Identification of Infant Crying Using Mel-Frequency Cepstral Coefficient (MFCC) and Artificial Neural Network (ANN) Methods

Ahmad Azhari [a,1,*], Intan Destiyanti [a,2]

[a] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1] ahmad.azhari@tif.uad.ac.id
* corresponding author

ARTICLE INFO

Keywords
Dunstan Baby Language
Artificial Neural Networks
Mel-Frequency Cepstral Coefficient
Feature extraction

ABSTRACT

The crying of infants aged 0-3 months can be classified according to their needs, as identified by Dunstan Baby Language, which consists of specific sounds denoting different needs. These sounds include "eairh" for discomfort caused by fart, "neh" indicating hunger, "heh" representing general discomfort, "owh" signaling tiredness or sleepiness, and "eh" expressing the need to burp. The baby crying sound data was obtained from the Dunstan Baby Language (DBL) database, which includes educational videos about infants and a collection of babies crying sounds. These sounds were converted into *.wav audio format and divided into 5-second segments. A total of 188 audio data segments were collected. The research employed the Artificial Neural Network (ANN) classification method and the Mel-Frequency Cepstral Coefficient (MFCC) feature extraction method. The collected data underwent feature extraction, aiming to identify distinctive characteristics using the librosa library in the Python programming language. This process allowed us to obtain specific information from the acquired sound data. The results of this study achieved an accuracy level of 90%. This research contributes to the understanding and classification of infant crying based on the Dunstan Baby Language, offering insights into their various needs. The implementation of ANN and MFCC techniques showcases the effectiveness of this approach in accurately classifying infant cries and provides a foundation for further research in the field of infant communication.

1. Introduction

Newborn infants exhibit primitive reflexes, including spontaneous sounds and movements. Similar to human reflexes such as hiccups, burping, and sneezing, the patterns of these sound signals can be recognized. During the ages of 6 to 10 months, the primitive reflexes in infants begin to fade as their environment starts influencing their adaptive abilities [1], [2]. Therefore, the majority of infants aged 0-3 months generally have similar initial cries that can be easily identified through analysis. If not promptly responded to by their caregivers, infants will start crying hysterically. The cries of infants aged 0-3 months can be classified according to their needs using the Dunstan Baby Language, which includes specific sounds such as "eairh" indicating fart, "neh" signifying hunger, "heh" representing discomfort, "owh" denoting tiredness or drowsiness, and "eh" expressing the need to burp [3].

