Journal of Human Centered Technology

https://www.humentech.utm.my E-ISSN: 2821-3467

Baby Crying Sound Classification using Convolutional Neural Network

Naufal Fikri Muhammad[1,2], Raimi Dewan[1,2,3*], Jaysuman Pusppanathan[1], Faishal Adilah Suryanata[1,2]

[1]Department of Biomedical Engineering and Health Sciences, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia
[2]Advanced Radio Frequency & Microwave Research Group, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia
[3]IJN-UTM Cardiovascular Engineering Centre, Institute of Human Centered Engineering, Universiti Teknologi Malaysia, Malaysia

*Corresponding Author raimi.dar@utm.my

Cite: https://doi.org/10.11113/humentech.v3n1.66 Research Article

Abstract:
Crying is a crucial means of communication for newborns, crying is a newborn's early form of communication. Many individuals are unable to recognise a baby's intention from cry unless they have the appropriate training or expertise, such as nurses, paediatricians, and childcare professionals. Accurately interpreting a baby's cry can be challenging. In this research paper, the study uses a method for classifying baby crying sounds using a Convolutional Neural Network (CNN) and the dataset includes belly pain, burping, discomfort, hungry, and tired for total of 3,495 one-second-long audio clips. The research methodology involves preprocessing the audio data, extracting Mel-Frequency Cepstral Coefficients (MFCC) as features, and training the CNN model. To determine the optimal architecture, two different configurations of the CNN model are evaluated. The settings for both configurations are the same, except for the layers. The first configuration utilizes 100, 200, and 100 neurons for the respective layers, while the second configuration employs 256, 512, and 256 neurons for each layer. the results have already been evaluated that the second configuration, with deeper and more complex layers, achieves higher accuracy (86%) compared to the first configuration (84%). The study demonstrates the effectiveness of CNNs in classifying baby cries and highlights the importance of model architecture in achieving accurate classification results. Future research could explore larger and more diverse datasets to improve generalizability.

Keywords: Baby cry; Convolutional neural network; Machine learning; Mel-frequency cepstral coefficient; Sound classification

1. INTRODUCTION

Crying is a newborn's early form of communication. Many individuals are unable to recognise a baby's intention from cry sound unless they have the appropriate training or expertise, such as nurses, paediatricians, and childcare professionals (1). The non-invasive technique of analysing baby cries for medical diagnosis in clinical settings is gaining popularity. Attention has been drawn to the complexity and dynamics of baby cries for their potential to diagnose health issues without the need of invasive techniques (2-3). The analysis of a baby's cry is based on listening to the audio signal and analysing its shape and spectrogram, which requires clinicians with extensive training. However, the subjective nature of listening to a baby's cries can lead to inconsistent and potentially incorrect diagnoses, especially when managed by different caretakers, in less-than-ideal conditions, or in a short amount of time (4).

Baby cries are non-linear, indicating that they do not conform with traditional statistical models or reflect a predictable pattern (5). The varied dynamics of baby cries refer to the various elements that may add to the cry, including pitch, volume, and duration, which can rapidly and unpredictably change (5). The issue of cry recognition has already been addressed in the literature; numerous approaches to the problem have been proposed, and numerous corpora have been used to evaluate them (6). Previous research has delved into this field, employing a variety of techniques and strategies to address the difficulties as shown on Table 1.

Due to advances in child development, baby screams are no longer a mystery; babies cry in specific ways to express specific needs that may require the attention of a caretaker. However, mastering these specific expressions takes time and can be challenging (7). Current research focuses on the learning of baby language to know their intention. Since Pricilla Dunstan discovered patterns in her baby's cries and founded the Dunstan Baby Language (DBL) program, this involves identifying a baby's underlying need as communicated through weeping. The original DBL classified five fundamental needs: colic, diaper, burp, appetite, and discomfort. Recent research indicates that a growing diversity of newborn cries now necessitates the representation of distinct wavelets.

© 2024 Penerbit UTM Press. All rights reserved

Received 08 December 2023; Accepted 02 February 2024
Available online 06 February 2024

In accordance with a similar way of thinking, this study proposes an approach with the potential to enable a machine to comprehend the complex meaning communicated by an infant's cry. The cry emitted by babies is influenced by their physiological and psychological state, as well as external or internal stimuli (8). Furthermore, the vocalisation of baby might serve as an indicator for the presence of pathological conditions, enabling early detection and intervention, which is crucial for the well-being of the infant (9).

