Baby Cry Sound Detection for Baby Health and Wellness Monitoring (Spectrogram,2DCNN/LSTM)

Shubhan Mehrotra
School of Electrical Engineering
Indian Institute of Technology Palakkad
Pudussery West, Palakkad, India, 678623
122101037@smail.iitpkd.ac.in

Dr. Sabarimalai Manikandan
School of Electrical Engineering
Indian Institute of Technology Palakkad
Pudussery West, Palakkad, India, 678623
msm@iitpkd.ac.in

***Abstract*—The aim of this paper is to investigate and classify baby crying sounds for health and wellness monitoring.In this paper we have classified the data into 5 classes which are 'bellypain', 'burping', 'discomfort', 'hungry', and 'tired. The prepoccessing for the data starts with augmenting the data with Gaussian noise, pitch shifting and frequency attenuation. We analyse the spectogram of the data and use its short time Fourier transformation whose amplitudes are used further. We utilize spectrogram a 2D Convolutional Neural Network (2D CNN) and Long short-term memory(LSTM) to extract its features and classify the cries with respect to these labels. Achieving a testing accuracy of 94.57% with 2D CNN and a testing accuracy of 91.47% with LSTM. Both the approach efficacy in real-time health monitoring of infants.**

***Index Terms*—Baby crying sounds, health monitoring, 2D Convolutional Neural Network, 2D CNN, Long short-term memory, LSTM**

I. Introduction

A. Significance of Project

The sound of a baby's cry is a universal language that communicates a range of needs and emotions. For parents and guardians, interpreting these cries accurately is critical looking after the well-being of infants. Regardless of all it becomes very difficult to understand the cause of a baby's cry. With this challenge in consideration in recent years with the advancements in technology we can automate baby sound cry detection systems which can help figure out the need of the baby. The application of baby cry sound detection is way more than a matter of convenience. Despite the numerous advances in healthcare a vast of amount of death in children happens amongst newborns which accounts for 47% of mortality of the total deaths of children under five years [1]. From a pediatric perspective, infant cry is the reflection of complex neurophysiological functions that can allow assessing a newborn's psychological and clinical status [2]. In the recent years a lot of studies have incorporated Machine Learning models to help healthcare professionals in early detection of several pathological conditions. Some of these studies have targeted deafness [3], autism [4], [5], [6] and other pathologies [7], [8]. Also in daycare where infant caregivers have a lot of babies to look over a detection model like this would help for the caregiver to prioritize their attention to ensure welfare of all infants. We will be looking at two machine learning models here for baby sound detection 2D CNN and LSTM. Both of these models are widely used for audio processing due to their effectiveness in learning hierarchical representations of the data along with keeping up with the variations in the data.

Identify applicable funding agency here. If none, delete this.

B. Related Work

In [12] ,the authors projected a novel approach to baby cry categorization by joining the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures accompanying cross-confirmation. The judgment of their procedure complicated equating allure accomplishment with a fundamental CNN-located approach utilizing the Dunstan Baby Language dataset [15].

In the world of visual and audio entertainment transmitted via radio waves signal acknowledgment,[13] have proved that voice conversion and turbulence adding cause upgraded acting. These improving methods embellish the model's talent to correctly categorize and change baby weeping patterns. Also, feature option methods are working to lighten the closeness of repetitious face that can prevent the model's discriminatory capacity in baby cry categorization tasks.

[14] et al. (2021) carried out multi-class classification using classes for pain, hunger, and tiredness on baby cry classification data. The STFT technique was largely used to transform the audio signal into a spectrum picture. Using CNN, the picture was fed for automatic feature extraction. To determine the final class, the collected characteristics were fed into the SVM classifier. Various SVM kernels were assessed. With 88.89% accuracy, the RBF-SVM produced the best results.

C. Motivation

By developing an mechanized method for baby cry sound detection, using deep learning models such as 2DCNN/LSTM (Convolutional Neural Network/Long Short-Term Memory) models, the project aims to determine a trustworthy and effective way for monitoring baby's health and wellness. This science can assist parents and caregivers in understanding the underlying cause of the baby crying.