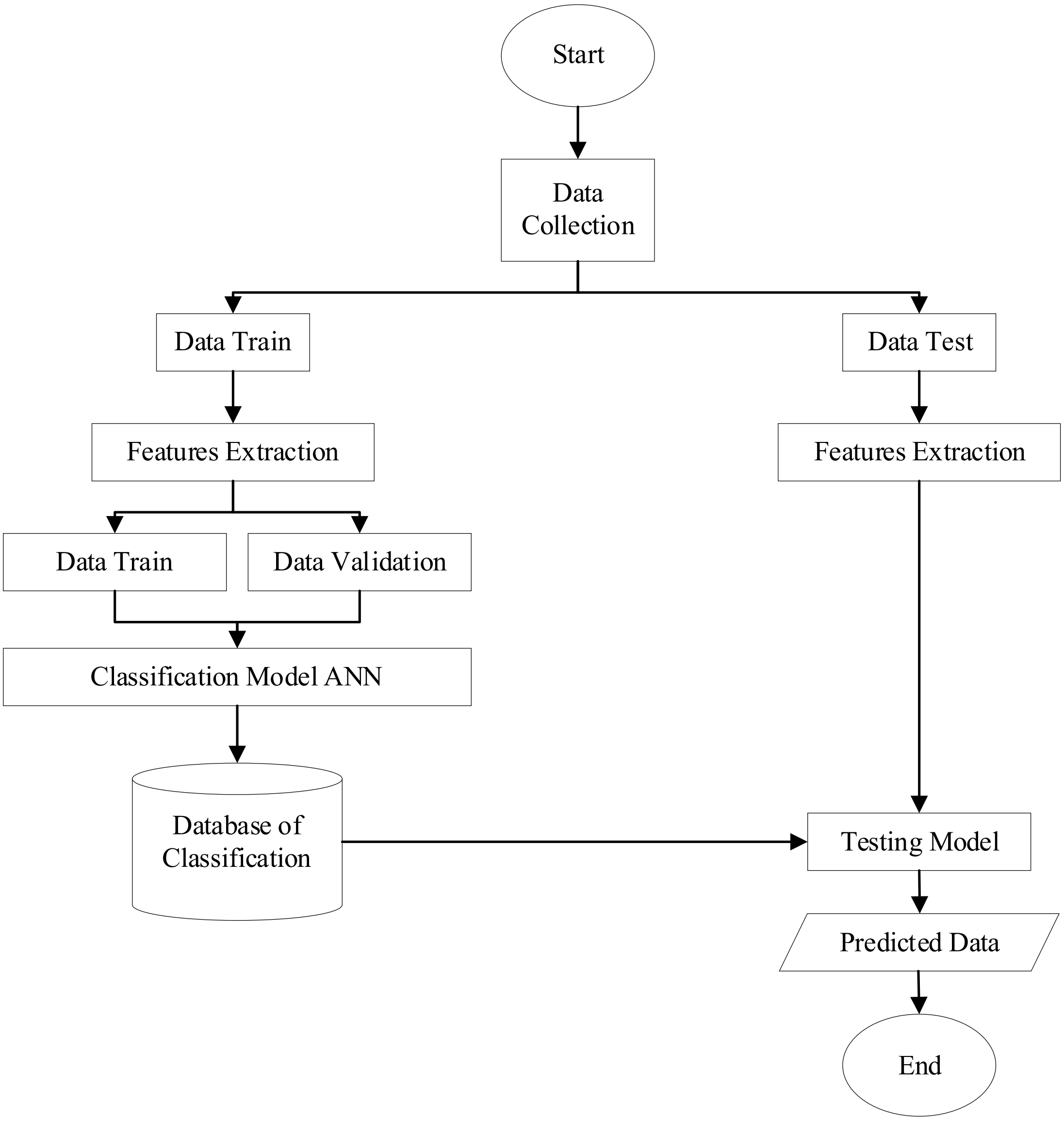
On average, infants tend to cry for about 2 hours per day during the first two weeks. When infants reach 6 weeks of age, they cry more frequently, approximately 2 hours and 15 minutes per day. By the time they reach 12 weeks of age, the duration of their crying decreases to an average of 1 hour and 10 minutes. There is significant variation in the duration of infant crying, with some infants crying for only 30 minutes per day while others cry for more than 5 hours per day [4].

This research focuses on developing an application capable of predicting the needs of infants. The application utilizes the Mel-Frequency Cepstral Coefficient (MFCC) method, which involves extracting features from infant crying sounds by dividing the data into different frequency components. Subsequently, the analysis of each component is performed based on the five categories defined by the Dunstan Baby Language. Additionally, the Artificial Neural Network (ANN) method, specifically the Multi-layer Perceptron (MLP) architecture with the Backpropagation training algorithm, is employed to predict the meaning behind infant cries.

2. Methods

2.1. System Design

The system design used is shown in the Fig. 1.



[redacted] System design diagram

2.2. Data Collection

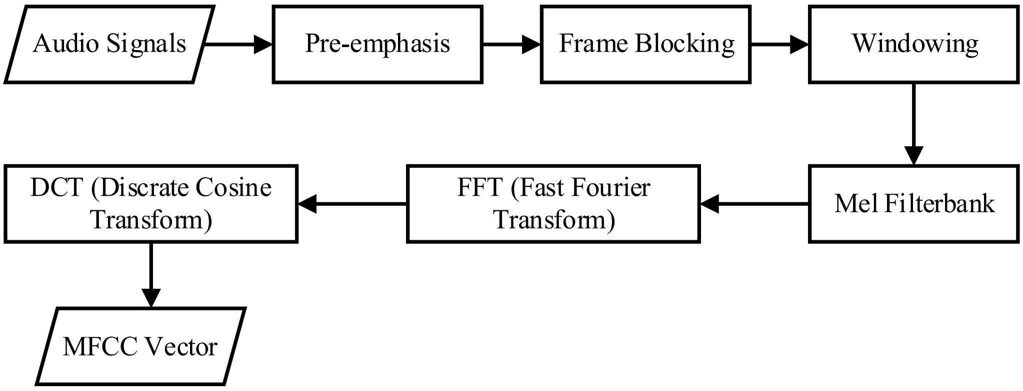
The data utilized in this research was extracted from the study conducted by Priscilla Dunstan between 1998 and 2006, which explored the meaning behind infant cries known as Dunstan Baby Language (DBL). The data consisted of videos that were processed to yield a total of 188 audio samples of infant cries. The data was divided into two subsets, namely the testing data and training data, as presented in Table 1.

