2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP) | 979-8-3503-0245-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSP58490.2023.10248903

Analysis of Multiple Types of Baby Cries Based on LSTM

Weijie You
School of Mathematics and Big Data
Foshan University
Foshan, China

Xuncan Xiao
School of Mathematics and Big Data
Foshan University
Foshan, China

*Dongyang Tu
School of Mathematics and Big Data
Foshan University
Foshan, China
*tudongyang@gmail.com

Zhaoji Dai
School of Mathematics and Big Data
Foshan University
Foshan, China

Zixuan Huo
School of Mathematics and Big Data
Foshan University
Foshan, China

***Abstract*：The literature currently provides limited analyses of multiple emotions conveyed by infant cries and associated training sets. To address this gap, we collected six common infant cry speech types, which include awake, needing a diaper change, needing to be held, hungry, sleepy, and uncomfortable, from online sources and constructed them into a training set. LSTM (Long Short-Term Memory Neural Network) was utilized for the analysis of infant cry speech in this study. Voice features based on spectral and MFCC features were extracted and inputted into the LSTM network, supplemented by three fully connected layers to enhance generalization ability and robustness. Our results demonstrate that the LSTM model achieved a maximum accuracy of 92.39% in identifying multiple types of infant cries while plateauing after 75 rounds, indicating both efficiency and good performance. Additionally, the confidence probability of the validated speech obtained by the model was 92.9%, demonstrating good generalization ability and robustness based on actual speech data. Therefore, our developed LSTM model has good potential in classifying multiple types of emotions conveyed by infant cries with a high degree of accuracy and robustness.**

***Keywords*：LSTM; Sequential Model; Spectrum Feature; MFCC Feature; Baby Cry Recognition.**

I. INTRODUCTION

Babies are very important to the world. They are the future of human society and a vital force in the inheritance and development of human civilization. According to UNICEF, approximately 160 million babies are born each year worldwide, and their growth and development are essential to the stability and prosperity of society as a whole, so caring for babies is an essential task.

Infants can only transmit information through their cries, but for young parents who are inexperienced in caring for infants, they are not able to accurately identify the meaning of their infant's cry signals [1], and accurate recognition and analysis of infant cries can help parents to better care for and understand their children. In most studies nowadays, studies have focused on the identification of two to three of the cries of pathological [1], hungry [2][3], uncomfortable [2][3], sleepy [2][3], and painful [3] cries, while few studies have analyzed multiple classifications of infant cries. This is partly because there are few publicly available databases of relevant infant cries, and databases of infant cries are part of the critical research [1]. By understanding and responding to cries one can effectively navigate through the stages of infant crying [3].

With the development of computer hardware and algorithms, more and more new algorithms can be widely used in the recognition and analysis of infant cries. For example, in Mukhopadhyay's study, machine learning algorithms achieved 80.56% accuracy in recognizing four types of cries (pain, hunger, birth, and pleasure) [4]. Chunyan Ji also used various machine learning and deep learning methods, such as SVM, K-means, CNN, and KNN, to achieve 70%-94.97% accuracy using different training sets of infant cries, but the classification accuracy is still relatively low and limited due to the lack of a standard common training set [1].

In this study, we use convolutional neural networks (CNNs) as well as LSTM models to analyze speech features in infant cries, such as pitch, sound intensity, frequency, short time energy, and MFCC. By combining these speech features, we aim to more accurately infer the emotional categories behind the cries, thus improving the quality and effectiveness of infant care. This approach will enable us to take timely steps to care for and protect the health and well-being of infants.

