module. The biggest feature of the convolutional layer is its local connectivity and weight sharing, which makes it possible to extract local features and greatly reduces the difficulty and complexity of

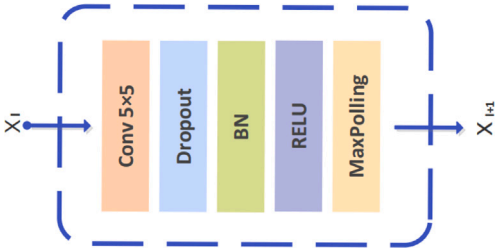

Fig. 3. Structure of the local feature extraction block.

network training. A dropout layer with a weight of 0.3 was designed after the convolutional layer. The batch normalization (BN) layer keeps the input distribution of each layer of the neural network similar during training, greatly improving the performance and stability. The pooling layer further reduces the number of parameters and computational complexity of the model while further preventing overfitting and improving generalization.

Global feature learning network

For the global feature learning network, two layers of multi-head attention, four LSTM layers, and one Softmax layer have been set up. The multi-head attention layer serves as the core to learn global dependency information, with the LSTM serving as a supplement and the Softmax layer used to ultimately set the result as a 6-class classification.

The MultiHeadAttention layer is an important component of the model and can help learn global dependencies in sequence data. The inspiration for this layer comes from the self-attention mechanism proposed in 'Attention Is All You Need' (Vaswani et al., 2017). The basic idea of self-attention is to compute a weighted sum of the input sequence, where the weights are determined by a learned attention function that considers the relationships between all positions in the sequence. This allows the model to focus on different parts of the sequence based on the importance of each position to the task. MultiHeadAttention is an extension of the self-attention mechanism that allows multiple attention heads to work in parallel. Mathematically,

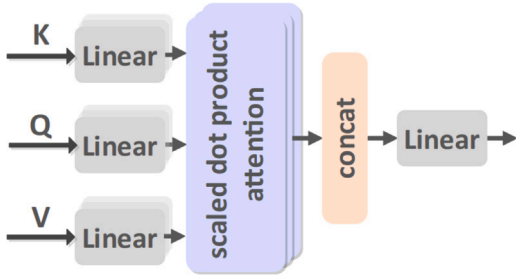

Fig. 4. MultiHeadAttention block.

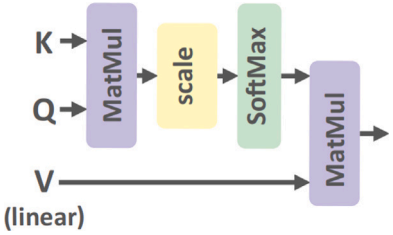

Fig. 5. Scaled dot-product attention layer.

the MultiHeadAttention layer can be defined as:

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,\text{head}_2,\dots,\\ &\quad \text{head}_h)W^O\\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V),\\ i &= 1,\dots,h. \end{aligned} \tag{1}$$

Where Q, K and V are query, key and value matrices with dimensions d_q, dk and d_v respectively. Specifically, in the MultiHeadAttention layer, we perform linear transformations on the query, key, and value matrices, respectively, and then split them into multiple heads, each with its own learnable weights and biases. Next, each head calculates the attention weights separately, multiplies the weights with the value matrix and adds them up, and finally concatenates the outputs of each head to form the final output tensor.

As in Fig. 4, the algorithm mainly consists of the following steps: (1)Perform linear transformations on the input Q, K, and V to obtain Q', K', and V'. (2)Split Q', K', and V' into num_heads heads. (3)Perform the scaled dot-product attention operation on each head to obtain the attention output. (4)Concatenate the attention outputs of multiple heads. (5)Perform a linear transformation on the concatenated attention output to obtain the final output.

Scaled dot-product attention is a type of attention mechanism used in the MultiHeadAttention layer. It calculates the weighted sum of values V given queries Q and keys K. As in Fig. 5, here is the structure and algorithmic steps of scaled dot-product attention: (1) Calculate the attention weights W using the following formula:

$$W = softmax(\frac{QK^T}{\sqrt{d_k}})$$

Where d_k is the dimension of the key matrix K, $softmax$ is a standard normalization function that ensures the sum of the weights W is 1. QK^T represents the product of the query matrix Q and the key matrix K, where T denotes transpose.(2) Calculate the weighted sum of the query matrix Q and the value matrix V as the output A.