Table 1. Data collection results

Classes	Data Train	Data Test	Total
Fart	44	6	50
Hungry	25	6	31
Uncomfortable	25	6	31
Sleepy	35	6	41
Burp	29	6	35
Total	158	30	188

2.3. Features Extraction

Feature extraction is the process of capturing distinctive characteristics from a given value or vector. In this study, the feature extraction process employed the Mel-Frequency Cepstral Coefficient (MFCC) method, which calculates cepstral coefficients while taking into account human auditory perception. The flow diagram of the MFCC can be observed in Fig. 2.



[redacted] Flowchart MFCC

1. Pre-emphasis is a process that reduces noise by emphasizing the high-frequency components, aligning them with the low-frequency and high-frequency components. This step utilizes Equation (1) to achieve the desired effect.

$$y(n) = s(n) - a.s(n-1) \tag{1}$$

2. Frame blocking [5] is the process of analyzing the signal in the form of frames. In this stage, the pre-emphasized speech signal is segmented into frames. The frame length used in the signal processing is typically set to a default range of 10-30 ms, with a constant time lapse of 20 ms throughout the process.

3. Windowing [6]–[9] is the process of smoothing the spectrum after frame blocking to reduce the discontinuity effects at the edges of the frames resulting from frame blocking. Its purpose is to obtain accurate signals within very short time intervals. This stage employs Equation (2), where N represents the number of frames present in each sample.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right) \tag{2}$$

4. Mel-Filterbank [10]–[12] is a process that transforms the audio signal from the frequency domain to the Mel-frequency domain. The Mel-Filterbank involves several stages aimed at determining the upper and lower boundaries of the filters, dividing the range between the

upper and lower boundaries according to the number of filters, and converting the upper and lower boundaries to the mel scale for each filter bank. The conversion from the mel scale back to the linear frequency scale is performed for both boundaries using Equation (3), where f represents the frequency.

$$mel(f) = 1125\, x \ln{(1 + \frac{f}{700})} \tag{3}$$

5. FFT (Fast Fourier Transform) is a technique used to convert a digital signal from the time domain to the frequency domain [13], with the aim of improving computational efficiency. In this stage, the FFT is performed on all frames of the signal that have undergone the windowing process. The FFT stage can be observed in Equation (4).

$$f(n) = \sum_{k=0}^{n-1} W_k e \frac{2\pi jkn}{N}, 0 \leq n \leq N-1 \tag{4}$$

6. DCT (Discrete Cosine Transform) is a process used to convert a signal from the frequency domain back to the time domain using the DCT [14]–[16]. Taking the logarithm of the multiplication result in the time domain yields the Mel Frequency Cepstral Coefficients (MFCCs), which are the output of the MFCC process and can be represented by Equation (5).

$$C_j = \sum_{i=1}^{M} X_i \cos{(j(i-1)/2 \frac{\pi}{M})} \tag{5}$$

2.4. Artificial Neural Network Classification

Artificial Neural Network (ANN), also known as a multi-layer perceptron (MLP), consists of multiple layers, including the input layer, hidden layer(s), and output layer [17], [18]. The input layer is responsible for receiving data, the hidden layer performs computations and can have multiple layers or none at all, and the output layer generates the final output based on the input and hidden layers. ANN can have different architectural configurations, such as single-layer, multi-layer, and competitive layer.

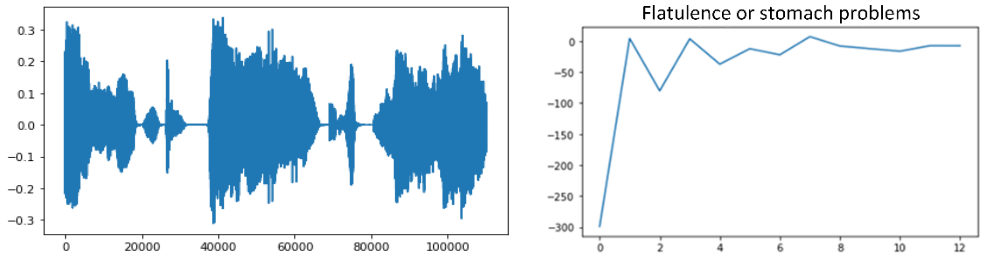
The Multi-layer Perceptron (MLP) is a feed-forward artificial neural network that consists of one or more hidden layers. Each layer in the MLP has a specific function [19]. The input layer receives input signals/vectors from external sources and distributes them to all neurons in the hidden layer. The output layer receives the output signals (or pattern stimuli) from the hidden layer and produces the final output signal/value/class of the entire network.