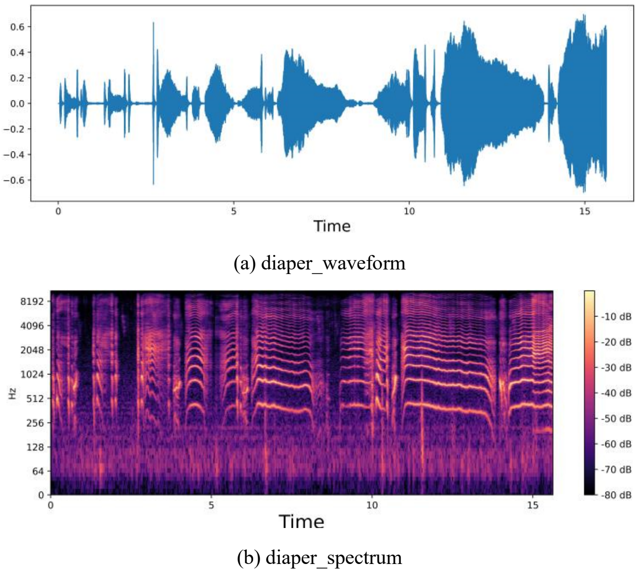
In previous studies, infant physiological variables such as facial expressions and sleep quality have been analyzed as parameters for studying infant needs [3][5]. This study combines the physiological mechanism of infant crying with a temporal model based on LSTM to more accurately and quickly analyze and classify six infant crying emotions (awake, diaper, cuddle, hungry, sleepy, and uncomfortable). This study not only provides better tools and methods for new parents or nannies but also presents new ideas and directions for research in the fields of speech signal processing and emotion analysis. With the continuous progress and development of science and technology, we believe that this method will be more widely used and promoted. The innovation of this study lies not only in its technical means but also in the practical value and impact it brings to human society.

II. Database Construction

The quality and accuracy of model training rely heavily on having a high-quality training set. For the study of infant cries, most of the available data is stored in hospitals or private databases. Some recent studies have used databases such as Baby Chillanto (2004), SPLANN (2015), Hypothyroid database(2009), and Donate A Cry (2015), among others [1]. In this study, the training sets were collected from two main sources: publicly available training sets online and training sets obtained from relevant technical forums using Python.

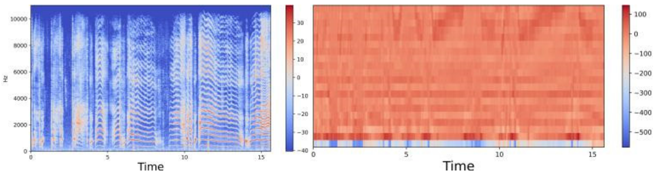
For this study, we chose the WAV format as the recording format for the training set to ensure high-quality audio. However, due to potential noise in the acquired audio that could interfere with recognition and classification, the audio needs to be pre-processed. To obtain a clear training set without noise interference, we carefully screened the audio data affected by noise.

The crying signal of infants is different from that of adults. In this study, we analyzed and extracted the crying signal of infants before training the model. In this study,we used the librosa library to analyze the signal for audio waveform plot, spectrum plot, acoustic spectrum plot, and Mel frequency cepstrum coefficient plot. Figure 1 represents crying indicating the need for a diaper change, while figure 2 displays crying indicating the need for a hug. The sound features reveal that different emotions in infants are reflected in different crying signals.



(a) diaper_waveform

(b) diaper_spectrum

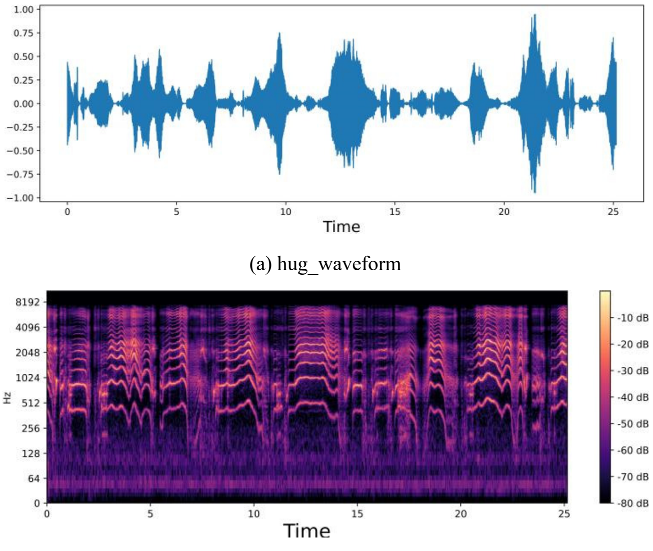


(c) diaper_acoustic spectrum (d) diaper_ Mel frequency cepstrum

Figure 1. the audio about waveform, spectrum, acoustic spectrum, and Mel frequency cepstrum of a crying infant when a diaper change is needed.