D. Objective and Work Plan

In this paper we will preview audio signals using spectogram. We use deep neural networks , working with them as follows:

1) Pre-Processing of the data where we augment the data
2) Extracting features of the audio signals using its spectogram
3) Using 2 Dimensional Convolutional Neural Network and Long short-term memory for creating a model and checking their performances

II. Materials and Methods

A. System Architecture with Description

The whole code block consists of the following: Data Acquisition, Data Prepoccessing, Feature Extraction, Training block of 2DCNN/LSTM, Visualisation of the Results.

The preprocessing part of the data is the augmentation where we add noise to our data and later is connected to module of features extraction which then gets sent to training module.

The 2DCNN and LSTM architecture specifications and their parameters are summarized in the following tables:

TABLE I
Model Architecture of 2D CNN

Layer (type)	Output Shape	Param #
conv2d$_1$2$(Conv2D)$	(None, 126, 430, 32)	320
max$_p ooling2d_1$2$(MaxPooling2D)$	(None, 63, 215, 32)	0
conv2d$_1$3$(Conv2D)$	(None, 61, 213, 128)	36,992
max$_p ooling2d_1$3$(MaxPooling2D)$	(None, 30, 106, 128)	0
conv2d$_1$4$(Conv2D)$	(None, 28, 104, 128)	147,584
max$_p ooling2d_1$4$(MaxPooling2D)$	(None, 14, 52, 128)	0
conv2d$_1$5$(Conv2D)$	(None, 12, 50, 128)	147,584
max$_p ooling2d_1$5$(MaxPooling2D)$	(None, 6, 25, 128)	0
flatten$_3(Flatten)$	(None, 19200)	0
dense$_6(Dense)$	(None, 1024)	19,661,824
dense$_7(Dense)$	(None, 5)	5,125

TABLE II
Model Architecture for LSTM

Layer (type)	Output Shape	Param #
lstm$_4$9$(LSTM)$	(None, 1024, 128)	93,696
dropout$_3$4$(Dropout)$	(None, 1024, 128)	0
lstm$_5$0$(LSTM)$	(None, 128)	131,584
dropout$_3$5$(Dropout)$	(None, 128)	0
dense$_4$8$(Dense)$	(None, 64)	8,256
dense$_4$9$(Dense)$	(None, 5)	325

B. System Specification Table

C. Flowchart

Refer to Fig-[1]

TABLE III
System Specification

Component	Specification
Sampling Rate	22.05 kHz
Processor	Intel i7-10750H CPU 2.59 GHz
Memory	16 GB RAM
Storage	1024 GB
Operating System	Windows
Deep Learning Framework	Keras

D. Differnet Modules of Proposed Methods

1) Data Acquisition: We acquired a data set named dont_cry_corpus from [17]. The data is widely used and has 457 clean audio files, which we will be using for our purpose. The data is already categorised under this dataset into 5 labels which are 'bellypain', 'burping', 'discomfort', 'hungry', and 'tired.

2) Data Prepossessing: This method augments the clean data we have gotten from data acquisition. The augmentation is done by adding Gaussian Noise ,Pitch Shifting and Frequency attenuation. The augmentation rate is defined for all and the total dataset size is expanded to 1939 files.

3) Feature Extraction: Here we generate spectogram graphs for all the augmented audio files and also extract features for all of them. We calculate the short time fourier transformation of audio files and convert the audio files amplitudes into decibels and normalize it. We reshape this array to fit how we further process our data and store the values in a tuple corresponding to the labels , which is further down broken into two lists one with normalised stft matrix is stored in one list and corresponding label of that data into another

4) Training of Data: Trains the machine learning model using the stft matrix data and then. We train the data with two models one is 2D CNN and another is LSTM.The dimensions of input that does into these layers are differnet as 2D CNN provides more depth which LSTM does not.

5) Visualisation of Data: We generate plots of the training process plot loss curves, accuracies, and confusion matrices.These provide insights into the model's behavior and performance.

6) Evaluation Module: We evaluate the trained model's performance using methods such as accuracy, precision, recall, and F1-score. then we conduct testing on separate datasets to assess the model's generalization capability.