In ANN, the learning or training process involves determining the optimal weights to be used in the testing phase [20]. There are various training algorithms available, with Backpropagation being the most popular one. The training procedure in Backpropagation is similar to that of a Perceptron. A set of training data is provided as input to the network. The network computes the output, and if there is an error (the difference between the desired target output and the actual output), the network's weights are updated to minimize that error.

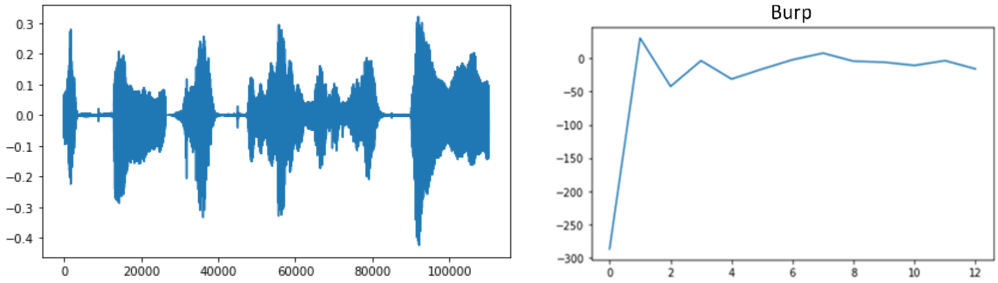
3. Results and Discussion

3.1. Feature Extraction

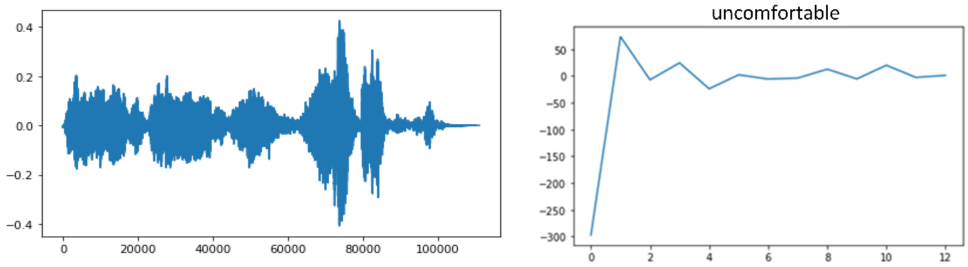
In this study, audio samples from Dunstan Baby Language (DBL) were utilized. DBL consists of educational videos on infant crying in the age range of 0-3 months, from which only the segments containing baby cries were extracted. The study focused on 5 classes based on DBL categories: gas or stomach-related issues, hunger, sleepiness, discomfort, and burping. The patterns within each class were identified using the Mel-Frequency Cepstral Coefficient (MFCC) method with 13 coefficients and a sample rate of 22050Hz. The number of MFCC coefficients typically ranges from 9 to 13 since the majority of signal energy is concentrated in the first few coefficients due to the nature of cosine transformation. The results of the MFCC process are presented in Fig. 3.



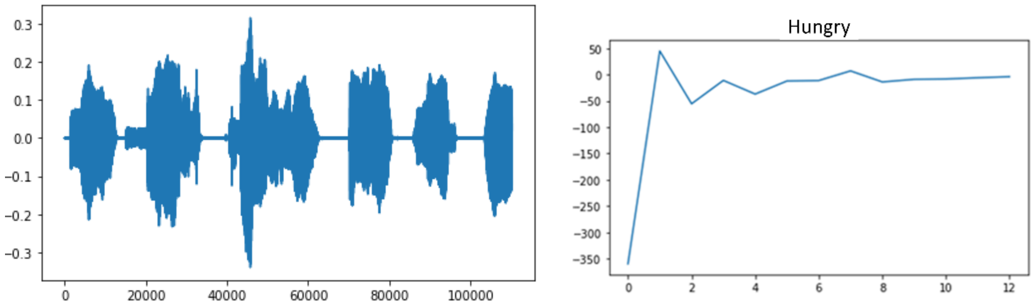
[redacted] Sound signals to the sound pattern on the fart label



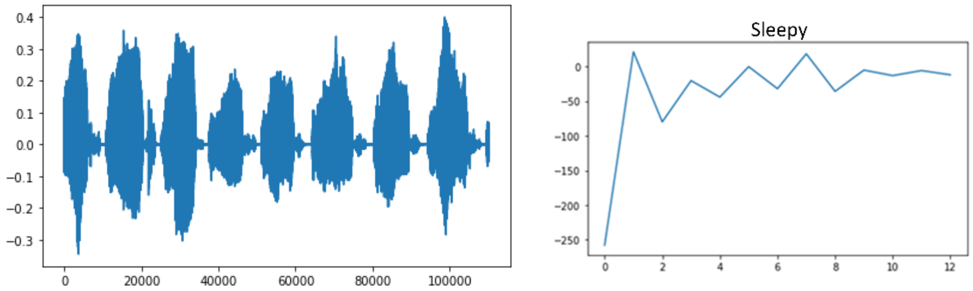
[redacted] Sound signals to the sound pattern on the burp label