Table 1. Previous studies focused on classifying infant crying sounds.

First author	Dataset	Features	Classifiers	Accuracy
Felipe *et al.* (10)	iCOPE (pain vs. no pain)	Mel Scale (MS), MFCC, Constant-Q Chromagram (CQC), Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Robust Local Binary Pattern (RLBP) extracted from spectrogram	SVM	71.68%
Franti *et al*. (11)	Dunstan Baby Database (pain, hunger)	Spectrogram	CNN	89%
Liu *et al*. (12)	NICU recorded (draw attention cry, diaper change needed cry, and hungry)	LPC, LPCC, MFCC, BFCC	Nearest Neighbor Artificial Neural Network	76.4%

The majority of research on categorizing newborn cries concentrates on a limited number of variables, such as time domain and time-frequency approaches (13). Mel Frequency Cepstral Coefficient (MFCC) is a common and well-known method for extracting features from a signal input that two research from the Table 1 using MFCC feature. It illustrates the spectral envelope defined by a set of frequencies. In each stage of the algorithm, multiple mathematical equations are utilized (14). The features are inspired by the behavior of the human auditory system. Humans are substantially better at detecting minor pitch variations at low frequencies than at high frequencies. The Mel-scale makes characteristics more closely correspond to human-perceivable / audible sounds (14-15). The human auditory system is nonlinear, meaning that it perceives sounds of different frequencies differently. Human auditory is typically more sensitive to sound waves with lower frequencies and less sensitive to those with higher frequencies. The relationship between Mel and Hertz frequencies is depicted by the Equation 1, and Figure 1 shows the nonlinear relationship between Mel and Hertz frequencies (16).

$$Mel(f) = 2595\, log\left(1 + \frac{f}{700}\right) \quad (1)$$

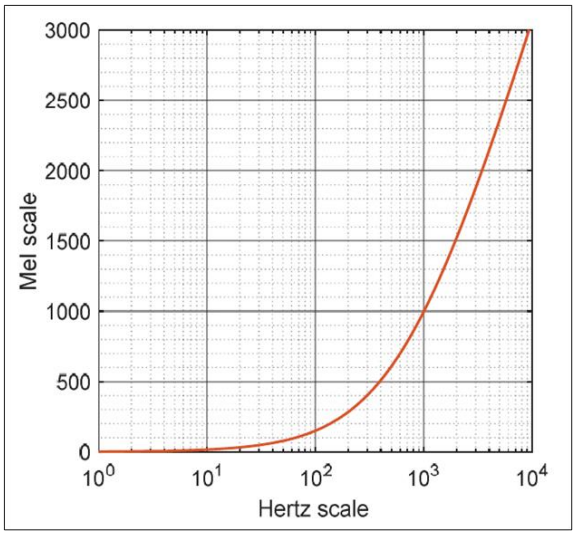


Figure 1. Non-linear relationship between Mel and Hertz (16).

CNN will be implemented in this study. The research findings indicate that a CNN with random initialization can autonomously learn diverse sets of edge detectors for the local extraction of low-level features (17). Extensive experiments show that CNNs are superior to Deep Neural Network (DNN) in four domains: channel-mismatched training-test conditions, noise robustness, distant speech recognition, and low-footprint models (17).

For sound recognition tasks, machine learning classifiers, such as a Convolutional neural network (CNN), are required once sound features have been defined. CNNs are based on the architecture of multi-layer perceptron based on Figure 2 with a number of significant modifications. The three dimensions, height, width, and depth are grouped into layers. In

© 2024 Penerbit UTM Press. All rights reserved

addition, not every node in each layer is connected to every node in the next layer. The architecture permits the CNN model to operate in two phases. Initially, during the phase of feature extraction in which a filter window traverses the input and extracts the total convolution at each position, the feature map retains these extracted features from each window. In between the CNN layers is a pooling procedure (18).

