Proceedings of the 9th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. ICBES 133
DOI: 10.11159/icbes23.133

Infant Cry Signal Detection And Classification Using Deep Learning

Omnia Badr eldine[1], Nagia Ghanem[2], Mohamed Selim[2], Nagwa El-Makky[2]
[1] Biomedical engineering department -Medical Research Institute, Alexandria University
Alexandria, Egypt
[2]Computer and systems engineering department-Faculty of Engineering-Alexandria University
Alexandria , Egypt
Omniabadr92@alexu.edu.eg
nagia.ghanem, mohammed.selim, nagwamakky@ alexu.edu.eg

Abstract - Detection of infant cries in noisy environments such as homes, hospitals and clinics is vital to determine the reason of baby's cry. Also, It is crucial to classify the detected cry signals into normal or pathological cries especially in the first months of the baby life. This paper proposes a deep learning automatic infant cry detection and classification system under noisy conditions. It classifies the detected cry signals into normal , asphyxia and deaf cry signals . The overall system is composed of two stages ; cry detection stage and cry classification stage. In first stage, features(Mel-frequency cepstrum coefficients MFCC) are extracted from audio signals collected from a daily life dataset and passed into a 2D-two layers convolutional neural network(2DCNN) to be classified into cry and non cry signals. In second stage , a 2D-three layers CNN is used to classify cry signals collected from dataset with cry segments only into Normal (N) , Asphyxia (A) and Deaf (D) signals according to extracted MFCC features . In first stage, Testing results show that the cry detection system reaches an accuracy of 99.59 % for classifying the signals into cry and non-cry . In second stage, Due to the lack of pathological cry signals datasets that are collected in noisy environments, we added different levels of white noise to the training dataset. This way, we were able to get more realistic results. In particular, our cry classification system achieves accuracy of 91.3% ,94.2% , 95.07 % (under white noise) with signal-to-noise ratio(SNR) of 5db, 10db and 15db, respectively.

Keywords*: *Infant Cry; Asphyxia; Deaf ; Mel Frequency Cepstrum ; 2DCNN;SNR

1. Introduction

Crying is the first sign of life at birth. It is the easiest way to determine the survival state, health and development condition of the infant[1].The purpose of cry detection algorithms is to efficiently identify infant cries in a variety of settings, e.g., a car, house, hospital when other sounds are occurring at the same time[2]. For several years, classification of infant cry signal played an important role in detecting several diseases such as asphyxia , deafness , apnea, hypothyroidism ,etc[2]. Asphyxia is a disease caused by malfunctioning of the central nervous system and occurs when a baby doesn't receive enough oxygen before, during or just after birth. It was found that asphyxia affects the sound of infant cry which makes it vital to early detect the asphyxiated cry signals[3]. Hearing disorders affect the performance of child's learning and development stages, especially if not recovered at early stages[4].

1.1. Literature review (recent first)

Cry sound detection techniques

In [5] X.Yao et al. developed four different detection models. The machine learning model that makes use of both deep spectrum and acoustic information achieves the best F1 score 0.613 .They applied continuous signal segmenting using sliding window, The prediction for each second was labelled "crying" if it appeared in any sliding window of 1 sec containing that second. In [6] L.Sze et al. proposed algorithm to detect cry in real noisy environment depending on acoustic features. Five acoustic features ; average frequency, pitch frequency, short-time energy (STE) acceleration, zero-crossing rate (ZCR), and Mel-Frequency cepstral coefficients (MFCC). Using noisy crying and non-crying samples, the algorithm was tested, and an accuracy of 89.20% was achieved for the offline testing. Also, three specially created noisy samples that included both crying and non-crying segments were used in [6] in online tests. The online accuracy was 80.77%. In [7] S.dewi et al. extract 19 linear frequency cepstral coefficients used as input to KNN classifier to determine whether the baby is crying or not. They achieved 90% as best detection .In [8] K. Manikanta et al. introduce a baby cry sound detection (BCSD) approach with various background noise types. MFCC coefficients and 3 machine learning classifiers 1D-CNN, feed-forward neural networks (FFNNs), and multi-class support vector machines were developed for cry sound detection. The 1DCNN, FFNN,

and SVM models achieved accuracy of 98.86%, 98.46%, and 97.97%, respectively, for detecting cry sounds for frames of 500 milliseconds.