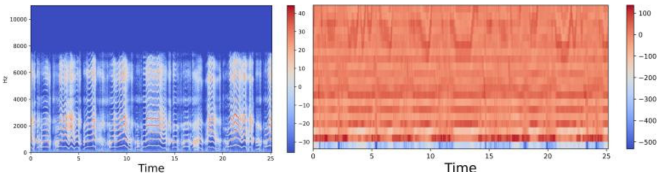
The crying signals indicating the need for a hug has a higher average decibel level during the first 5 seconds compared to the crying signals indicating the need for a diaper change. Moreover, the former exhibits 11 wave frequencies, whereas the latter displays 9.The frequency spectrum plot confirms that the speech indicating the need for a hug has more abrupt frequency features, particularly the prominent 2048Hz frequency component. Conversely, the audio indicating the need for a diaper change has a more noticeable frequency component of 1024Hz.Since infant speech has a higher pitch than adults, high-frequency components are crucial. The spectrogram shows that the high-frequency components of the speech indicating the need for a hug remain stable throughout the audio, while the high-frequency components of the speech indicating the need for a diaper change increase with time after 5 seconds. The MFCC plot reveals higher and more frequent high-frequency components in the speech indicating the need for a hug.

These differences in sound features signify that different infant emotions convey distinct information. By accurately recognizing and analyzing these sound features, timely responses to the emotional state of infants can be provided.



(a) hug_waveform

(b) hug_spectrum



(c) hug_acoustic spectrum (d) hug_ Mel frequency cepstrum

Figure 2. the audio about waveform, spectrum, acoustic spectrum, and Mel frequency cepstrum of a crying infant when a hug is needed.

Amulya A. Dixit's study concluded that Mel frequency cepstrum coefficients (MFCCs) are more effective for feature extraction [3][6-9]. Therefore, in this study, MFCCs were utilized to extract features from infant crying audio. Various types of infant crying sounds were analyzed, and a set of 40 MFCC coefficient features was obtained through calculation. Upon analyzing these MFCC coefficients, it can be observed that the crying sound is mainly concentrated in the frequency band ranging from 300 Hz to 400 Hz. The distribution of frequencies in the frequency domain is relatively even, and it can extend to the 500 Hz band around the third second of the crying sound.

According to the formula of MFCC:

$$C_i = \sum_{j=1}^{M} \log(\sum_{n=1}^{N} S_n(j) H_i(j)) \cos[\frac{\pi i}{M}(j - \frac{1}{2})] \quad (1)$$

In MFCC feature extraction, C_i represents the i^{th} cepstral coefficient which highlights the energy of the signal at Mel frequency scale; $S_n(j)$ represents the j^{th} frequency component of the n^{th} frame, corresponding to the magnitude of the signal in the frequency domain;$H_i(j)$ denotes the response of the i^{th} Mel filter on the j^{th} frequency component, indicating how much the frequency component contributes to the i^{th} filter's output.N refers to the total number of frames in the speech signal,while M denotes the dimensionality of cepstral coefficients.Our analysis of the crying sound revealed that the fundamental frequency is high at around 500 Hz. In addition, the MFCC coefficients indicated that the frequency and spectral changes in the infant's cry were more abrupt, suggesting a higher degree of emotional variability. These findings may indicate that the infant's crying reflects a sense of fragility, helplessness, and nervousness.