The main advantage of the MultiHeadAttention layer is its ability to capture long-range dependencies in the input sequence. By attending to different parts of the sequence in parallel, the layer can identify important relationships between distant positions that may be overlooked by a single attention function.

Parameter setting

For the local feature learning network, the number of neurons in the convolutional layers of the stacked local feature learning modules was set to 32, 64, and 128, respectively. The kernel size is set to 5 and the strides are set to 2. In addition, the weight of dropout is set to 0.3. For the global feature learning network, the parameters of the multi-head attention layer are set. The number of heads is set to 8 and the key dim is set to 16. Then four stacked LSTM layers are set behind it to further learn global dependencies. The number of neurons in the LSTM layer is set to 256, 128, 64, and 32 respectively. A dropout layer with a weight of 0.3 is added again to further prevent overfitting. The optimizer uses stochastic gradient descent and sets the learning rate to 8e−1. The loss function uses sparse categorical crossentropy. In order to evaluate the effectiveness of our AlgNet, a set of control CRNN models was designed for comparison. The control model, as a baseline, removes the multi-head attention layer and replaces the LSTM layers with four layers of RNN layers. The number of neurons in the RNN layers is set to 256, 128, 64, and 32 respectively.

Data

Currently, the mainstream datasets for infant cry classification include the Baby Chillanto Database (Reyes-Galaviz et al., 2008), Donate a Cry (Veres, 2023), Dunstan Baby Language (Dunstan, 2012; Anon, 0000), COPE /iCOPE databases (Brahnam et al., 2006). The Baby Chillanto Database and the COPE/iCOPE databasesmainly focus on a specific type of data, making them commonly used for binary classification of infant cries. On the other hand, the Donate a Cry Corpus and Dunstan Baby Language datasets are more balanced, making them suitable for classification of three or more types of infant cries (Özseven, 2022). Therefore, in our research, these two datasets as well as the baby_crying database (hngynjy, 2023) were used.

Dunstan baby language dataset (Dunstan, 2012; Anon, 0000)

The Dunstan Baby Language dataset is derived from the research conducted by Priscilla Dunstan and her team. They discovered that regardless of language, culture, or race, all infants universally use five words during the first three months of pregnancy. These five universal words are: "neh" for hungry, "owh" for sleepy, "eh" for the infant wanting to burp, "eairh" for the infant experiencing stomach cramps due to trapped gas, and "heh" for the infant feeling discomfort internally. Originally, the Dunstan Baby Language was a video that included examples of the five types of infant crying recorded by Dunstan, along with explanations of their characteristics and tips on how to soothe the crying based on their types. The cries of all infants in the video were recorded in a studio to eliminate noise and resonance. All the babies are from the United States or Australia.

In 2021, Bratan et al. (2021) conducted testing on Dunstan baby cry. He collected cry sounds from 78 Romanian infants aged 0 to 1 month from two maternity hospitals in Bucharest, Romania. These cry sounds were then listened to and categorized by an analyst certified in the Dunstan method. Subsequently, a deep learning model was trained using the Dunstan Baby Language dataset. Finally, the model was used to classify a newly collected dataset of Romanian infant cry sounds, achieving an accuracy of over 80%. This demonstrated that baby cry sounds from different countries can be classified using the same method and model, and proved that a realistic tests protocol can be run.

To ensure fairness in the experimental comparison, the approach outlined by Maghfira et al. (2020) was followed, removing irrelevant content and only retaining the cries for classification. The audio samples have a sampling rate of 16,000 Hz, mono channel. The extracted infant cry audio consists of a total of 315 sounds, representing 56 "neh" cries, 106 "owh" cries, 55 "eh" cries, 61 "heh" cries, and 37 "eairh" cries. More information can be found at the URL: https://www.dunstanbaby.com/.

Table 2
The categories of infant cry.

	Hunger	Sleepy	Burping	Stomachache	Diaper change	Craving hug	Awake	Discomfort
Dunstan Baby Language	✓	✓	✓	✓				✓
Donate a Cry	✓			✓	✓			
Baby cry	✓	✓			✓	✓	✓	✓

Donate a cry dataset (Veres, 2023)

The Donate a Cry dataset is an open-source dataset aimed at helping people identify the needs of babies based on the type of their crying. The dataset is comprised of 1128 audio recordings, each containing 20 features and 9 labels (hunger, needs burping, stomachache, discomfort, tired, lonely, cold/hot, scared, unknown), and also includes the age and gender of the baby. The age of the babies in the dataset ranges from 0 to 2 years, and they come from various countries around the world. The dataset is collected entirely through user uploads using the 'Donate-a-cry' mobile application for IOS or Android. The audio files should contain samples of babies crying, with the corresponding label information encoded in the file names. Due to the voluntary nature of user uploads, the dataset is not balanced across all categories. To ensure fairer comparison of experimental results, Rosen et al. (2021)'s method was used, using the more balanced portions of the dataset. Specifically, three categories of infant crying reasons were selected for the research: hunger, discomfort (need diaper change), and stomachache. More information can be found at the URL: https://github.com/gveres/donateacry-corpus.