Cry classification techniques

In [9] C.JI proposed a method for accurately classifying asphyxiated baby crying by creating weighted prosodic features paired with acoustic features to create a merged feature matrix. The MFCC and prosodic features are concatenated and used as input to a deep learning neural network to classify the cries into normal or asphyxiated. The testing accuracy achieved 96.74%. In [10] A.K et al. classify neonatal cries into pain, hunger, and sleepiness. The spectrogram of the neonatal cries is fed into the DCNN. Extracted features from the fully connected layer of the DCNN are used as input to a support vector machine (SVM) classifier. In this paper, the SVM-RBF achieves the best classification accuracy of 88.89%. In [11] C.Chang and L.Tsai applied two stages techniques; cry detection and classification .In first stage, the spectrum of baby cry signals is used as input to a 2DCNN to detect the cry signal. In second stage, 1DCNN is used to classify the detected cry into four categories (pain, hunger, tiredness, and damp diaper). The detection accuracy is 99.83%. The classification model achieved a recognition accuracy of 78.27%.

In contrast, Our work proposes a two phases infant cry detection and classification system in a noisy environment based on deep learning techniques. In the cry detection phase, MFCCs coefficients are extracted from the audio recordings and used as input to a two layers convolutional neural network. The CNN classifies the signals into cry and non-cry signals. In the classification phase, The extracted MFCCs of the detected cry signal are fed into three layers CNN to classify the detected cry signal into normal, asphyxia and deaf cry signal under different noise levels.

2. DATASETS AND MODELS

2.1. Datasets

In this work, two different datasets are used , deBarbaro cry corpus [12] used in [5] and the Baby Chillanto dataset [13]. DeBarbaro cry corpus is used in training the model in the cry detection phase and Baby chillanto database is used in training the model in the cry classification phase. The deBarbaro cry corpus consists of audio recordings of twenty-two infants aged 1-10 months [14]. A 16-bit PCM channel sampled at 22.05kHz was used to capture the audio[15]. The noise of daily life was captured in this authentic audio recordings. Each episode of crying was divided into five second segments (with four-second overlap between neighbouring segments). The whole data collection contains approximately 61h of annotated data, including nearly seven hours of unique crying data. Each 5 second training segment is labelled with either “crying” or “not crying”. We used Debarbaro cry dataset in our work because it is rich, recent and collected in noisy environment.

The Baby Chillanto database, is a property of the Instituto Nacional de Astrofisica Optica y Electronica (INAOE) – CONACYT, Mexico [13]. The database consists of different categories of infant cry signals: Normal (N) ,deaf (D) , asphyxia (A) , hunger (H) and pain (P) cry signals. Each cry signal is divided into non overlapping segments of 1 sec length .The signals were recorded of newborn up to 6 months old babies by specialized physicians and were sampled at different sampling frequencies (8 kHz, 11.025 kHz and 22 kHz). The Baby Chillanto database is widely used in the pathological cry classification research.

TABLE 1 .STATISTICS OF BABY CHILLANTO DATABASE

Infant Cry Types	Number of 1 sec samples	Number of babies
Normal (N)	507	5
Asphyxia(A)	340	6
Deaf(D)	885	6
Hunger(H)	350	32
Pain(P)	192	21

2.2. Methodology

Figure 1 illustrates the complete infant cry detection-classification system. Briefly the system is composed of two stages; "Stage1. cry detection stage" and "Stage 2. cry classification stage". In Stage1, 5 sec audio segments from the deBarbaro cry dataset are used to train the model. Mel frequency cepstrum coefficients (MFCCs) are extracted from the audio segments. These coefficients are passed to a two layers convolutional neural network (CNN) to classify signals into cry and non-cry . In stage 2, 1 sec audio cry signals from Baby Chillanto database, augmented by white noise, are used to extract the MFCC features. The MFCC coefficients are used as input to a three layers CNN model. The signals are classified into normal (N), asphyxia (A) and deaf (D).