III. METHOD

In this study, both traditional convolutional neural networks (CNNs) and long short-term deep warp networks were utilized to perform learning analysis on the training and test sets.

The Long Short-Term Memory Neural Network (LSTM) is a type of recurrent neural network that combines memory units and gating mechanisms to effectively capture and utilize long-term contextual information. This makes LSTM well-suited for handling long-term dependencies and learning data features for speech extraction by establishing mapping relationships between speech features and acoustic features.

The Long Short-Term Memory Neural Network (LSTM) overcomes the problems of gradient disappearance and gradient explosion in long sequences by incorporating input gates, forgetting gates, and output gates. Each gate is controlled by a sigmoid function that regulates the flow of information in and out, while the dot product operation controls the strength of the gate. These gates are integrated with the cell states of the LSTM, which is the core element of the network and what distinguishes it from standard recurrent neural networks. The cell states are responsible for passing information, while the gates control the update and removal of the cell states. Together, these mechanisms enable LSTM to effectively handle long-term dependencies and learn features from speech data.

A. Input Gates

Input gates in the LSTM model determine the extent to which the current input should contribute to updating the cell state. The input gates are controlled by a sigmoid activation function and a dot product operation, which take the current input and previous cell states as inputs and produce a vector between 0 and 1. If the output of the input gate is close to 0, the current input is not updated for the cell state, while an output close to 1 indicates that the cell state will be fully updated. These vectors are multiplied by the current input to give a score of 0 or 1, thereby specifying which input items should be updated and which should not. The input gates can be represented mathematically by Eq:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

In this equation, x_t represents the input at the current time step, while h_{t-1} represents the output at the previous time step. $i_t$denotes the vector that controls the flow of information into the cell state, W_i andU_i and $b_i$represents the trained model parameters, and σ denotes the sigmoid activation function.

B. Forgetting Gates

The forgetting gate is another crucial element of the LSTM model, responsible for determining how much previous updates of the cell state should be retained or discarded. Like the input gate, the forgetting gate is controlled by a sigmoid activation function and a dot product operation. For each element, the forgetting gate produces a vector ranging from 0 to 1. A value close to 0 indicates that a small fraction of the previous cell state will be forgotten, while a value close to 1 means that no information will be forgotten. The equation for the forgetting gate can be represented as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3)$$

In this equation,the forget gate control vector is denoted by f_t, and controls the flow of information going out of the cell state.W_f and U_f and b_f represent the trained model parameters for the forget gate, and $\sigma$is the sigmoid activation function used to calculate the forget gate vector.

C. Output Gates

The output gate is the final component of the LSTM model, responsible for generating a vector of outputs based on the current cell state. Like the input and forgetting gates, the output gate is controlled by a sigmoid activation function and a dot product operation, taking the current cell state and previous inputs as input to produce a vector ranging from 0 to 1. This vector is multiplied by the current cell state to obtain an output for the current time step. The equation for the output gate can be represented as follows:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (4)$$

In this equation, o_t denotes the vector that controls the flow of information into the cell state,W_o and U_o and b_o represents the trained model parameters, σ denotes the sigmoid activation function.

The input gate, forgetting gate, and output gate in the LSTM model work collaboratively to update and control the flow of information within cell states. Specifically, the input gate determines how much of the present input should be added to the long-term memory of the LSTM network, while the forgetting gate determines how much of the previous state should be forgotten. The output gate, on the other hand, determines how much of the current state's information should be output. By regulating the inflow and outflow of information within cell states, these gates ensure that the network can remember relevant information while avoiding overload and confusion caused by irrelevant or redundant data. This mechanism helps the model achieve better robustness and generalization in handling long sequences.