Baby cry dataset (hngynjy, 2023)

In the study, the publicly available dataset called baby_crying was also utilized. This dataset consists of six categories of cries with different sampling rates: 16,000 Hz, 1,600 Hz, and 44,100 Hz, with 16,000 Hz being the most predominant. The data was uniformly resampled to 16,000 Hz for our study. The total number of samples in the dataset is 918, with 124 cries labeled as Hungry, indicating that the infant is crying due to hunger. There are 144 cries labeled as Sleepy, indicating that the infant is crying because they are sleepy. There are 160 cries labeled as uncomfortable, indicating that the infant is crying due to discomfort. Additionally, there are 134 cries labeled as diaper, indicating that the infant is crying because they need a diaper change. There are 160 cries labeled as hug, indicating that the infant is crying because they desire to be held. Lastly, there are 160 cries labeled as awake, indicating that the infant is crying upon waking up. The average duration of the cries is 20 s. More information can be found at the URL: https://aistudio.baidu.com/datasetdetail/84370.

The categories of infant cry sounds included in each dataset are recorded in Table 2.

Experimental design

In order to evaluate the performance of our method compared to other methods on different datasets, as well as the individual performance of our designed efficient graph structure and AlgNet model, four experimental configurations were designed. These configurations consisted of: AlgNet model with the efficient graph structure, AlgNet model with standalone MFCC features, a regular CRNN model with the efficient graph structure and a regular CRNN model with standalone MFCC features. Each of these four configurations was tested on each dataset, resulting in a total of twelve experiments. To ensure the reliability of the experimental results, ten-fold cross-validation was employed in each experiment. The dataset will be divided into approximately ten equally sized subsets. For each round of model evaluation, one of these subsets will be used as the testing set, while the other nine subsets will be used as the training set, ensuring that no data is simultaneously allocated to both the testing and training sets. The model will be trained using these nine training sets, and then evaluated on the reserved testing set. This process will be repeated, selecting a different validation set each time, until each subset has been used as the testing set. Finally, the evaluation results from each round will be averaged.

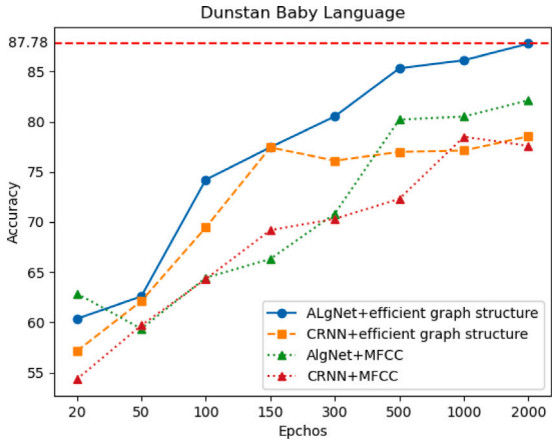

Fig. 6. Experimental results on the Dunstan Baby Language.

Results and discussion

Experimental results

Through experiments, it is observed that our method achieves satisfactory results on multiple datasets (Figs. 6 7 8). Among all the datasets, the combination of AlgNet and efficient graph structure consistently performs the best. The performance of the combination of regular CRNN and efficient graph structure, as well as AlgNet model with standalone MFCC features, fluctuates to some extent depending on the dataset, but it is consistently higher than the performance of regular CRNN with standalone MFCC features. These results are consistent with our preliminary experiments and expectations, which supports the effectiveness of our proposed method. In all experimental groups, the AlgNet+ efficient graph structure group using the donate a cry dataset achieved the highest accuracy, with a accuracy of 93.83%. This is also the highest accuracy reported in previous studies using this dataset. To facilitate the comparison of experiments, The epoch was standardized for all experiments. Therefore, it can be observed that some groups tended to overfit or even exhibit overfitting phenomena after a certain number of epochs. To prevent this phenomenon from interfering with the experimental results, a ten-fold cross-validation method was adopted, randomly dividing the dataset into ten subsets. In each experiment, one subset was used as the test set while the remaining nine subsets were used as the training set. The model was trained using the training set and evaluated on the test set. This process was repeated ten times, with each subset used as the test set once. Finally, the average value of the ten experiments was calculated.