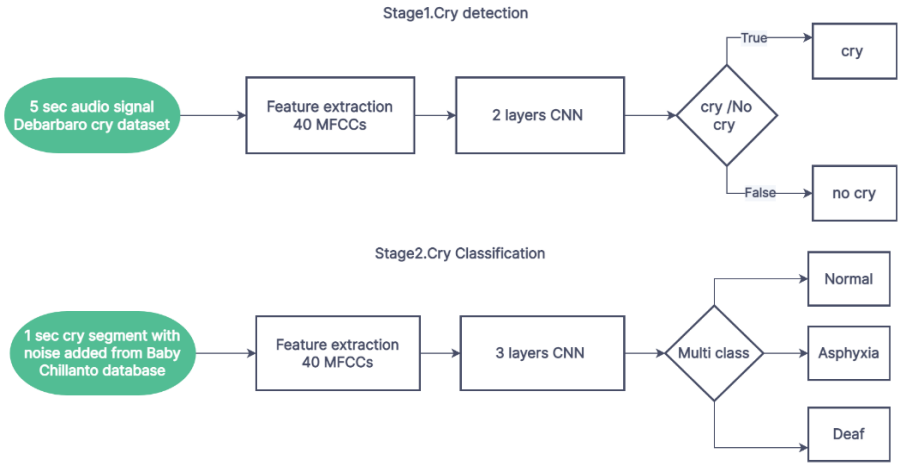

Fig.1 :The complete infant cry detection-classification system architecture

2.3. Mel frequency cepstum coefficients (MFCC)

The Mel frequency detects frequencies below 1 KHz in a linear scale and detects frequencies above 1 KHz in a logarithmic scale[16]. The Mel scale for a given frequency (f) in Hz is calculated as shown in equation (1):

$$mel(f) = 2952log_{10}(1 + \frac{f}{700}) \quad (1)$$

In MFCC, the signal is divided into short frames where each frame is multiplied by hamming window to overcome the discontinuity of the signal. Fast Fourier transform (FFT) is used to represent the windowed signal into the frequency domain. The power spectrum obtained from the FFT is mapped onto the Mel filter bank by a set of triangular overlapped filters banks shown in figure 2[17]. Discrete cosine transform (DCT) is applied to the log of Mel filter banks for representing the signal's feature into the time domain, then the mel coefficients are generated[18]. MFCC can be computed using (2):

$$C_n = \sum_{i=1}^{k}(logE_i)\cos[n(i-0.5)\frac{\pi}{k}] \quad (2)$$

Where C_n refers to the MFCC, n=1,2,…M. M represents the total number of coefficients, E_i represents the logarithmic energy output for the i^{th}Mel filter and K is the total number of filters [19].

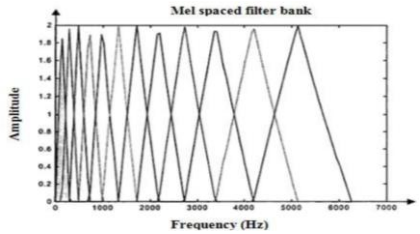

Fig.2 :Example of Triangular Mel spaced filter banks[17]

2.4. Stage 1. Cry detection

Stage 1 shown in Figure 3 is the cry detection stage which is implemented as a binary classifier. The method takes input audio recordings(deBarbaro dataset) and extracts MFCC coefficients of each input. The extracted features are to a two layers 2D-CNN to be classified into cry and non cry samples. 40 MFCCs coefficients are extracted for each audio signal. The chosen frame duration value is 50 msec, successive frames are overlapped by 25% and a Mel filter of 24 filters is used. The audio signals are resampled at 8000 Hz. These coefficients are taken as input to the CNN network and are resized before entering the CNN into (40,134,1) to transform the audio features to images. The deep learning network consists of 2 convolutional layers followed by a fully connected layer and the output layer. In the first convolutional layer , we used 120 filters , kernel size of (10,2) , with activation function "Relu". Maximum pooling is done with the size of 2 * 1 and with stride (2,2) . In the Second convolutional layer, we used 100 filters , kernel size of (6,2) , with activation function "Relu". Maximum pooling is done with size of 2*1 and with strides (2,2). The output of the convolutional layers is flattened before passing to the fully connected dense layer. To prevent overfitting ,a dropout layer is added after each convolution layer. The typical rectifier, which corresponds to rectified linear unit (ReLU) layers, serves as the CNN's activation function and "Adam optimization" is used.