E. Mathematical Expressions Used

For most of our preprocessing we are using libraries so we are not directly using any mathemtical expressions which is true for the visualisation part as well. but generally the formulaes for calculating precision, recall, f1-score are as follows:

$$P = \frac{TP + FP}{TP} \tag{1}$$

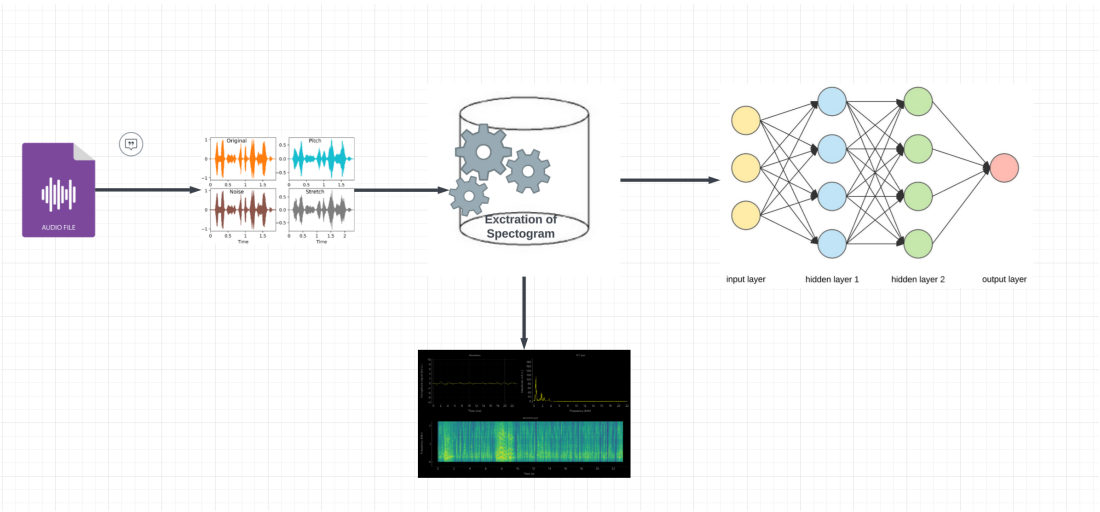

Fig. 1. Flowchart of the Model

where P is Precision with TP are True positives and FP are False Positives similarly,

$$R = \frac{TP + FN}{TP} \tag{2}$$

where R is recall also known as sensitivity and FN here is False negatives

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

The F1-score is the harmonic mean of precision and recall

Speceficity:

$$Specificity = \frac{TN + FP}{TN} \tag{4}$$

F. User Interface Related to Project Tasks

We created a very basic browsing window for the model we have created and we run it to load any of the audio files even the ones not in the dataset and we get accurate results where the model predicts one of the 5 labels on the audio. Refer to Fig.2 and Fig.3

G. Performance Matrix

TABLE IV
PERFORMANCE MATRIX FOR 2D CNN

Labels	Precision	Recall	F1-score	Specificity	Support
belly_pain	1.00	0.96	0.98	1.0	25
burping	1.00	1.00	1.00	1.0	28
discomfort	0.92	0.80	0.86	0.990	15
hungry	0.86	0.94	0.90	0.948	33
tired	0.96	0.96	0.96	0.990	28
accuracy			0.95		129
macro avg	0.95	0.93	0.94		129
weighted avg	0.95	0.95	0.95		129