Comparison with other methods

Through experiments, it was observed that our method outperforms the competing methods on all three datasets (Table 3). On the Dunstan Baby Language Dataset, our method achieves an accuracy of 87.78, which is a 1.75 improvement over the method proposed by Maghfira. On the donate a cry dataset, our method achieves an accuracy of 93.83, which is 12.56 higher than the method proposed by Sharma. Since Jiang's study only utilized a subset of the dataset and incorporated their own collected data during the experimental process, it was not included as a part of the experimental comparison. On the baby_crying dataset, an impressive accuracy of 93.14% was also achieved.

Table 3
Comparison with other methods.

Dataset	Study	Feature	Classifier	Accuracy
Dunstan Baby Language	Maghfira	Spectrogram	CNN-RNN	86.03
	Franti	Spectrogram	CNN	81.56
	Bano	MFCC, Pitch, ENG	k-NN	85.00
	Bănică	MFCC	GMM	81.80
	Our Experiment 1	Efficient graph structure	AlgNet	87.78
	Our Experiment 2	MFCC	AlgNet	82.13
	Our Experiment 3	Efficient graph structure	CRNN	78.54
	Our Experiment 4	MFCC	CRNN	77.59
Donate a cry	Sharma	Frequency, Entropy, Spectral	GMM	81.27
	Rosen	MFCC	SVM	88.00
	Our Experiment 1	Efficient graph structure	AlgNet	93.83
	Our Experiment 2	MFCC	AlgNet	85.13
	Our Experiment 3	Efficient graph structure	CRNN	88.31
	Our Experiment 4	MFCC	CRNN	67.41
Baby_crying	Our Experiment 1	Efficient graph structure	AlgNet	93.14
	Our Experiment 2	MFCC	AlgNet	89.04
	Our Experiment 3	Efficient graph structure	CRNN	91.04
	Our Experiment 4	MFCC	CRNN	81.58

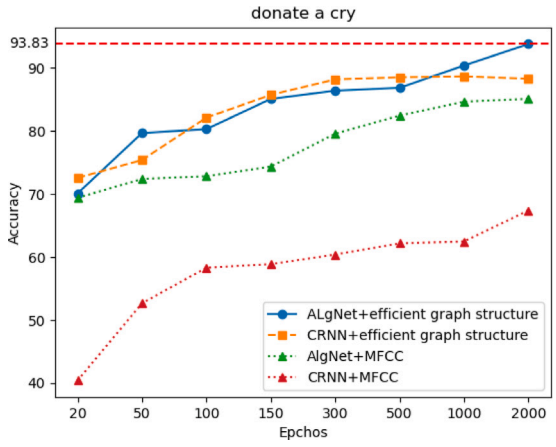

Fig. 7. Experimental results on the Donate a cry dataset.

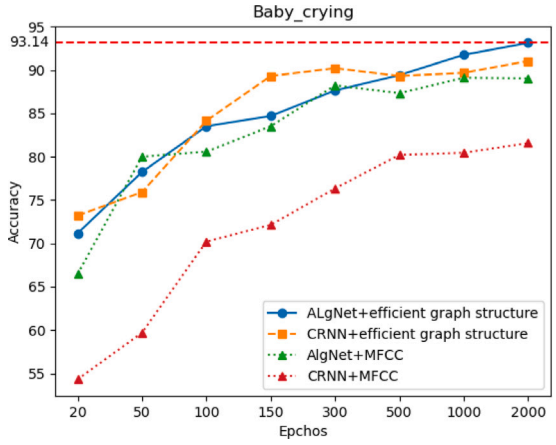

Fig. 8. Experimental results on the baby_crying dataset.