[redacted] The sound signal to the sound pattern on the label is uncomfortable



[redacted]Sound signal to the sound pattern on the hungry label



[redacted] Sound signals to the sound pattern on the sleepy label

3.2. Training, Validation and Testing

The data in this study were divided into several parts, namely the training data and validation data, which were split from the dataset in an 80% to 20% ratio. The training data consisted of 126 samples, while the validation data comprised 32 samples. Additionally, a separate set of 30 samples was reserved for the testing data. The data partitioning can be observed in Table 2.

Table 2. Separating audio data

Classes	Data Train	Data Validation	Data Test	Total
Classes	36	8	6	50
Fart	17	8	6	31
Hungry	22	3	6	31
Uncomfortable	30	5	6	41
Sleepy	21	8	6	35
Burp	126	32	30	188

3.3. Artificial Neural Network

This study employed the Sequential Artificial Neural Network (ANN) method, specifically utilizing the Multi-layer Perceptron (MLP) architecture with the Backpropagation training algorithm. The constructed ANN architecture consisted of one input layer, three hidden layers, and one output layer. The input layer comprised 40 neurons, which were derived from the MFCC coefficient values. The first hidden layer consisted of 100 neurons with the ReLU activation function, the second hidden layer comprised 200 neurons with ReLU activation, and the third hidden layer included 100 neurons utilizing the ReLU activation function. The output layer contained 5 neurons, corresponding to the 5 labels/classes, and employed the softmax activation function. For a detailed overview, including the layer types, output shapes, number of parameters (weights) in each layer, and the total number of connected parameters (weights) in the neurons overall, please refer to Table 3.

Table 3. Model Summary

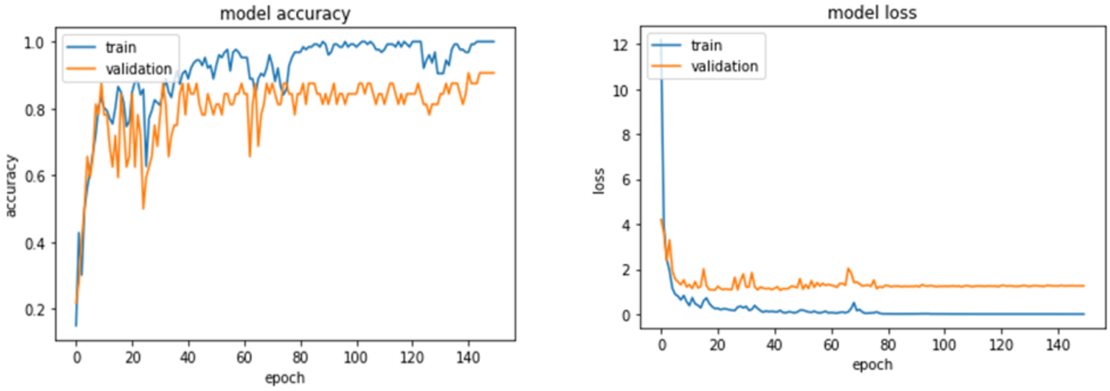
Layer (Type)	Output Shape	Param #
dense (Dense)	(None, 100)	1400
activation (Activation)	(None, 100)	0
dense_1 (Dense)	(None, 200)	20200
activation_1 (Activation)	(None, 200)	0
dense_2 (Dense)	(None, 100)	20100
activation_2 (Activation)	(None, 100)	0
dense_3 (Dense)	(None, 5)	505
activation_3 (Activation)	(None, 5)	0
Total params: 42.205		
Trainable: 42.205		
Non-trainable params: 0		

3.4. Training Model ANN

During the training process, the model was compiled by specifying the loss function, accuracy metric, and optimizer. The categorical_crossentropy loss function was employed to calculate the loss between the labels and predictions. The accuracy metric, utilizing the accuracy type, was employed to measure how often the predictions matched the labels during the training process. Lastly, the Adam optimizer was utilized for the optimization algorithm. To ensure compatibility with the constructed model, the batch size, which denotes the number of data samples propagated through the neural network, and the number of epochs were specified. Table 4 provides a summary of the model compilation and model fit employed in this study. The training process yielded an accuracy of approximately 90% and a loss value of 1.25, as illustrated in Fig. 8a for accuracy visualization and Fig. 8b for loss visualization.