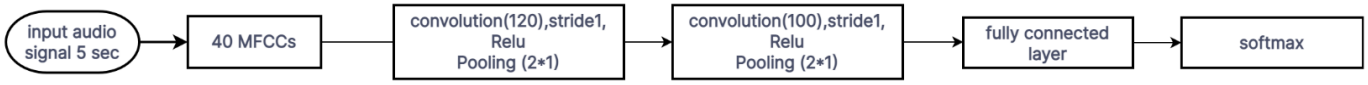

Fig.3: block diagram of the CNN architecture used in the cry detection model.

2.5. Stage 2 .Cry Classification

In stage 2 shown in figure 4, Infant cry signals are classified into 3 classes :normal (N) , asphyxia (A)and deaf (D) based on MFCC features extracted from Baby Chillanto dataset augmented by adding white noise before extracting the features as follows: Noise signals are added to study the performance of the model in environments with different signal-to-noise ratios (SNR) mimicking the real environment. The $RMS_{noise}$is calculated according to the chosen SNR by equation(3) [20]. In this work, the chosen values of SNR are (5db for high noise level , 10 for medium noise level and 15 for low noise level) and each experiment is performed under the three levels of noise.

$$SNR = 10\log\left(\frac{RMS_{signal}^2}{RMS_{noise}^2}\right) \quad (3)$$

where $RMS_{signal}$is the root mean square value of the signal and RMS_{noise} is the root mean square of the noise.

In this work, 40 MFCC coefficients are extracted for each input cry signal. The chosen frame duration value is 50 msec, successive frames are overlapped by 25% and a Mel filter bank of 24 filters is used. The cry signals are resampled at 8000 Hz. These coefficients are used as input to the CNN network and are resized before entering the CNN into (40,27,1). The 40 MFCCs coefficients are fed to the CNN network. The deep learning network consists of 2D three convolutional layers followed by a fully connected layer and the output layer. The structure of the first 2 layers of the network is same as the network in stage1 explained in section 2.4 . In the third convolutional layer , we used 50 filters , kernel size of (3,2) , with activation function "Relu". Maximum pooling is done with size of 2*1 and with strides (2,2).The output layer is modified by adding kernel regularizer L2 to avoid the overfitting of neural networks , thus the networks will converge faster during the training process and update the weights in less time. Adding "L2" Regularization in last layer tends to improve the model[21] ."Adam optimization" is used for optimization.

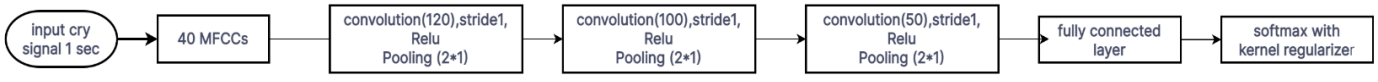

Fig.4: block diagram of the CNN architecture used in the cry classification model

We have applied conventional validation to assess the reliability of the classification techniques used in the two stages. Namely, 80 % of data are used for training and the remaining 20 % are used for testing. To analyse the performance of the proposed model , several performance metrics extracted from the confusion matrix are calculated for each experiment[22]

a) Overall accuracy**:** is the number of the correct predictions made by the proposed algorithm over all signals.

$$overall\ accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

b)Precision :is a measure of a classifier exactness indicating the proportion of true positive samples over the total predicted positive observations.

$$Precision\ = \frac{TP}{TP+FP} \quad (5)$$

c)F1 score :is the weighted average of the precision and recall (or sensitivity). It is a measure of the test accuracy .

$$F1\ score = 2 * \frac{(Recall*Precision)}{Recall+Precision} \quad (6)$$

d)Recall : is the percentage of true positives that a diagnostic test successfully detects

$$Recall\ = \frac{TP}{TP+FN} \quad (7)$$

The attributes used are TP (true positive) , TN (true negative), FP (False positive) and FN (False negative)

3. Results and discussion

3.1. Cry detection stage

In the detection stage, samples are used from deBarbaro cry corpus[12] . To maintain the balance of the set during network training, 603 crying samples and 603 non crying samples are randomly selected. The experiments were conducted using keras python library by google COLAB. 40 MFCCs coefficients extracted from the input audio signals are used as input to stage1 - CNN model. Using 30 epochs and batch size 256 , the F1 score obtained is 1 (accuracy =99.59%) while the proposed system in [5] achieves F1 score 0.613 in detecting the cry events per second using same dataset. Our system outperforms the approaches proposed in [6], [7] and [8] achieving (89.2 %, 90 % and 98.8%) using different datasets.