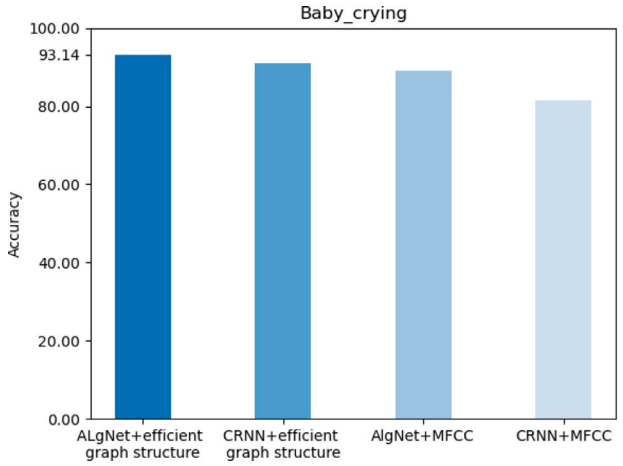

Fig. 9. Comparison of four experimental on the baby_crying dataset.

Comparison of four experimental groups

Through experiments, it is observed that our four experimental configurations yield different results on the three datasets (Table 4, and Figs. 9–11). Across all datasets, the combination of AlgNet and efficient graph structure consistently achieves the best results. Conversely, the combination of regular CRNN and standalone MFCC features consistently yields the least impressive performance.

In the Dunstan Baby Language Dataset, the results of the AlgNet model with standalone MFCC features are superior to those of the regular CRNN with efficient graph structure. However, on the donate a cry dataset and baby_crying dataset, the performance of the regular CRNN with efficient graph structure surpasses that of the AlgNet model with standalone MFCC features. This indicates that both modules play a role in the overall performance, and the effectiveness of this role varies depending on the dataset and data distribution.

By comparing the results of these four experimental configurations, it is evident that both the efficient graph structure and AlgNet have significantly positive impacts on the final classification accuracy. The extent of their impact, however, may vary depending on the distribution or characteristics of the data. In some datasets, the AlgNet model yields better performance, while in others, the efficient graph structure leads to higher improvements in the experimental results.

The comparison of different features

During the research process, in order to find features more suitable for infant cry classification, features such as spectrogram, mel spectrogram, MFCC, LFCC, BFCC, Cochleagram, and GFCC were extracted and various combinations were explored. Ultimately, due to the continuity and complementarity among MFCC, spectrogram, and mel

Table 4
Comparison of four experimental groups.

Group	Dataset		
	Dunstan baby language	donate a cry	Baby_crying
AlgNet model and the efficient graph structure	87.78	93.83	93.14
AlgNet model and MFCC features	82.13	85.13	89.04
CRNN model and the efficient graph structure	78.54	88.31	91.04
CRNN model and MFCC features	77.59	67.41	81.58

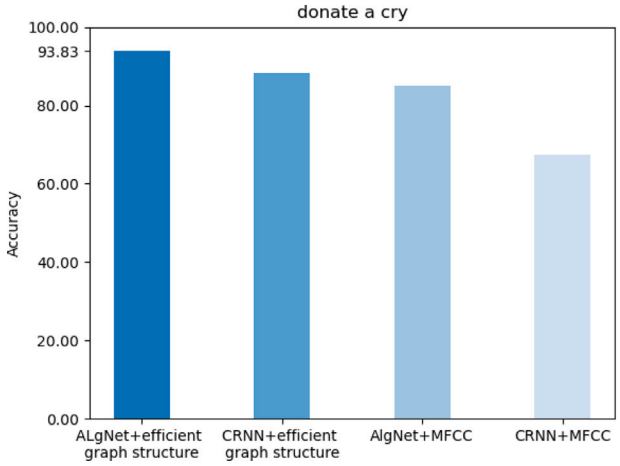

Fig. 10. Comparison of four experimental on the donate a cry dataset.

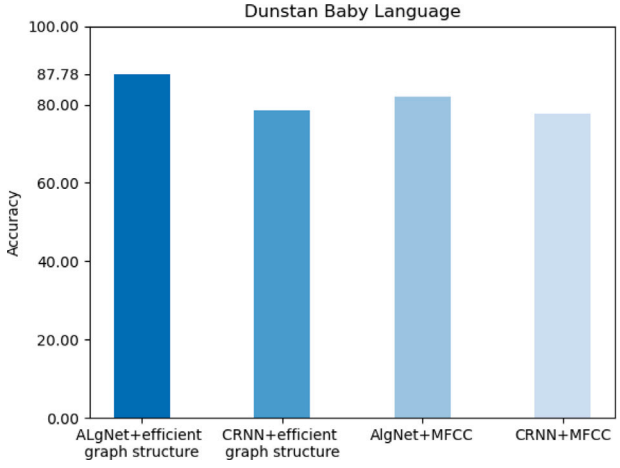

Fig. 11. Comparison of four experimental on the Dunstan baby language.

spectrogram, they were chosen to be combined into a more efficient graph structure. As the Dunstan dataset is the most balanced among the three datasets and is widely used in various studies, experiments were conducted on different features using this dataset, and recorded the five best-performing results in Table 5.