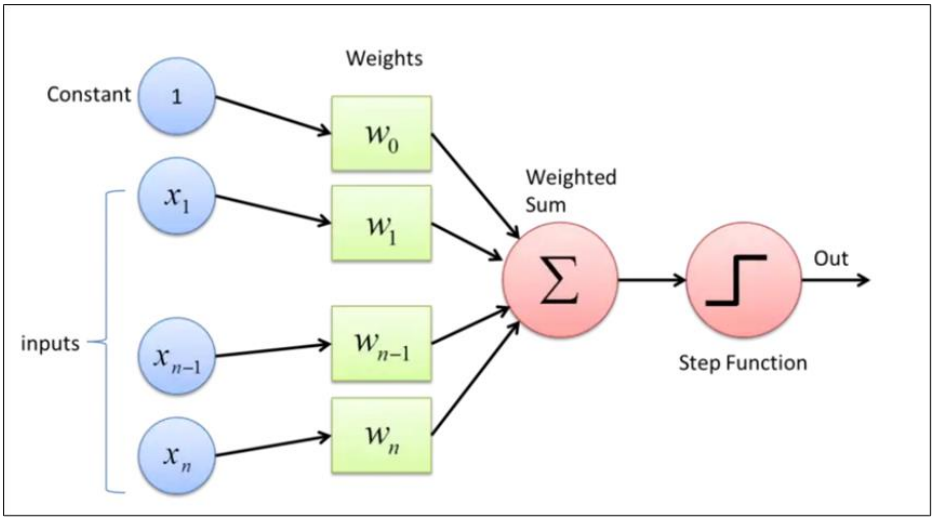


Figure 2. Illustration of perceptron, basic unit in neural networks, used for pattern classification (19).

Pattern recognition, mean the separation of significant features from irrelevant peripheral information (20). Pattern recognition typically employs, a well-known deep learning technique. A transmission line intelligent diagnosis method based on a convolutional neural network that improved defect classification precision. CNN, a multi-layered neural network, connects all feature maps, allowing it to learn based on its weights (21). Illustrative example of One-dimensional structure of CNN is shown in Figure 3 (22).

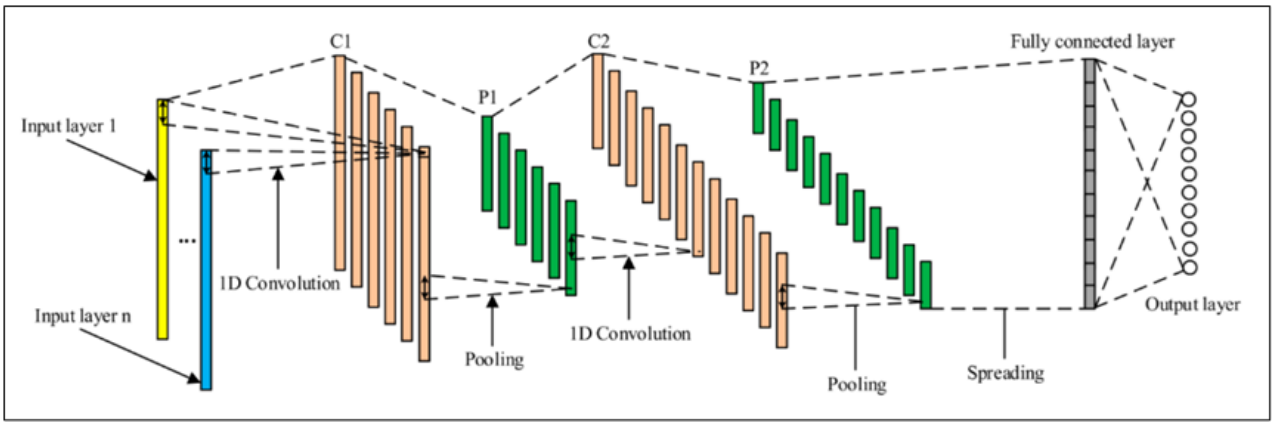


Figure 3. Illustrative of one-dimensional CNN structure (22).

MFCC as the feature and CNN's machine learning had been studied in previous speech recognition research. The work from Mustaqeem *et al*. (23) with title “A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition” combining a spectrogram with a CNN model for speech recognition is the objective of this endeavour. The work obtains an impressive recognition rate of 79.5% for emotions such as anger, happiness, natural, and sadness. Another work from Pelchat *et al.* (24) which also extracts the MFCC features and spectrogram as a representation of the MFCC feature, which is then employed in the CNN model for music genre classification. The outcome was 85% accuracy.

2. EXPERIMENTAL METHODS

This section concentrates on the research methodology, which includes data collection, machine learning processing, and the investigation of various configurations. Figure 4 provides an overview of the research process, highlighting sequential steps taken. The initial step involves preprocessing the input data, followed by conversion process to extract Mel-frequency cepstral coefficients (MFCC) features. These MFCC features serve as the foundation for building the Convolutional Neural Network (CNN) model. Finally, the output, or results of the model are evaluated, employing confusion matrix to assess the performance and effectiveness of the classification model.