Table 4. Compilation model and model fit

Classes Attributes	Parameters	Value
Model Compilation	Loss Function Metrics Accuracy Optimizer	categorical_crossentropy accuracy adam
Model Fit	batch_size epochs	20 150

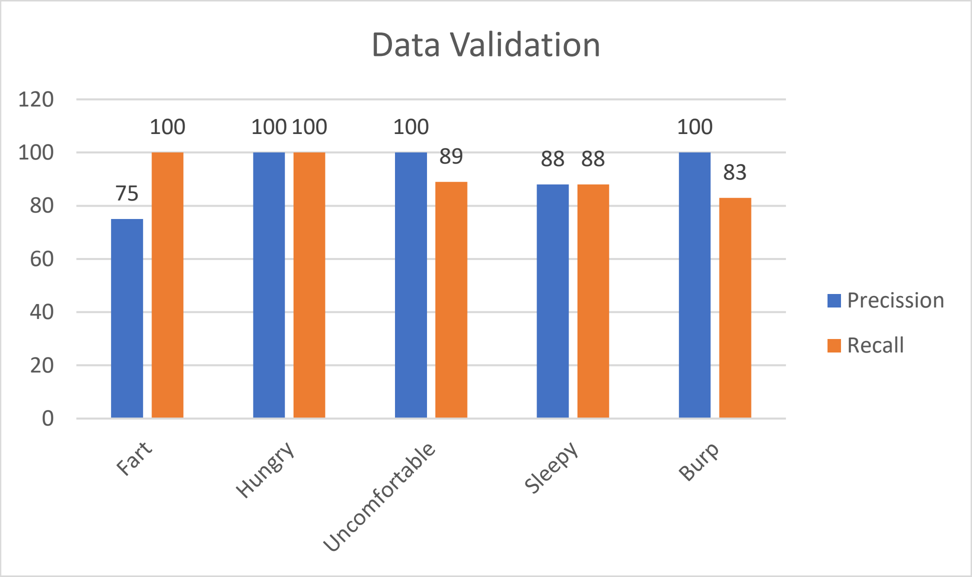


(a) (b)

[redacted] (a) Model Accuracy (b) Model Loss

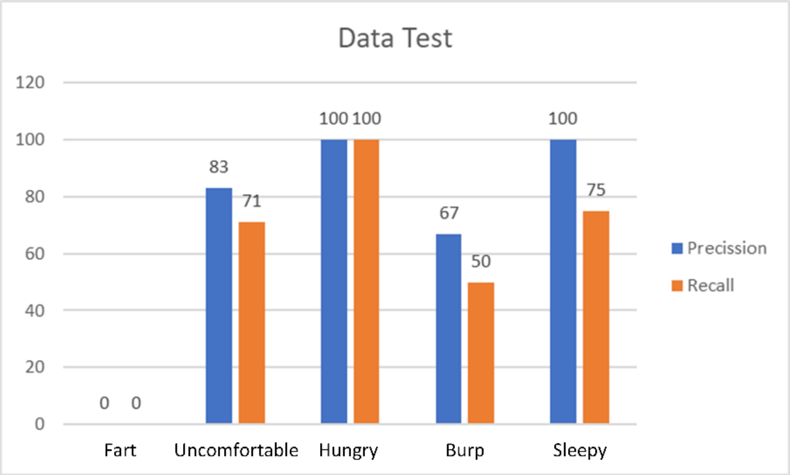
3.5. Evaluation Model

The research evaluation employed the confusion matrix method to assess the performance of the developed system by testing the training data, validation data, and test data. Fig. 9 presents the precision and recall results for each label, indicating that the "burping" label achieved the highest precision and recall, both at 100%.



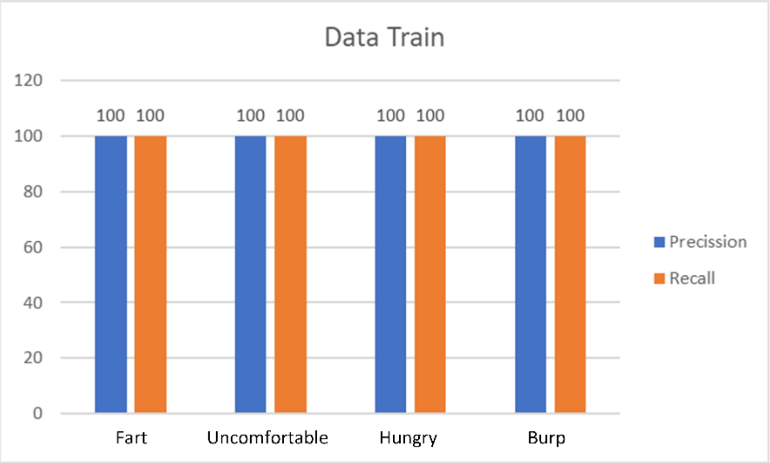
[redacted] Confusion matrix results each label on Data Validation