Conclusion

This paper presents an efficient classification method for classifying infant cries into three or more categories, aiming to infer the needs of infants based on their cries. To achieve this, an efficient graph structure and a local-to-global neural network based on convolutional neural networks and attention mechanisms are proposed, called AlgNet.

Table 5
The comparison of different features.

Method	Highest accuracy
the efficient graph structure	87.78%
MFCC and mel spectrogram	85.43%
MFCC	82.13%
Spectrogram and LFCC	74.53%
Mel spectrogram	72.13%
Spectrogram	63.11%

For the graph structure, which is based on multi-dimensional hybrid features. spectrogram, mel spectrograms and MFCC are generated through a series of related operations. Subsequently, we performed selective concatenation in different dimensions to eliminate redundant information caused by correlation between these features. Meanwhile, in order to preserve their mutually complementary information as much as possible during the synthesis process, a more efficient multi-dimensional graph structure is proposed. Ultimately, experimental results demonstrated that this efficient multi-dimensional spectrogram structure is more suitable for classifying infant cries compared to individual MFCC features.

In AlgNet, a convolution-based local feature learning network and a global feature learning network based on multi-head attention mechanisms is designed. The multi-head attention serves as the core of the model, leveraging its ability to learn global dependencies for the multi-dimensional hybrid features. The final experimental results demonstrate that the proposed network architecture is more suitable for classifying the resulting efficient graph structure. The combination of these two methods has achieved promising results on different datasets.

To validate the effectiveness of our proposed methods, comparative experiments and ablation experiments are conducted on multiple infant cry datasets. The approach achieves 87.78% accuracy on the Dunstan baby language dataset, 93.83% accuracy on the donate a cry dataset, and 93.14% accuracy on the baby_crying dataset. Furthermore, the use of the efficient graph structure improved the accuracy by an average of 8.01% compared to using standalone speech features, and the AlgNet model achieved an average accuracy improvement of 5.62% compared to traditional deep learning model. The experimental results demonstrate that the method proposed in this paper can effectively classify the needs of infants based on their cries.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abou-Abbas, L., Alaie, H.F., Tadj, C., 2015. Automatic detection of the expiratory and inspiratory phases in newborn cry signals. Biomed. Signal Process. Control 19, 35–43. http://dx.doi.org/10.1016/j.bspc.2015.03.007.

Alaie, H.F., Abou-Abbas, L., Tadj, C., 2016. Cry-based infant pathology classification using GMMs. Speech Commun. 77, 28–52. http://dx.doi.org/10.1016/j.specom.2015.12.001.

Alkema, L., Raftery, A.E., Gerland, P., Clark, S.J., Pelletier, F., Buettner, T., Heilig, G.K., 2011. Probabilistic projections of the total fertility rate for all countries. Demography 48 (3), 815–839. http://dx.doi.org/10.1007/s13524-011-0040-5.

Anon, 0000. Dunstan Baby Language | As seen on Oprah. URL https://www.dunstanbaby.com/.

Ashwini, K., Vincent, P.D.R., Srinivasan, K., Chang, C.Y., 2021. Deep learning assisted neonatal cry classification via support vector machine models. Front. Public Health 9, http://dx.doi.org/10.3389/fpubh.2021.670352.

Bănică, I.-A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C., 2016. Automatic methods for infant cry classification. In: 2016 International Conference on Communications. COMM, IEEE, pp. 51–54. http://dx.doi.org/10.1109/ICComm.2016.7528261.

Bano, S., RaviKumar, K., 2015. Decoding baby talk: A novel approach for normal infant cry signal classification. In: 2015 International Conference on Soft-Computing and Networks Security. ICSNS, IEEE, pp. 1–4. http://dx.doi.org/10.1109/ICSNS.2015.7292392.

Brahnam, S., Chuang, C.-F., Shih, F.Y., Slack, M.R., 2006. SVM classification of neonatal facial images of pain. In: Fuzzy Logic and Applications: 6th International Workshop, WILF 2005, Crema, Italy, September 15-17, 2005, Revised Selected Papers 6. Springer, pp. 121–128.