In this study, we utilized a neural network architecture comprising a single LSTM layer with 64 neurons and four feedforward network layers, alongside an additional layer with 32 dimensions, to extract features from time series data. The extracted features were then processed by two layers with 16 dimensions and ReLU activation functions, followed by a fully connected layer with six dimensions and a softmax activation function for nonlinear transformation. To prevent overfitting during training, three dropout layers with a rate of 20% were added to randomly discard some neurons. Finally, the outputs were mapped to the output space for feature-enhanced classification of the training set. Regularization was also implemented by adding a regularization term as a combination of the square of L1 and L2 parameters to the loss function to reduce overfitting and improve the model's generalization performance. The following Figure 3 demonstrates the general architecture of the LSTM model.

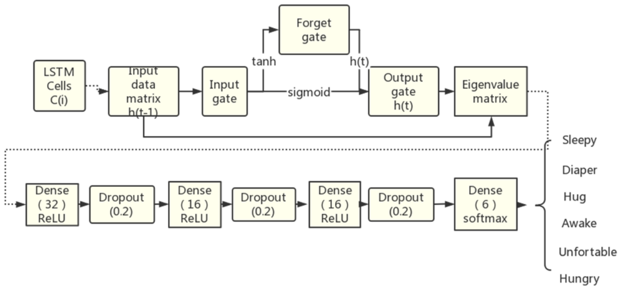


Figure 3. the experimental model architecture

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Introduction to this training environment: ①GPU: RTX1050 with 8 GB of video memory; ②CPU: Internet i5-9300H; ③RAM: 15.9 GB; ④Hard disk memory: 1 TB. The experiment is also equipped with a parallel computing platform CUDA v11.2, a deep learning framework PyTorch-GPU v1.10, and other related third-party dependencies for Python.

To implement the LSTM and fully connected layers, we used a sequential model with linear stacking. The input round of the LSTM model consisted of 40-dimensional MFCC features. During network training, we selected Adam as the optimizer for the model, setting the learning rate to 0.01 . The algorithmic flow of this process was as follows:

$$u_d w = \beta_1 u_d w + (1-\beta_1) dW \tag{5}$$

$$s_d = \beta_2 s_d w + (1-\beta_2) dW^2 \tag{6}$$

$$W = W - \alpha \frac{u_d W}{\sqrt{s_d W + \epsilon}} \tag{7}$$

In this formula, $u_d w$ and s_d and w represents respectively the momentum and squared gradient of the dW^{th} layer. The parameters β_1 and β_2 are hyperparameters, and W is the weight applied to the network. Furthermore, α represents the learning rate used in the training process, while ϵserves as a regularization term that prevents the denominator from becoming zero.

During training, we set the batch size to 113, which indicates the number of training samples input synchronously in each training. We also defined the number of iterations (Epoch) as 500. To assess the performance of the model, we used the Softmax cross-entropy loss function in our experiments. The algorithmic flow of this process was as follows:

$$\text{loss}(x, \text{cass}) = -x_{\text{class}} + \log \sum_{j=1}^{k} \exp(x_j) \tag{8}$$

Among them,loss represents the category to which the current sample belongs. This formula implements the operation of converting an input vector into a probability distribution by using an exponential function to transform raw scores into positive numbers, and then dividing them by the sum of all scores to normalize them into probabilities.

B. Experimental results:After performing feature extraction on the training set with both LSTM and CNN models, we observed from Figure 4 that the LSTM model tended to plateau after about 75 rounds of iterations for both the training and validation sets, achieving an accuracy of around 92%. In contrast, the CNN model continued to learn even after 200 rounds of iterations for the training set, while the validation set had stabilized after about 50 rounds. However, the accuracy of the CNN model was only around 30%. These results suggest that the LSTM model performed better, with a faster learning rate and stronger model generalization ability. The LSTM model captured the dependencies between multiple features through memory units, and utilized this information to make more accurate predictions. Moreover, it controlled the flow of feature data through gates to reduce the interference of noisy features, further improving the robustness of the model.