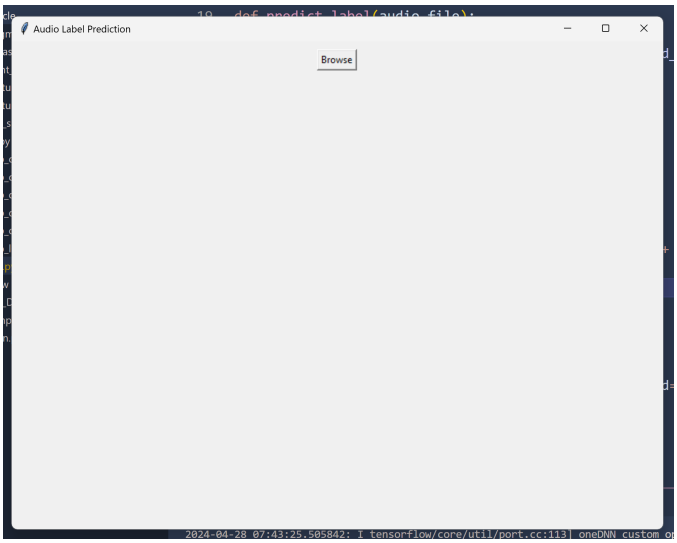

Fig. 2. User Interface Windowscreen

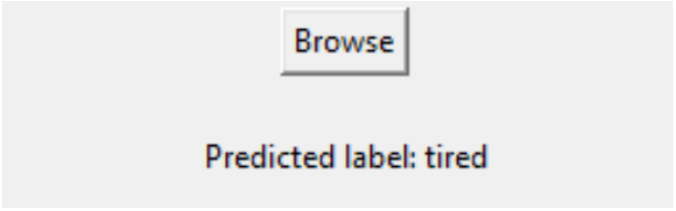

Fig. 3. Output when the file has been added

III. Results and Discussions

A. Experimental Setup and Data Base Collection

We got the Database from Ji et al.[17].This dataset was one of the most widely used datasets that had been properly labelled. Getting this data was one of the initial parts of the experimental setup then we move forward to the preprocessing

TABLE V
PERFORMANCE MATRIX FOR LSTM

Labels	Precision	Recall	F1-score	Specificity	Support
belly_pain	0.96	0.96	0.96	0.989	25
burping	1.00	0.96	0.98	1.0	28
discomfort	0.83	0.67	0.74	0.982	15
hungry	0.79	0.94	0.86	0.916	33
tired	1.00	0.93	0.96	1.0	28
accuracy			0.91		129
macro avg	0.92	0.89	0.90		129
weighted avg	0.92	0.91	0.91		129

parts where we Augment the data using the library "audiomentations" then the augmented audio files are saved on the local device to run further computations on it. After that we use "librosa" library to extract features and to plot the spectograms for this augmented audio we used "matplotllib".For training our model we use Keras for both 2D CNN and LSTM. Again visualisations of data using "matplotlib" and confusion matrix is plot using "seaborn". These are some of the major libraries we have used in our experimental setup.

B. Tables and Graphs

1) 2D CNN: Now for 2D CNN we have the following tables an graphs :

TABLE VI
PERFORMANCE MATRIX FOR 2D CNN

Labels	Precision	Recall	F1-score	Specificity	Support
belly_pain	1.00	0.96	0.98	1.0	25
burping	1.00	1.00	1.00	1.0	28
discomfort	0.92	0.80	0.86	0.990	15
hungry	0.86	0.94	0.90	0.948	33
tired	0.96	0.96	0.96	0.990	28
accuracy			0.95		129
macro avg	0.95	0.93	0.94		129
weighted avg	0.95	0.95	0.95		129

Testing Loss: 0.38239726424217224 Testing Accuracy: 0.9457364082336426

and accuracy: 0.9825 - loss: 0.0696 - val_accuracy: 0.9614 - val_loss: 0.2468 at the final epoch. Plot for Validation Accuracy and Training Accuracy is: Fig.4

Plot for Validation Loss and Training Loss is : Fig.5

Plotting Confusion Matrix: Fig.6

2) LSTM: Now for the second model which is LSTM we have the following table and graphs:

TABLE VII
PERFORMANCE MATRIX FOR LSTM

Labels	Precision	Recall	F1-score	Specificity	Support
belly_pain	0.96	0.96	0.96	0.989	25
burping	1.00	0.96	0.98	1.0	28
discomfort	0.83	0.67	0.74	0.982	15
hungry	0.79	0.94	0.86	0.916	33
tired	1.00	0.93	0.96	1.0	28
accuracy			0.91		129
macro avg	0.92	0.89	0.90		129
weighted avg	0.92	0.91	0.91		129