Bratan, C.A., Gheorghe, M., Ispas, I., Franti, E., Dascalu, M., Stoicescu, S.M., Rosca, I., Gherghiceanu, F., Dumitrache, D., Nastase, L., 2021. Dunstan baby language classification with CNN. In: 2021 International Conference on Speech Technology and Human-Computer Dialogue. SpeD, pp. 167–171. http://dx.doi.org/10.1109/SpeD53181.2021.9587374.

Dunstan, P.J., 2012. Calm the Crying: Using the Dunstan Baby Language, vol. 240, Penguin Books Ltd.

Felipe, G.Z., Aguiar, R.L., Costa, Y.M., Silla, C.N., Brahnam, S., Nanni, L., McMurtrey, S., 2019. Identification of infants' cry motivation using spectrograms. In: 2019 International Conference on Systems, Signals and Image Processing. IWSSIP, IEEE, pp. 181–186. http://dx.doi.org/10.1109/IWSSIP.2019.8787318.

Franti, E., Ispas, I., Dascalu, M., 2018. Testing the universal baby language hypothesis-automatic infant speech recognition with cnns. In: 2018 41st International Conference on Telecommunications and Signal Processing. TSP, IEEE, pp. 1–4. http://dx.doi.org/10.1109/TSP.2018.8441412.

hngynjy, 2023. The Baby_crying Database. URL https://aistudio.baidu.com/datasetdetail/84370.

Ji, C., Mudiyanselage, T.B., Gao, Y., Pan, Y., 2021. A review of infant cry analysis and classification. EURASIP J. Audio Speech Music Process. 2021 (1), 1–17. http://dx.doi.org/10.1186/s13636-021-00197-5.

Lahmiri, S., Tadj, C., Gargour, C., Bekiros, S., 2022. Deep learning systems for automatic diagnosis of infant cry signals. Chaos Solitons Fractals 154, 111700. http://dx.doi.org/10.1016/j.chaos.2021.111700, URL https://www.sciencedirect.com/science/article/pii/S0960077921010547.

Liang, Y.-C., Wijaya, I., Yang, M.-T., Juarez, J.R.C., Chang, H.-T., 2022. Deep learning for infant cry recognition. Int. J. Environ. Res. Public Health 19 (10), 6311. http://dx.doi.org/10.3390/ijerph19106311.

Maghfira, T.N., Basaruddin, T., Krisnadhi, A., 2020. Infant cry classification using cnn–rnn. In: Journal of Physics: Conference Series, vol. 1528, no. 1, IOP Publishing, 012019. http://dx.doi.org/10.1088/1742-6596/1528/1/012019.

Messaoud, A., Tadj, C., 2010. A cry-based babies identification system. In: Image and Signal Processing: 4th International Conference, ICISP 2010, Trois-RiviÈRes, QC, Canada, June 30-July 2, 2010. Proceedings 4. Springer, pp. 192–199.

Moharir, M., Sachin, M.U., Nagaraj, R., Samiksha, M., Rao, S., 2017. Identification of asphyxia in newborns using gpu for deep learning. In: 2017 2nd International Conference for Convergence in Technology. I2CT, pp. 236–239. http://dx.doi.org/10.1109/I2CT.2017.8226127.

Mukhopadhyay, J., Saha, B., Majumdar, B., Majumdar, A., Gorain, S., Arya, B.K., Bhattacharya, S.D., Singh, A., 2013. An evaluation of human perception for neonatal cry using a database of cry and underlying cause. In: 2013 Indian Conference on Medical Informatics and Telemedicine. ICMIT, IEEE, pp. 64–67. http://dx.doi.org/10.1109/IndianCMIT.2013.6529410.

Özseven, T., 2022. A review of infant cry recognition and classification based on computer-aided diagnoses. In: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications. HORA, pp. 1–11. http://dx.doi.org/10.1109/HORA55278.2022.9800038.

Ozseven, T., 2023. Infant cry classification by using different deep neural network models and hand-crafted features. Biomed. Signal Process. Control 83, 104648. http://dx.doi.org/10.1016/j.bspc.2023.104648, URL https://www.sciencedirect.com/science/article/pii/S1746809423000812.

Reyes-Galaviz, O.F., Cano-Ortiz, S.D., Reyes-García, C.A., 2008. Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In: 2008 Seventh Mexican International Conference on Artificial Intelligence. IEEE, pp. 330–335. http://dx.doi.org/10.1109/MICAI.2008.73.