© 2024 Penerbit UTM Press. All rights reserved

In particular, the study examines variations in the design, such as the configuration of dense layers within the CNN model and the selection of the number of epochs. The goal is to evaluate how these variations affect the model's performance and identify the configurations that produce the most accurate classification results. This study contributes to an increased understanding of the significance of architectural decisions and training duration in CNN-based infant sound classification.

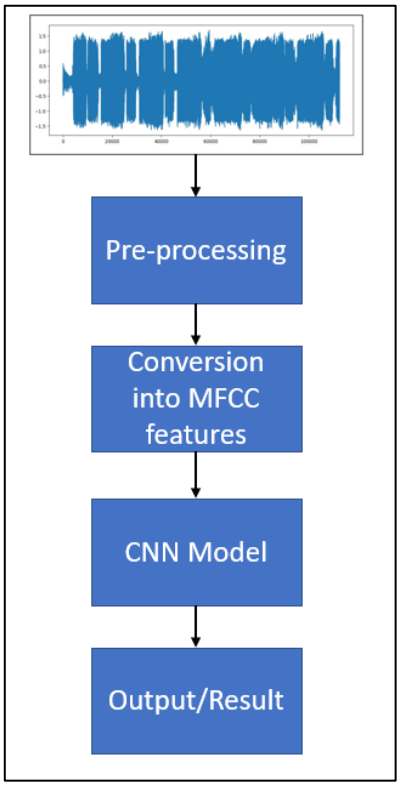


Figure 4. General process of sound classification.

2.1. Data Collection and Software

Integrated development environment (IDE) and a clean audio dataset with the correct labelling of infant cry intention were required for project execution. The collection of audio data was acquired from GitHub (25) while another source is from the Baby Chillanto Data Base to complement the datasets (26). Github audio dataset is a known as the Donate-a-cry campaign's infant cry audio, The audio files contain baby cry samples and were tagged by the contributors themselves. Baby Chillanto Data Base was developed by the recording of medical doctors, which is a property of NIAOE-CONACYT, Mexico (26). The bit rate and sampling rate of all the data has 128 kbps and 8 kHz, respectively.

The respective sources have performed a data cleaning process to eliminate non-based-cries sounds from the available datasets, including white noise, baby chatter, adult mimicking baby cries, etc. To ensure the quality, dependability, and precision of the dataset used to train the model, audio data must be cleaned from potential noises or unwanted data. All audio files are organized into folders based on their respective Dunstan infant language classification categories (tired, belly pain, burping, discomfort, and hunger). It should be noted that the dataset only contains male and female infants aged 0 to 2 years.

This research paper presents a dataset of baby noises comprised of 3,495 one-second-long audio clips, which were painstakingly compiled. The dataset contains distinct categories such as belly pain, burping, discomfort, hungry, and tired. Notably, the dataset exhibits substantial class distribution variations. The most prevalent category, with 2,808 occurrences, is hunger. After hunger, there are 302 samples of abdominal pain, 174 samples of discomfort, 157 samples of tired, and 54 samples of burping. This dataset is a valuable resource for future research and the development of classification and comprehension algorithms for prevalent infant cries.

Jupyter notebook is the IDE for machine learning classification. It is a web-based, open-source environment for interactive computation that supports Python, R, LaTeX, and Javascript, among others. It enables the incorporation of theoretical explanations, computer codes, simulation results, and plots in a single file, resulting in a workspace where people unfamiliar with programming languages can learn concepts by observing simulation results (27).

The CNN model was created using Jupyter Notebook, a prominent environment that facilitates the use of the Python programming language. Python provides a collection of machine learning-specific libraries that were instrumental in the development of the model. These libraries include Scikit-learn (sklearn), TensorFlow, librosa, and matplotlib. Additionally, Jupyter notebook can be utilized to visualize the model's output.

© 2024 Penerbit UTM Press. All rights reserved

2.2. Data Preprocessing

A primary goal of the preprocessing phase is to ensure uniformity and consistency in the subsequent extraction of MFCC features. The first step of the preprocessing phase entails converting the audio from a two-channel audio format to a single-channel audio format to reduce complexity. Because the audio and music signal bandwidth are limited to 20 kHz, a sampling rate of 16 bits has been selected. The sample frequency rates are 22050 Hz based on the Nyquist frequency calculation as shown in Equation 2 (28).

$$f_{nyq} = f_{in.max} \qquad (2)$$

2.3. MFCC Features Extraction

The extraction of Mel-frequency cepstral coefficients (MFCCs) involves transforming the entire audio dataset into MFCC features using the librosa library and its 'librosa.feature.mfcc' function. The function n_mfcc=40 specifies that 40 MFCC should be extracted from the audio signal. MFCCs are coefficients that represent the short-term power spectrum of a sound signal. The choice of 40 coefficients is a common practice, and the specific number may vary based on the application and the characteristics of the audio data.