Fig. 4. Training and validation accuracy

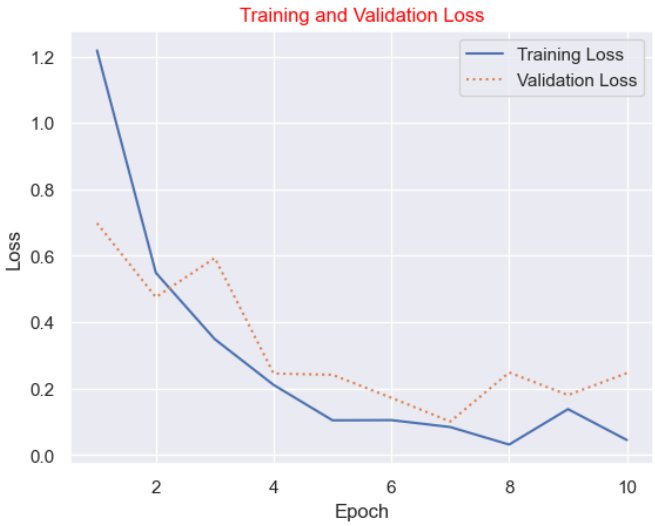

Fig. 5. Training and validation loss

Testing Loss: 0.34383639693260193 Testing Accuracy: 0.9147287011146545

and accuracy: 0.9271 - loss: 0.2463 - val_accuracy: 0.8958 - val_loss: 0.3082 at the final epoch Plot for Validation Accuracy and Training Accuracy is: Fig.7

Plot for Validation Loss and Training Loss is : Fig.8

Plotting Confusion Matrix: Fig.9

C. Result Comparison

1) LSTM Performance: :

- Testing Loss: 0.3438
- Testing Accuracy: 0.9147

Final Epoch Metrics:

- Accuracy: 0.9271
- Validation Accuracy: 0.8958
- Precision: 0.92

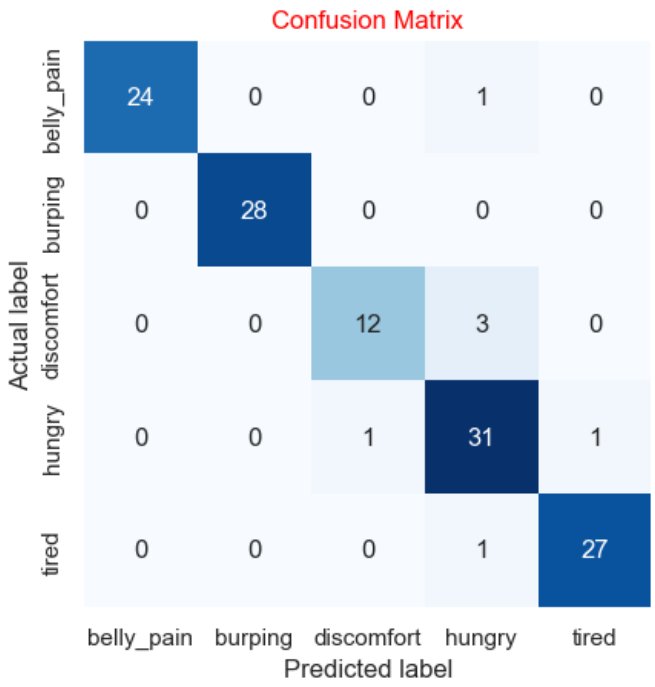

Fig. 6. Confusion Matrix

Fig. 7. Training and validation accuracy

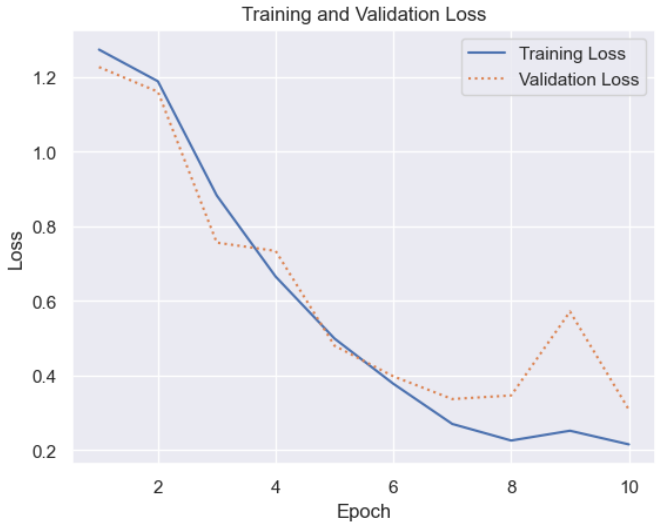

Fig. 8. Training and validation loss

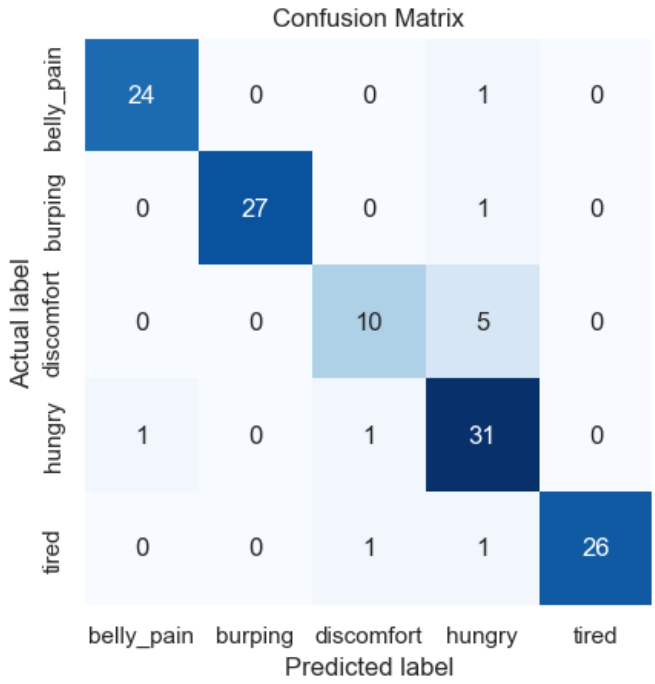

Fig. 9. Confusion Matrix

- Recall: 0.91
- Specificity: 0.9774

2) CNN Performance: :

- Testing Loss: 0.3824
- Testing Accuracy: 0.9457

Final Epoch Metrics:

- Accuracy: 0.9825
- Validation Accuracy: 0.9614
- Precision: 0.95
- Recall: 0.95
- Specificity: 0.9856

Both LSTM and 2D CNN performed well and achieved high testing accuracy, which indicates these models are highly competent in classifying data The CNN outperformed LSTM model with a better testing accuracy i.e 94.57 which is better in 91.47 Training accuracy was also better achieved by 2D CNN model with an accuracy of 98.25 and minimal loss. The precision ,recall and specificity of both models are pretty high showing good performance. With this in mind we see that 2D CNN performed better than LSTM slightly and had better

accuracy

IV. Conclusion

With the results and comparisons we have come to the conclusion that 2D CNN performed well for the baby crying dataset.Even when we refer to the confusion matrix we can see 2D CNN performing better on the testing dataset than the LSTM model.For future works on this project we can work on the following:

- We only did some basic augmentation to the data we can explore new ways of augmentation to train it better for real life audiosets.
- We can fine tune the 2D CNN model even more and play with parameters and fine tune them in such a way such that it performs better for this dataset.
- Implementing this model onto hardware
- Testing it with more data and expanding on the dataset we are working with as it is a small dataset with limited labels. Enhancing our model and training it with bigger and better dataset would help make the model work with more labels as well .
- Try including more features or representations during feature extraction to get a better performance.
- We can also explore combing multiple training models to improvethe accuracy of the model.
- Conducting user studies to evaluate real world performance and usability of this model.

References

[1] G. Coro, S. Bardelli, A. Cuttano, et al., "A self-training automatic infant-cry detector," Neural Comput & Applic, vol. 35, pp. 8543–8559, 2023. [Online]. Available: https://doi.org/10.1007/s00521-022-08129-w

[2] Ainsworth, M.D.S., Blehar, M.C., Waters, E., & Wall, S.N. (2015). Patterns of Attachment: A Psychological Study of the Strange Situation (1st ed.). Psychology Press. https://doi.org/10.4324/9780203758045

[3] J. O. Garcia and C. A. Reyes Garcia, "Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks," Proceedings of the International Joint Conference on Neural Networks, 2003., Portland, OR, USA, 2003, pp. 3140-3145 vol.4, doi: 10.1109/IJCNN.2003.1224074. keywords: Cepstrum;Pathology;Feedforward systems;Neural networks;Feedforward neural networks;Pediatrics;Acoustic signal detection;Pain;Ear;Signal analysis,

[4] G. Esposito and P. Venuti, "Comparative Analysis of Crying in Children with Autism, Developmental Delays, and Typical Development," Focus on Autism and Other Developmental Disabilities, vol. 24, no. 4, pp. 240-247, 2009. doi: 10.1177/1088357609336449.