Rosales-Pérez, A., Reyes-García, C.A., Gonzalez, J.A., Reyes-Galaviz, O.F., Escalante, H.J., Orlandi, S., 2015. Classifying infant cry patterns by the genetic selection of a fuzzy model. Biomed. Signal Process. Control 17, 38–46. http://dx.doi.org/10.1016/j.bspc.2014.10.002.

Rosen, R.J., Tagore, D., Iyer, T.J., Ruban, N., Raj, A.N.J., 2021. Infant mood prediction and emotion classification with different intelligent models. In: 2021 IEEE 18th India Council International Conference. INDICON, IEEE, pp. 1–6. http://dx.doi.org/10.1109/INDICON52576.2021.9691601.

Sahak, R., Mansor, W., Lee, Y., Yassin, A.M., Zabidi, A., 2010. Orthogonal least square based support vector machine for the classification of infant cry with asphyxia. In: 2010 3rd International Conference on Biomedical Engineering and Informatics, Vol. 3. IEEE, pp. 986–990. http://dx.doi.org/10.1109/BMEI.2010.5639300.

Salehian Matikolaie, F., Kheddache, Y., Tadj, C., 2022. Automated newborn cry diagnostic system using machine learning approach. Biomed. Signal Process. Control 73, 103434. http://dx.doi.org/10.1016/j.bspc.2021.103434, URL https://www.sciencedirect.com/science/article/pii/S1746809421010314.

Saraswathy, J., Hariharan, M., Vijean, V., Yaacob, S., Khairunizam, W., 2012a. Performance comparison of daubechies wavelet family in infant cry classification. In: 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications. IEEE, pp. 451–455. http://dx.doi.org/10.1109/CSPA.2012.6194767.

Saraswathy, J., Hariharan, M., Vijean, V., Yaacob, S., Khairunizam, W., 2012b. Performance comparison of daubechies wavelet family in infant cry classification. In: 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications. pp. 451–455. http://dx.doi.org/10.1109/CSPA.2012.6194767.

Satapathy, S., Chang, C.-Y., Bhattacharya, S., Raj Vincent, P.M.D., Lakshmanna, K., Srinivasan, K., 2021. An efficient classification of neonates cry using extreme gradient boosting-assisted grouped-support-vector network. J. Healthc. Eng. 2021, 7517313. http://dx.doi.org/10.1155/2021/7517313.

Shaikh, M.B., Chai, D., Islam, S.M.S., Akhtar, N., 2023. PyMAiVAR: An open-source python suite for audio-image representation in human action recognition. Softw. Impacts 17, 100544. http://dx.doi.org/10.1016/j.simpa.2023.100544, URL https://www.sciencedirect.com/science/article/pii/S2665963823000817.

Sharan, R.V., Xiong, H., Berkovsky, S., 2021. Benchmarking audio signal representation techniques for classification with convolutional neural networks. Sensors (Basel) 21 (10), 3434. http://dx.doi.org/10.3390/s21103434.

Sharma, K., Gupta, C., Gupta, S., 2019. Infant weeping calls decoder using statistical feature extraction and gaussian mixture models. In: 2019 10th International Conference on Computing, Communication and Networking Technologies. ICCCNT, IEEE, pp. 1–6. http://dx.doi.org/10.1109/ICCCNT45670.2019.8944527.

Ting, H.-N., Choo, Y.-M., Ahmad Kamar, A., 2022a. Classification of asphyxia infant cry using hybrid speech features and deep learning models. Expert Syst. Appl. 208, 118064. http://dx.doi.org/10.1016/j.eswa.2022.118064, URL https://www.sciencedirect.com/science/article/pii/S0957417422012696.

Ting, H.-N., Choo, Y.-M., Kamar, A.A., 2022b. Classification of asphyxia infant cry using hybrid speech features and deep learning models. Expert Syst. Appl. 208, 118064. http://dx.doi.org/10.1016/j.eswa.2022.118064.

Trapanotto, M., Nanni, L., Brahnam, S., Guo, X., 2022. Convolutional neural networks for the identification of African lions from individual vocalizations. J. Imaging 8 (4), 96. http://dx.doi.org/10.3390/jimaging8040096.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Veres, G., 2023. donateacry-corpus. URL https://github.com/gveres/donateacry-corpus.