In Fig. 10, the per-label confusion matrix results for the test data can be observed. It can be seen that the "burping" label has the lowest precision and recall scores, both at 0%. This indicates that for audio samples labeled as "burping," the predictions did not find any matches or the predictions were incorrect. On the other hand, the "hungry" label achieved the highest precision and recall scores, both at 100%.



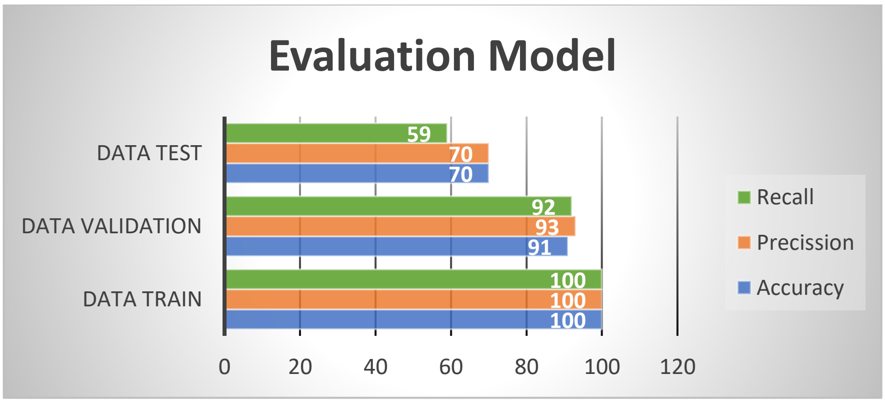
[redacted]Confusion matrix results each label on Data Test

In Fig. 11, the confusion matrix results for the training data are displayed. The precision and recall scores for all labels show a perfect score of 100%. This indicates that during the prediction process on the training data, all the results matched the expected labels, resulting in accurate predictions for all samples.



[redacted]Confusion matrix results each label on Data Train

The results of accuracy, precision, and recall for all datasets can be observed in Fig. 12. The lowest scores are obtained from the test data, with a precision of 70%, recall of 59%, and accuracy of 70%. The second highest scores are achieved by the validation data, with a precision of 93%, recall of 92%, and accuracy of 91%. The highest scores are obtained from the training data, with a precision of 100%, recall of 100%, and accuracy of 100%.



[redacted]Confusion matrix results

4. Conclusion

The data obtained in this study consists of a total of 188 audio cry samples, divided into two subsets: the test data with 30 audio samples and the training data with 158 audio samples. These samples are further categorized into 5 labels/classes. The feature extraction method utilized in this research is MFCC, with a coefficient count of 13, which is employed as input nodes in the ANN model. The trained model achieved an accuracy rate of 90%. The evaluation results for system performance using the confusion matrix method show that the test data achieved an accuracy of 70%, the validation data achieved an accuracy of 93%, and the training data achieved a perfect accuracy of 100%. The implementation of the ANN method in the system demonstrates the ability to classify cry samples based on the categories in the DBL database and accurately predict the meaning behind infant cries.

References

[1] E. Franti and M. Dascalu, "Testing the Universal Baby Language Hypothesis-Automatic Infant Speech Recognition with CNNs," in *Telecommunications and Signals Processing (TSP)*, 2018, pp. 424–427. doi: https://doi.org/10.1109/TSP.2018.8441412.

[2] A. Chinello, V. Di Gangi, and E. Valenza, "Persistent primary reflexes affect motor acts: Potential implications for autism spectrum disorder," *Res Dev Disabil*, vol. 83, pp. 287–295, Dec. 2018, doi: 10.1016/j.ridd.2016.07.010.

[3] Eva Rhea Moeckel and Noori Mitha, *Textbook of Pediatric Osteopathy*. Elsvier Health Sciences, 2008. Accessed: Jun. 17, 2023. [Online]. Available: https://books.google.co.id/books?id=Y9Dpcqr7PZMC&dq

[4] D. Wolke, rer nat hc, A. Bilgin, and M. Samara, "Systematic Review and Meta-Analysis: Fussing and Crying Durations and Prevalence of Colic in Infants," *J Pediatr*, vol. 185, pp. 55–61, 2017, doi: https://doi.org/10.1016/j.jpeds.2017.02.020.