This function from the librosa library transforms audio data in a way that facilitates the computation of MFCCs. This conversion process is required to obtain a consistent representation of the audio signals in the form of MFCC features, which are necessary inputs for the classification process's subsequent analysis and model tasks. Then the datasets are imported and the MFCC extractions are applied to all the audio files apply all the MFCC extraction to all the files. Subsequently, the files are converted into data frame consist of extracted MFCC features and the class classification, as well as label encoding to convert class to become a set of numerical values.

2.4. Train Test Split

The first step in creating a model is separating a dataset into training and testing sets, a prerequisite for all classification tasks. As depicted in Figure 5, the "train_test_split" function from the'sklearn.model_selection' module facilitates this procedure. The input data is denoted by 'X' and 'y', where 'X' represents the MFCC feature vectors or attributes of the audio samples and 'y' represents the corresponding target variables. By specifying a test size of 0.2 (20% of the entire dataset), the function separates the data into training and test sets.

```
##Train test split
#split data, test set 20%, 80% training set
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
```

Figure 5. Train test split from sklearn.model_selection.

The training and test data utilised in this study were created in a random manner from the whole dataset. The training dataset contains a total of 2,796 instances of baby cry sounds, representing 80% of the entire dataset. In contrast, the test set consists of 699 instances of baby cry sounds, which accounts for 20% of the overall dataset. The training and test datasets consist of baby cry sounds from various classes, ensuring a thorough and representative sample for assessing the effectiveness of algorithms and models in the classification and analysis of child cries.

2.5. Model Creation

Deep Learning strategies that have been proven to be highly effective in the field of image classification are proposed as a solution for this undertaking. MFCC are initially extracted from the sound samples. It summarizes the relationship between the perceived frequencies of an audio sample and the actual measured frequency values, enabling it to analyse the sample's temporal and frequency properties. These audible representations aid in differentiating characteristics required for classification.

Next are the outlines of the model classification's architecture, constructed using the keras library and the 'sequential' class. It consists of multiple layers, with the activation function 'relu' being used to induce non-linearity and an input shape of (40,), followed by a dropout layer with a dropout rate of 0.3. The final layer is a dense layer containing 'num_labels' units, which correspond to the number of classes in the audio classification assignment. The 'softmax' activation function is used to generate probability scores for each class, which represent the model's predicted probabilities for each descriptor. The training lasts for a variable number of epochs, the adam optimizer and the categorical cross-entropy loss function are employed, and the batch size for each dataset is 32.

This study examines the two configurations of the dense layer presented in Table 2 in order to determine the optimal configuration for obtaining the highest accuracy and performance in the CNN model. In the first configuration, 100, 200, and 100 neurons are allocated to each layer, whereas in the second configuration, 256, 512, and 256 neurons are allocated to each layer. By comparing these configurations, we hope to identify the one that produces the highest levels of precision and overall model performance.

© 2024 Penerbit UTM Press. All rights reserved

Table 2. Different configuration of dense layer units in CNN model.

CNN Configuration	Dense layer		
	Layer 1	Layer 2	Layer 3
1st Config	100	200	100
2nd Config	256	512	256

2.6. Epoch

During the model training process, the choice of epoch refers to the number of times the model iterates over the entire training dataset. In this case, a total of 100 epochs were choose after several tries, meaning the model goes through the training data 100 times to adjust its parameters and improve its performance through repeated learning and optimization.

2.7. Confusion Matrix

A confusion matrix is a matrix that summarises the classification efficiency of an algorithm or model. It displays the predicted and actual classifications for a set of data elements organised into various classes. 'n', which represents the number of distinct classes, determines the size of the confusion matrix (29).

In the given scenario, the confusion matrix is of size n x n, with n being equal to 2. This indicates that the classification problem under examination involves two distinct classes. Each element within the matrix corresponds to a unique combination of predicted and actual classifications, offering valuable insights into the model's performance and its ability to accurately classify instances (29). Table 3 represents the confusion matrix for the case where n=2. The entries in the matrix have the following interpretations:

- ‘a’ represents the number of correct negative predictions.
- ‘b’ represents the number of incorrect positive predictions.
- ‘c’ represents the number of incorrect negative predictions.
- ‘d’ represents the number of correct positive predictions.