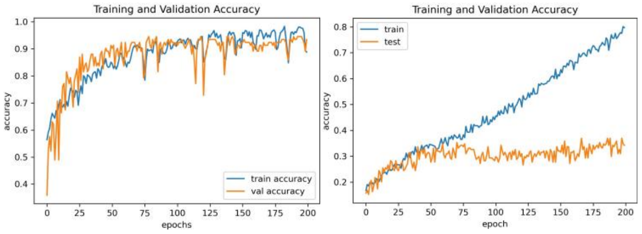


Figure 4. depicts the training set and validation set accuracies of both the LSTM model(on the left) and CNN model(on the right)

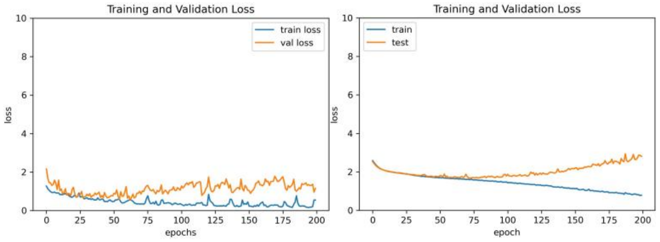
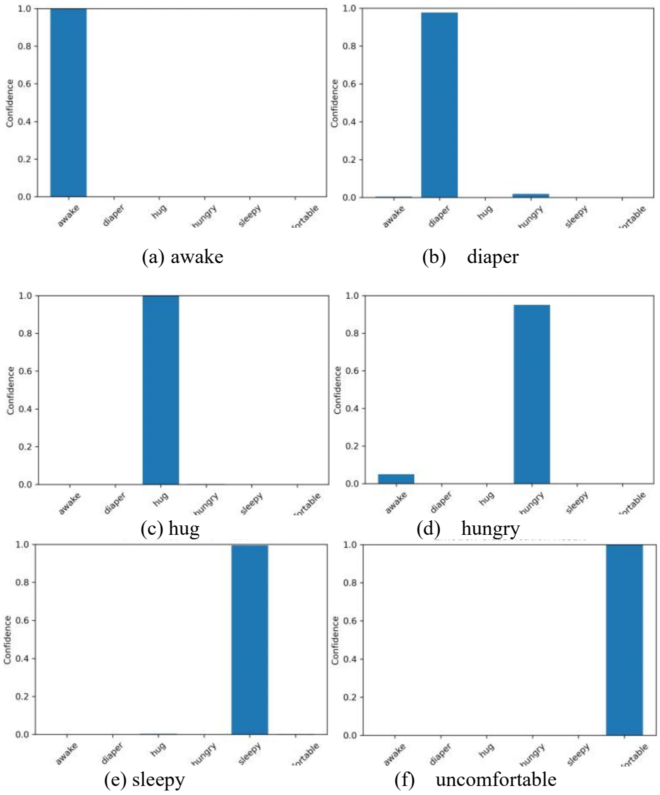


Figure 5. illustrates the loss of the training and validation sets for both the LSTM(on the left) and CNN models(on the right)

From Figure 4, it can be observed that the LSTM model achieved a maximum accuracy of 94.57% in the validation set, while the CNN model only achieved a maximum accuracy of 36.96% in the validation set. The difference between the maximum accuracies of the two models in the validation set was 57.61%. These results clearly demonstrate the superior performance of the LSTM model over the CNN model in terms of accuracy on this particular dataset.

Figure 5 shows the loss curves of both the LSTM and CNN models. We can observe that both models have a relatively fast convergence rate. However, the loss of the LSTM model is smaller than that of the CNN model. Furthermore, after 50 iterations, the loss of the CNN model starts to bounce back, indicating overfitting, and suggesting poor model generalization ability.



(a) awake (b) diaper

(c) hug (d) hungry

(e) sleepy (f) uncomfortable

Figure 6. displays the confidence analysis of the six classifications, awake, diaper, hug, hunger, sleepy, uncomformable

Figure 6 presents the confidence analysis of the six classifications using the LSTM model. The figure consists of 6 graphs, each representing one of the six sentiment classifications . These graphs display the confidence evaluation results of the corresponding classification. Based on our analysis of the LSTM model and the data, we observed that all 6 sentiment classifications have high confidence scores, indicating that the LSTM model is effective in accurately classifying the emotions conveyed by the audio signals.