[5] G. Esposito and P. Venuti, "Comparative Analysis of Crying in Children with Autism, Developmental Delays, and Typical Development," Focus on Autism and Other Developmental Disabilities, vol. 24, no. 4, pp. 240-247, 2009. doi: 10.1177/1088357609336449.

[6] S. Orlandi, C. Manfredi, L. Bocchi and M. L. Scattoni, "Automatic newborn cry analysis: A Non-invasive tool to help autism early diagnosis," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 2012, pp. 2953-2956, doi: 10.1109/EMBC.2012.6346583. keywords: Pediatrics;Autism;Variable speed drives;Resonant frequency;Frequency control;Protocols;Acoustics

[7] A. Kachhi, P. Gupta and H. A. Patil, "Features Motivated From Uncertainty Principle for Classification of Normal vs. Pathological Infant Cry," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 1253-1257, doi: 10.23919/EUSIPCO55093.2022.9909858. keywords: Pathology;Time-frequency analysis;Uncertainty;Cepstral analysis;Europe;Transforms;Signal processing;Infant cry classification;Heisenberg's uncer-tainty principle;Time-Bandwidth Product;u-vector;latency

[8] M. Z. Mohd Ali, W. Mansor, L. Y. Khuan and A. Zabidi, "Simulink model of Mel Frequency Cepstral Coefficient analysis for extracting asphyxiated infant cry features," 2012 International Conference on Biomedical Engineering (ICoBE), Penang, Malaysia, 2012, pp. 475-478, doi: 10.1109/ICoBE.2012.6179062. keywords: Mel frequency cepstral coefficient;Mathematical model;Feature extraction;Analytical models;Application software;Pediatrics;Digital signal processing;infant cry;mel frequency cepstral coefficient;matlab simulink,

[9] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN – RNN," J. Phys.: Conf. Ser., vol. 1528, no. 1, p. 012019, Apr. 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1528/1/012019.

[10] R. Torres, D. Battaglino, and L. Lepauloux, "Baby Cry Sound Detection: A Comparison of Hand Crafted Features and Deep Learning Approach," in *Proceedings of the International Conference on Image Analysis and Processing*, 2017, pp. 168-179. doi: 10.1007/978-3-319-65172-9_15.

[11] D. Widhyanti and D. Juniati, "Classification of Baby Cry Sound Using Higuchi's Fractal Dimension with K-Nearest Neighbor and Support Vector Machine," *Journal of Physics: Conference Series*, vol. 1747, no. 1, p. 012014, Feb. 2021. doi: 10.1088/1742-6596/1747/1/012014

[12] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN – RNN," *Journal of Physics: Conference Series*, vol. 1528, no. 1, p. 012019, Apr. 2020. doi: 10.1088/1742-6596/1528/1/012019.

[13] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data Augmentation Improves Recognition of Foreign Accented Speech," in Proc. Interspeech 2018, pp. 2409-2413, 2018. doi: 10.21437/Interspeech.2018-1211.

[14] V. R. Joshi, K. Srinivasan, P. M. D. R. Vincent, V. Rajinikanth, and C. Y. Chang, "A Multistage Heterogeneous Stacking Ensemble Model for Augmented Infant Cry Classification," *Frontiers in Public Health*, vol. 10, p. 819865, Mar. 24, 2022. doi: 10.3389/fpubh.2022.819865.

[15] C. A. Bratan et al., "Dunstan Baby Language Classification with CNN," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2021, pp. 167-171, doi: 10.1109/SpeD53181.2021.9587374. keywords: Pediatrics;Hospitals;Databases;Convolutional neural networks;Cultural differences;Australia;Dunstan baby language;convolutional neural network;infant cry classification,

[16] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN – RNN," Journal of Physics: Conference Series, vol. 1528, no. 1, p. 012019, Apr. 2020. doi: 10.1088/1742-6596/1528/1/012019.

[17] C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan, "A review of infant cry analysis and classification," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2021, no. 1, p. 8, Feb. 2021. doi: 10.1186/s13636-021-00197-5.