Table 3. Confusion matrix for two-class classification problem (29).

	Predicted Negative	Predicted Positive
Actual Negative	a	b
Actual Positive	c	d

3. RESULTS AND DISCUSSION

This section provides a comprehensive presentation of the results and discussion derived from the employed detailed methodology. The methodology concentrates on classifying a baby's specific conditions by analysing their cries in various settings and observing the maximum accuracy. This section provides a comprehensive performance evaluation of the techniques discussed previously in the context of classifying baby sounds, taking into account various configurations.

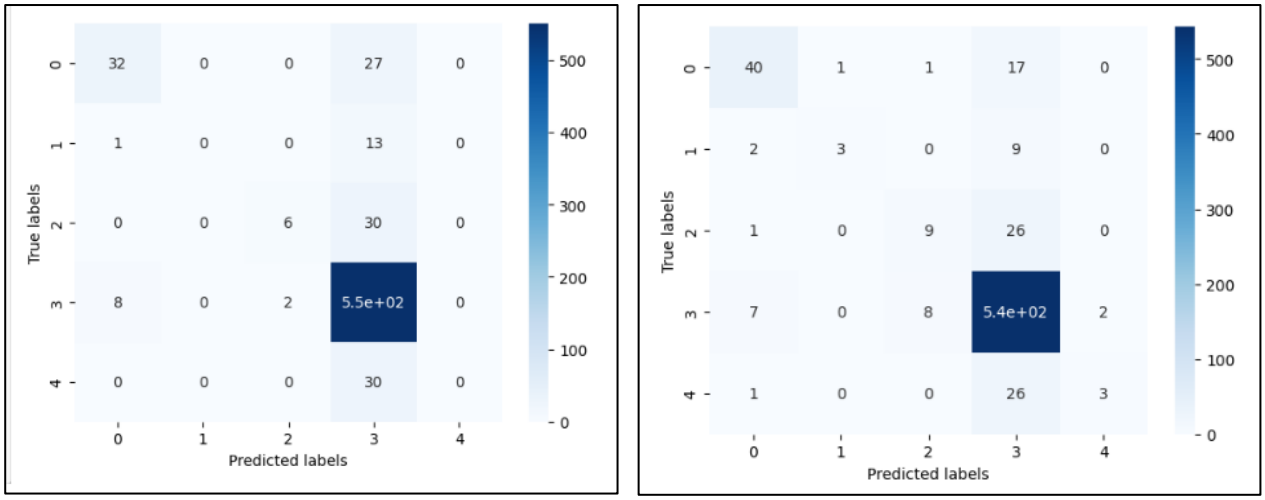
The study employed two distinct CNN model configurations to evaluate their efficacy in classifying baby sounds. The initial configuration consists of three CNN layers followed by dense layers containing 100, 200, and 300 neurons, respectively. The second configuration consisted of CNN layers with 256, 512, and 256 neurons, followed by layers with extensive connections. The confusion matrix is in numerical labels to represent different classes in the classification, as shown in Table 4.

The primary metric used to measure the performance of the models was accuracy, which indicates the proportion of instances that were correctly classified. The accuracy of the initial CNN configuration with three CNN layers and dense layers (100, 200, 300) was 84%. This indicates that 84% of the infant sounds in the dataset were correctly classified by the model. To acquire a deeper understanding of the performance, we utilised a confusion matrix to analyse the results.

Table 4. Label description for confusion matrix.

Number	**0**	**1**	**2**	**3**	**4**
Label	Belly pain	burping	Discomfort	Hungry	tired

© 2024 Penerbit UTM Press. All rights reserved



(a) (b)

Figure 8. (a) 1st Configuration result and (b) 2nd configuration result.

The confusion matrix illustrated the distribution of predicted and actual classes for the classification of baby sounds. Each row represents instances belonging to an actual class, while each column represents instances belonging to a predicted class. We observed, based on the confusion matrix, that the model performed well for some classes particularly. For instance, the model was highly accurate at classifying cries of hunger. The effectiveness of the second CNN configuration with CNN layers (256, 512, 256) followed by dense layers was superior to the first CNN configuration. It obtained an accuracy of 86%, signifying that it correctly categorized 86% of the baby sounds. Similar to the initial configuration, we utilised a confusion matrix to obtain insight into the performance of the model.