V. CONCLUSION

In this study, we utilized the LSTM model to analyze and classify a dataset of baby cries obtained from multiple sources on the Internet. As a result, we achieved a maximum validation accuracy of 94.57%, which is 57.61% higher than the results obtained from the convolutional neural network. Furthermore, the training time for each round was only 1s and 12ms. Our findings suggest that the LSTM model has better performance in learning the features of baby cries as compared to the CNN model. Additionally, the LSTM model tends to plateau after 75 iterations, indicating high efficiency in terms of learning rate and model generalization ability.

In order to enhance the generalizability and robustness of the model, further research is needed where the model can be trained on a larger dataset with more classifications. This will enable the model to be better applied to real-life scenarios for the classification of baby cries, which could promote the development of the baby care industry

AUTHOR

*Corresponding author: Dongyang Tu is currently working at School of Mathematics and Big Data,Foshan University,and his degree is a master's degree in computer engineering(MCE). He mainly studies Big Data Systems and Algorithms.

(E-mail) tudongyang@gmail.com

FUND PROJECTS:

2022 Annual University Student Innovation and Entrepreneurship Training Program S202211847111; Foshan University 2022 Academic Fund Project xsjj202204kjb09; Jihua laboratory scienctific project, X210101UZ210

REFERENCES

[1] Ji, C., Mudiyanselage, T.B., Gao, Y. et al. A review of infant cry analysis and classification. J AUDIO SPEECH MUSIC PROC. 2021, 8 (2021). https://doi.org/10.1186/s13636-021-00197-5

[2] K. S. Alishamol, T. T. Fousiya, K. J. Babu, M. Sooryadas and L. Mary, "System for Infant Cry Emotion Recognition using DNN," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 867-872, doi: 10.1109/ICSSIT48917.2020.9214198.

[3] A. A. Dixit and N. V. Dharwadkar, "A Survey on Detection of Reasons Behind Infant Cry Using Speech Processing," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2018, pp. 190-194, doi: 10.1109/ICCSP.2018.8524517.

[4] J. Mukhopadhyay, B. Saha, B. Majumdar, A. K. Majumdar, S. Gorain, B. K. Arya, S. D. Bhattacharya, A. Singh, in 2013 Indian Conference on Medical Informatics and Telemedicine, ICMIT 2013. An evaluation of human perception for neonatal cry using a database of cry and underlyingcause, (2013). https://doi.org/10.1109/IndianCMIT.2013.6529410

[5] Y. Skogsdal, M. Eriksson, and J. Schollin, “Analgesia in newborns given oral glucose,” Acta Paediatrica, vol. 86, no. 2, pp. 217-220, 1997.

[6] R. J. Rosen, D. Tagore, T. J. Iyer, N. Ruban and A. N. J. Raj, "Infant Mood Prediction and Emotion Classification with Different Intelligent Models," 2021 IEEE 18th India Council International Conference (INDICON), Guwahati, India, 2021, pp. 1-6, doi: 10.1109/INDICON52576.2021.9691601.

[7] N. Meephiw and P. Leesutthipornchai, "MFCC Feature Selection for Infant Cry Classification," 2022 26th International Computer Science and Engineering Conference (ICSEC), Sakon Nakhon, Thailand, 2022, pp. 123-127, doi: 10.1109/ICSEC56337.2022.10049328.

[8] R. Cohen and Y. Lavner, "Infant cry analysis and detection," 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 2012, pp. 1-5, doi: 10.1109/EEEI.2012.6376996.

[9] S. Sharma and V. K. Mittal, "Infant cry analysis of cry signal segments towards identifying the cry-cause factors," TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, Malaysia, 2017, pp. 3105-3110, doi: 10.1109/TENCON.2017.8228395.