[5] O. Kamil, "Frame Blocking and Windowing Speech Signal," *Journal of Information, Communication and Intellegence System (JICIS)*, vol. 4, no. 5, pp. 87–94, 2018, [Online]. Available: https://www.researchgate.net/publication/331635757

[6] S. Zhang, E. Loweimi, P. Bell, and S. Renals, "Windowed Attention Mechanism for Speech Recognition," in *IEEE Signal Processing Society*, 2019, pp. 7100–7104.

[7] A. Kumar, G. Verma, C. Rao, A. Swami, and S. Segarra, "Adaptive Contention Window Design using Deep Q-learning," in *IEEE International Conference on Coustics, Speech and SIgnal Processing (ICASSP)*, Nov. 2020. doi: https://doi.org/10.1109/ICASSP39728.2021.9414805.

[8] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 47795–47814, 2021. doi: 10.1109/ACCESS.2021.3068045.

[9] I. Kavalerov, S. Wisdom, Erdogan Hakan, and Patton Brian, "Universal Sound Separation," in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, 2019, pp. 175–179. doi: https://doi.org/10.1109/WASPAA.2019.8937253.

[10] L. Juvela *et al.*, "Speech Waveform Synthesis from MFCC Sequences with Generative Adversarial Networks," in *IEEE Signal Processing Society*, 2018, pp. 5679–5683. doi: https://doi.org/10.1109/ICASSP.2018.8461852.

[11] "Choice of Mel Filter Bank in Computing MFCC of a Resampled Speech," in *Universiti Teknologi Malaysia*, IEEE, 2010, pp. 121–124. doi: https://doi.org/10.1109/ISSPA.2010.5605491.

[12] C. Donahue and B. L. R. Prabhavalkar, "Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition," in *IEEE Signal Processing Society*, 2018, pp. 5024–5028. doi: https://doi.org/10.1109/ICASSP.2018.8462581.

[13] P. M. Kumar and K. Srinivas, "Real Time Implementation of Speech Steganography," in *International Conference on Smart System and Inventive Technology*, 2020, pp. 365–369. doi: https://doi.org/10.1109/ICSSIT46314.2019.8987785.

[14] X. Liu, M. Sahidullah, and T. Kinnunen, “Learnable MFCCs for speaker verification,” in *Proceedings - IEEE International Symposium on Circuits and Systems*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ISCAS51556.2021.9401593.

[15] L. Stankovic and M. Brajovic, “Analysis of the reconstruction of sparse signals in the DCT domain applied to audio signals,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 26, no. 7, pp. 1216–1231, Jul. 2018, doi: 10.1109/TASLP.2018.2819819.

[16] M. E. Rahaman, S. M. Shamsul Alam, H. S. Mondal, A. S. Muntaseer, R. Mandal, and M. Raihan, “Performance Analysis of Isolated Speech Recognition Technique Using MFCC and Cross-Correlation,” in *International Conferences on Computing, Communication and Networking Technologies (ICCCNT)*, 2019. doi: https://doi.org/10.1109/ICCCNT45670.2019.8944534.

[17] Mahāwitthayālai Songkhlānakharin. College of Computing, C. Electrical Engineering/Electronics, IEEE Thailand Section, and Institute of Electrical and Electronics Engineers, *A Light-Weight Artificial Neural Network for Speech Emotion Recognition using Average Values of MFCCs and Their Derivatives*. 2020. doi: https://doi.org/10.1109/ECTI-CON49241.2020.9158221.

[18] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, “Parkinson’s Disease Indentification using KNN and ANN Algorithm based on Voice Disorder,” in *International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2020. doi: https://doi.org/10.1109/IRASET48871.2020.9092228.

[19] M. Kahani, M. H. Ahmadi, A. Tatar, and M. Sadeghzadeh, “Development of multilayer perceptron artificial neural network (MLP-ANN) and least square support vector machine (LSSVM) models to predict Nusselt number and pressure drop of TiO2/water nanofluid flows through non-straight pathways,” *Numeri Heat Transf A Appl*, vol. 74, no. 4, pp. 1190–1206, Aug. 2018, doi: https://doi.org/10.1080/10407782.2018.1523597.

[20] S. Liu, A. Borovykh, L. A. Grzelak, and C. W. Oosterlee, “A neural network-based framework for financial model calibration,” *J Math Ind*, vol. 9, no. 1, Dec. 2019, doi: 10.1186/s13362-019-0066-7.