Compared to the first configuration, the second configuration's confusion matrix revealed an overall improvement in classification accuracy across all classes. The model performed well, resulting in fewer incorrect classifications. In the second configuration, the increased depth and complexity of the CNN layers appeared to have contributed to the improved accuracy. Overall, both configurations demonstrated promising results in classifying baby sounds. The second configuration, with its higher accuracy of 86%, outperformed the first configuration. This suggests that deeper and more complex CNN architectures can capture more intricate patterns in the baby sound data, leading to improved classification accuracy.

4. CONCLUSION

This project demonstrated the effectiveness of convolutional neural networks (CNNs) in classifying infant cries. By evaluating two distinct CNN configurations, this research was able to classify cry sounds accurately. Using a split of 80% training data and 20% testing data, obtained an astounding 86% accuracy, with the second configuration outperforming the first. According to the analysis of the confusion matrix, this highlights the significance of model architecture in achieving higher classification precision. However, it is important to recognize the limitations of this study, particularly in terms of the dataset's size and diversity. To improve the generalizability of the models, future research should investigate utilizing larger and more diverse data sets.

ACKNOWLEDGMENT

The authors would like to acknowledge Ministry of Higher Education, Universiti Teknologi Malaysia, and Research Management Centre for the support of this work under UTM Encouragement Research Grant with grant number Q.J130000.3851.20J74 and Q.J130000.3851.20J51. The Baby Chillanto Data Base is a property of the Instituto Nacional de Astrofisica Optica y Electronica – CONACYT, Mexico. We would like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their dedication of the collection of the Infant Cry database.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

(1) Liu L, Li W, Wu X, Zhou BX. Infant cry language analysis and recognition: An experimental approach. IEEE/CAA J. Autom. Sin. 2019; 6(3):778–788. https://doi.org/10.1109/JAS.2019.1911435.
(2) Zakaria NH, Phang FA, Pusppanathan J. Physics on the go: A mobile computer-based physics laboratory for learning forces and motion. Int. J. Emerg. Technol. Learn. 2019; 14 (24):167. https://doi.org/10.3991/ijet.v14i24.12063.