TABLE 2. PERFORMANCE RESULTS OF THE CRY DETECTION SYSTEM

	Precision	Recall	F1 score
Cry signals	1	0.99	1
Non cry signals	0.99	1	1

3.2. Cry Classification stage

In the classification stage, samples are used from Baby Chillanto database[13]. All samples of normal and asphyxia classes are used, for deaf we used 879 samples. Noise signals are added to each cry sample. In this work, we have performed two experiments . Number of CNN layers are changed in each experiment to determine the best accuracy achieved. In all experiments, we used 100 epochs, batch size of 256 , L2 of 0.3 and 40 MFCCs coefficients. In first experiment, MFCCs coefficients extracted from the input cry signals are used as input to the stage2-CNN model with 3 layers. The accuracy obtained is 91.3 % ,94.2% , 95.07 % under white noise's SNR of 5db, 10db and 15db, respectively. In second experiment, number of CNN layers is changed to only 2 layers. The accuracy obtained is 89.28 %, 91.3 %,92.46 % under white noise's SNR of 5db, 10db and 15db, respectively. It is clear that the accuracy of the 2 layers is significantly than 3 layers.

3.3. Comparison with previous work

By comparing our cry classification model to the previous proposed systems([9] to [11]) , the low level noise's accuracy is close to results obtained in [9] which achieve accuracy of 96.7% (noise free and same dataset) . Our model outperforms the proposed algorithm in [10] (noise free and different dataset) under the low , medium and high level's noise. Finally, Our model significantly outperforms the algorithm in [11] (using different dataset).

TABLE 3. PERFORMANCE RESULTS OF THE 3 LAYERS CNN MODEL WITH DIFFERENT NOISE LEVELS

SNR=5db				SNR=10db				SNR= 15db			
Accuracy=91.3%				Accuracy=94.2%				Accuracy=95.07%			
	Precision	Recall	F1 score		Precision	Recall	F1 score		Precision	Recall	F1 score
N	0.87	0.90	0.88	N	0.94	0.86	0.90	N	0.89	0.91	0.90
A	0.96	0.95	0.95	A	0.98	0.98	0.98	A	0.97	0.99	0.98
D	0.87	0.86	0.87	D	0.89	0.95	0.92	D	0.96	0.92	0.94

TABLE 4. PERFORMANCE RESULTS OF THE 2 LAYERS CNN MODEL WITH DIFFERENT NOISE LEVELS

SNR=5db				SNR=10db				SNR=15db			
Accuracy=89.28%				Accuracy=91.3%				Accuracy=92.46%			
	Precision	Recall	F1 score		Precision	Recall	F1 score		Precision	Recall	F1 score
N	0.74	0.83	0.78	N	0.83	0.81	0.82	N	0.77	0.96	0.86
A	0.97	0.98	0.97	A	0.98	0.98	0.98	A	0.99	0.95	0.97
D	0.87	0.79	0.83	D	0.85	0.87	0.86	D	0.93	0.87	0.90

4. Conclusion and Future work

In this paper, we have presented an infant cry detection and classification system. The system classifies detected cry signals into three categories(Normal ,Asphyxia, and deaf) using deep learning algorithms based on MFCC features. The detection stage achieved 99.5 %.

Due to the lack of pathological cry signals datasets that are collected in noisy environments, we augmented the dataset used by white noise of different levels. This leads to more reliable results. Different experiments are implemented by changing number of CNN layers in order to achieve the best classification accuracy. The best accuracy obtained is by using 40 MFCCs as input to 3 layers of CNN achieving 91.3% ,94.2% and 95.07% under signal-to-noise ratio(SNR) of 5db ,10db and 15db, respectively. In future work, we intend to experiment with different neural network topologies and try data augmentation techniques. Also, we intend to design and implement a prototype of a hardware medical device. Finally, we encourage researchers to publish and share pathological cry signals datasets in noisy environments. This will enable more realistic studies.

Acknowledgements

Thanks to Lara Andres, Nina Nariman, Brooke Benson, and Kara Kaur for their dedication of the collection of DeBarbaro cry cropus[12]. We 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 Baby Chillanto database[13].

References

[1] Y. Kheddache and C. Tadj, "Frequential Characterization of Healthy and Pathologic Newborns Cries," *American Journal of Biomedical Engineering*, vol. 3, no. 6, pp. 182–193, 2013.