© 2024 Penerbit UTM Press. All rights reserved

(3) Lahmiri S, Tadj C, Gargour C, Bekiros S. Deep learning systems for automatic diagnosis of infant cry signals. Chaos Solit Fractals. 2022; 154:111700. https://doi.org/10.1016/j.chaos.2021.111700.
(4) Manfredi C, Bandini A, Melino D, Viellevoye R, Kalenga M, Orlandi S. Automated detection and classification of basic shapes of newborn cry melody. Biomed Signal Process Control. 2018; 45:174–181. https://doi.org/10.1016/j.bspc.2018.05.033.
(5) Lahmiri S, Tadj C, Gargour C, Bekiros S. Characterization of infant healthy and pathological cry signals in cepstrum domain based on approximate entropy and correlation dimension. Chaos Solitons Fractals. 2021; 143:110639. https://doi.org/10.1016/j.chaos.2020.110639.
(6) Severini M, Ferretti D, Principi E, Squartini S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. IEEE Access. 2019; 7:51982–51993. https://doi.org/10.1109/ACCESS.2019.2911427.
(7) Herencsar N, Benedetto F, Crichigno J. Special issue "Selected papers from the 2019 42nd International Conference on Telecommunications and Signal Processing (TSP)". Appl Sci. 2020; 10(6):2088. https://doi.org/10.3390/app10062088.
(8) Ntalampiras S. Audio pattern recognition of baby crying sound events. J Audio Eng Soc. 2015; 63(5):358–369. https://doi.org/10.17743/jaes.2015.0025.
(9) Șandru ED, Buzo A, Cucu H, Burileanu C. Recent experiments and findings in baby cry classification. In: Lecture notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering. Springer Verlag: Berlin; 2018. p. 253–260. https://doi.org/10.1007/978-3-319-92213-3_37
(10) Felipe GZ, Aguiar RL, Costa YMG, Silla CN, Brahnam S, Nanni L, McMurtrey S. Identification of infants' cry motivation using spectrograms. 2019 Int Conf Syst Signals Image Process. 2019; 181–186. https://doi.org/10.1109/IWSSIP.2019.8787318.
(11) Herencsar N, Benedetto F, Crichigno J. Special issue "Selected papers from the 2018 41st International Conference on Telecommunications and Signal Processing (TSP)". Appl Sci. 2018; 1-4. https://doi.org/10.3390/app9102056.
(12) Liu L, Li Y, Kuo K. Infant cry signal detection, pattern extraction and recognition. 2018 International Conference on Information and Computer Technologies, 2018; 159–163. https://doi.org/10.1109/INFOCT.2018.8356861.
(13) "Dunstan Baby Language: What Is It and Does It Work?" [Internet] [Cited 2022, Dec 12]. Available from: https://www.healthline.com/health/baby/dunstan-baby-language.
(14) Hamza A, Javed ARR, Iqbal F, Kryvinska N, Almadhor AS, Jalil Z, Borghol R. Deepfake audio detection via MFCC features using machine learning. IEEE Access. 2022; 10:134018–28. https://doi.org/10.1109/ACCESS.2022.3231480.
(15) Alsayaydeh JAJ, Indra WA, Khang AWY, Hossain AKMZ, Shkarupylo V, Pusppanathan J. The experimental studies of the automatic control methods of magnetic separators performance by magnetic product. ARPN J Eng Appl Sci. 2020; 15(7);922–927.
(16) Tong Y, Zhang X, Ge Y. Classification and recognition of underwater target based on MFCC feature extraction. 2020 IEEE Int. Conf. Signal Process. Commun. Comput. 2020; 1–4. https://doi.org/10.1109/ICSPCC50002.2020.9259457.
(17) Huang J-T, Li J, Gong Y. An analysis of convolutional neural networks for speech recognition, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. 2015; 4989–4993. https://doi.org/10.1109/ICASSP.2015.7178920.
(18) Kikuchi T, Ozasa Y. Watch, listen once, and sync: Audio-visual synchronization with multi-modal regression CNN. IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc. 2018; 3036–3040. https://doi.org/10.1109/ICASSP.2018.8461853.
(19) Sagar Sharma. What the hell is perceptron? The Fundamentals of Neural Networks. Towards Data Science [internet]. 2017. [Cited 2022 Dec 14] Available from: https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53.
(20) Selfridge OG. Pattern recognition and modern computers. AFIPS '55 (Western): Proceedings of the March 1-3, 1955, Western Joint Computer Conference. 1955; 91–93. https://doi.org/10.1145/1455292.1455310.
(21) Liang YC, Wijaya I, Yang MT, Cuevas Juarez JR, Chang HT. Deep learning for infant cry recognition. Int J Environ Res Public Health. 2022; 19(10). https://doi.org/10.3390/ijerph19106311.
(22) Zan T, Wang H, Wang M, Liu Z, Gao X. Application of multi-dimension input convolutional neural network in fault diagnosis of rolling bearings. Appl Sci. 2019; 9(13). https://doi.org/10.3390/app9132690.
(23) Mustaqeem, Kwon S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors. 2019; 20(1):183. https://doi.org/10.3390/s20010183.
(24) Pelchat N, Gelowitz CM. Neural network music genre classification. Canadian J Elec Comp Eng. 2020; 43(3):170–173. https://doi.org/10.1109/CJECE.2020.2970144.
(25) Open Data Commons Open Database License (ODbL) v1.0 [Internet]. Open Data Commons. [Accessed May 24, 2023] Available from: https://github.com/gveres/donateacry-corpus/tree/master/donateacry_corpus_cleaned_and_updated_data.
(26) Reyes-Galaviz OF, Cano-Ortiz SD, Reyes-García CA. Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. 2008 7th Mexican International Conference on Artificial Intelligence - Proceedings of the Special Session. 2008; 330–335. https://doi.org/10.1109/MICAI.2008.73.
(27) Bascuñana J, León S, González-Miquel M, González EJ, Ramírez J. Impact of Jupyter Notebook as a tool to enhance the learning process in chemical engineering modules. Educ Chem Eng. 2023; 44:155–163. https://doi.org/10.1016/j.ece.2023.06.001.
(28) Mottaghi-Kashtiban M, Farazi S, Shayesteh MG. Optimum structures for sample rate conversion from CD to DAT and DAT to CD using multistage interpolation and decimation. Sixth IEEE International Symposium on Signal Processing and Information Technology. Institute of Electrical and Electronics Engineers Inc. 2006; 633–637. https://doi.org/10.1109/ISSPIT.2006.270877.
(29) Visa S, Ramsay B, Ralescu AL, van der Knaap E. Confusion matrix-based feature selection. Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference. 2011; 710(1):120–127.

© 2024 Penerbit UTM Press. All rights reserved