[2] 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, 2021.

[3] L. J. Millar, L. Shi, A. Hoerder-Suabedissen, and Z. Molnár, "Neonatal Hypoxia Ischaemia: Mechanisms, Models, and Therapeutic Challenges," *Front. Cell. Neurosci.*, vol. 11, p. 78, May 2017.

[4] Jr.Va´rallyay, “The melody of crying,” International Journal of Pediatric Otorhinolaryngology, Vol 71 , pp. 1699-1708 ,2007.
[5] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro, “Infant crying detection in real-world environments,” *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 131-135.
[6] L. S. Foo, W.-S. Yap, Y. C. Hum, Z. Kadim, H. W. Hon, and Y. Kai Tee, “Real-time baby crying detection in the noisy everyday environment,” *in Proceedings of the 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Shah Alam, Malaysia, 2020, pp. 26-31.
[7] S. P. Dewi, A. L. Prasasti, and B. Irawan, “Analysis of LFCC feature extraction in Baby Crying Classification using KNN,” *in Proceedings of the IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, Bali, Indonesia, 2019, pp. 86-91.
[8] K. Manikanta, K. P. Soman, and M. S. Manikandan, “Deep Learning based effective baby crying recognition method under indoor background sound environments,” in *Proceedings of the 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, India, 2019, pp. 1-6.
[9] Ji. Chunyan, "Infant Cry Signal Processing, Analysis, and Classification with Artificial Neural Networks," Ph.D. Dissertation, Dept. Comp. Science., Georgia State University.
[10] A. K, P. M. Vincent, K. Srinivasan, and C.-Y. Chang, “Deep learning assisted neonatal cry classification via support Vector Machine Models,” *Frontiers in Public Health*, vol. 9, 2021.
[11] C.Y. Chang and L.Y. Tsai, “A CNN-based method for infant cry detection and recognition,” *Advances in Intelligent Systems and Computing*, pp. 786–792, 2019.
[12] K. de Barbaro .(2022) [Online]. Available: https://homebank.talkbank.org/access/Password/deBarbaroCry.html
[13] O. F. Reyes-galaviz, U. Aut, M. Tlaxcala, S. D. Cano-ortiz, and C. A. Reyes-garc, “Evolutionary-Neural System to Classify Infant Cry Units for Pathologies Identification in Recently Born Babies,” Proc. Seventh Mexican International Conference on Artificial Intelligence, pp. 330–335, 2008.
[14] A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson, “A thorough evaluation of the Language Environment Analysis (LENA) system,” *Behavior Research Methods*, vol. 53, no. 2, pp. 467–486, 2020.
[15] M. Ford, C. Baer, D. Xu, U. Yapanel, and S. Gray, “The lena language environment analysis system: Audio specifications of the dlp-0121,” Sep 2008.
[16] A. Chittora and H. A. Patil, “Spectral analysis of infant cries and adult speech,” Int. J. Speech Technol., vol. 19, no. 4, pp. 841–856, 2016.
[17] S. K. Saksamudre and R. R. Deshmukh, “Comparative study of isolated word recognition system for Hindi language,” *Int. J. Eng. Res. Technol. (Ahmedabad)*, vol. V4, no. 07, 2015
[18] S. Gupta, J. Jaafar, and A. Bansal, “Feature extraction using MFCC,” Signal & Image Processing : An International Journal (SIPIJ), vol. 4, no. 4, pp. 101–108, 2013.
[19] N. S. A. Wahid, P. Saad, and M. Hariharan, “Akademia Baru Automatic Infant Cry Classification Using Radial Basis Function Network Akademia Baru,” vol. 4, no. 1, pp. 12–28, 2016.
[20] L. N. Wijayasingha.(2021,Jan 29). Adding noise to audio clips. [Online]. Available: https://medium.com/analytics-vidhya/adding-noise-to-audio-clips-5d8cee24ccb8.
[21] J. Brownlee.(2020,Aug 25). How to use weight decay to reduce overfitting of neural network in Keras.[Online]. Available:https://machinelearningmastery.com/how-to-reduce-overfitting-in-deep-learning-with-weight-regularization/.
[22] W. Zhu, N. Zeng, and N. Wang, “Sensitivity , Specificity , Accuracy , Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations,” *in Proceedings of the NESUG: Health Care and Life Sciences*, 2010, pp. 1–9.