Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

DETECTION OF ASPHYXIA IN INFANTS USING DEEP LEARNING CONVOLUTIONAL NEURAL NETWORK (CNN) TRAINED ON MEL FREQUENCY CEPSTRUM COEFFICIENT (MFCC) FEATURES EXTRACTED FROM CRY SOUNDS

A. Zabidi[1], I. M. Yassin[1,*], H. A. Hassan[2], N. Ismail[1], M. M. A. M. Hamzah[1], Z. I. Rizman[3] and H. Z. Abidin[1]

[1]Faculty of Electrical Engineering, UniversitiTeknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[2]Faculty of Engineering, Universiti Selangor, 45600 Bestari Jaya, Selangor, Malaysia

[3]Faculty of Electrical Engineering, UniversitiTeknologi MARA, 23000 Dungun, Terengganu, Malaysia

Published online: 10 September 2017

ABSTRACT

Deep Learning Neural Network (DLNN), is a new branch of machine learning with the ability for complex feature representation compared to traditional 4th-generation neural networks. Although it was mainly suited for image feature (since it was inspired by object recognition method of mammalian visual system), if any type of feature can be translate into image, other type of data could be fit for using DLNN. In this paper, we prove that Mel Frequency Cepstrum Coefficient (MFCC) feature generates from audio signal of infant cry could be used as input feature for the Convolution Neural Network (CNN).

Author Correspondence, e-mail: ihsan.yassin@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i3s.59

The result shows CNN can be used to classify between normal and pathological (asphyxiated) cry with 94.3% accuracy in training set and 92.8% accuracy in testing set.

Keywords: Deep Learning Neural Network (DLNN); Convolution Neural Network (CNN); Mel Frequency Cepstrum Coefficient (MFCC).

1. INTRODUCTION

Deep Learning Neural Network (DLNN) is a branch of machine learning [1] with the ability for complex feature representation relative to current fourth-generation neural networks. DLNN was proposed by [4] as an improvement to the conventional fourth-generation neural networks. The DLNN is inspired by object recognition method of the mammalian visual system in which information entering the retina to visual center undergoes several cascading layers that sequentially extract edge, part, shape features and then finally forming an abstract based on those features. Similarly, DLNN extract features layer-by-layer and combines low-level feature to become high-level feature during each layer's processes [2]. The model makes DLNN particularly excellent in problems such as image classification, object detection and semantic segmentation [3].

There are several types of DLNNs namely Deep Belief Neural Network (DBNN), Stacked Auto-Encoders (SAE) and Convolution Neural Network (CNN). Comparing the training methods with Multilayer Perceptron (MLP), DBNN and SAE used greedy layer-wise algorithms (e.g. Restricted Boltzmann Machine (RBM) for DBNN and Auto-Encoders for SAE) while MLP used initials random matrix in its weight [2]. The training for CNNs is relatively like MLP (Gradient Descent Back-Propagation). The work in this paper focuses on CNN as it was reported as an excellent tool in the deep learning framework [5].

Due to CNN's excellent visual-spatial recognition ability, the application areas of CNN is very broad. One of its famous usage examples is handwriting recognition [5]. CNN have also been explored for medical research purposes. For example, in [6-7], deep learning was applied on Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) images.

Although more suited for images, if other types of features can be represented like that of an image, other applications are also possible (for example, sound). In [8], a CNN was trained to

classify the sound of heart sound phonocardiogram. The MFCC algorithm was used to discover a-time frequency representations to be used as features for the CNN. The CNN structure consisted of two convolutional-pooling layer pairs and a fully-connected MLP at the back. The optimal parameters for the convolutional filters and strides were done using a combination of manual testing and random search. The authors reported good classification accuracy (above 80%), showing the CNN's capability for automatic feature extraction and ability of MFCC to distinguish between normal and abnormal heart sound from noisy data.

Mel Frequency Cepstrum Coefficient (MFCC) features generate two-dimensional features similar to images. MFCC also has been used as an input feature for CNN for baby cry detection in domestic environments [9] with 82.5% accuracy. Similarly, in [10], a CNN was trained to recognize three types of infant cries (hunger, pain and drowsiness) using spectrograms as the inputs. Results showed that the CNN managed to obtain 78.5% validation accuracy when trained for 25,000 iterations. A common pattern for all research using sound inputs was that all sound features could be fed to CNN if it can be represented as an image.

In this paper, we demonstrate the ability of the CNN to extract and classify features from MFCC features for detection of infants suffering from asphyxia based on their crying sound patterns. MFCC features was extracted from infant cry sounds and fed to the CNN. As will be shown in the results, the CNN managed to classify the feature accurately with above 90% accuracy. The remainder of this paper is as follows: Section II presents the methodology used, followed by the results and discussions in section III. Finally, concluding remarks are presented in section IV.

2. METHODOLOGY

2.1. Data Collection

Two type of dataset needed for classification purpose. Since the objective in this work is to detect asphyxiated baby from their cry signal, the normal cry (without any pathological condition) also needs to be collected. By this two signal, classification can be made using CNN. For normal cry, the dataset were obtained from the InstitutoNacional de Astrofísica, Óptica and for asphyxia the dataset was obtained from University of Milano-Bicocca. All the

signal came in form of Microsoft Wave (WAV) file. The signals were then segmented into 1-second samples, producing 284 asphyxia signals and 316 normal cry signals.

2.2. MFCC Feature Extraction

From the 600 cry signals, MFCC was used to extract the feature to be used in CNN. The feature were extracted using typical MFCC settings used in [11-14]. The number off$_B$ was obtained with used of the sampling frequency (f_s) using Equation **Erreur ! Source du renvoi introuvable.** which give the typical number of f_B = 26 forf_s = 8000Hz and typical value of n_c = 12.

$$f_B = 3 \times \log_{10} \times f_s \quad (1)$$

From this, an output of m x n_c size of a MFCC feature were generated (Fig.1). This output will be used as input part to the CNN.

2.3. CNN Parameter

A typical CNN architecture consists of a convolution layer, Rectified Linear Units (ReLU), max pooling layer and fully connected layer [1, 6, 15-16]:

1. Convolution Layer(s): The Convolution Layer(s) are a collection of filter banks (kernels) that are used to extract features from the inputs. Each kernel is responsible to detect specific patterns by examining small portions in the image as a feature [1, 17]. These kernels have tunable parameters which can be optimized for better feature extraction.
2. Rectified Linear Units (ReLU): ReLUs typically follow the Convolution Layers as an activation function. In neural network, the activation function is a function that determines whether a neuron fires or stays dormant. The commonly used MLP activation function (tangent-sigmoid) is less suitable for training CNNs since it has the problem of vanishing gradient (gradient essentially become zero after several iteration of training). Therefore, ReLU was proposed as an alternative to tangent-sigmoid activation function to avoid this issue [18]. Additionally, since the calculations inside ReLU layers is simpler, they are also able to speed up CNN training [19].
3. Pooling: Convolution and ReLU layers typically generate a significant amount of features. The features may impact CNN training performance if these features are sent directly to the deeper layers of the CNN without simplification. Pooling layers help to reduce the

resolution (scale) of the features thus decreasing the calculation costs for the following layers. Additionally, pooling layers can avoid local variances and smooth out the output. Three types of pooling methods are available namely max, min and average pooling [17]. The pooling method choice is then applied independently on each feature map produced by the convolution and ReLU layers [1-2].

4. Fully Connected Layer: The fully connected layer serves as the classification layer to complete the CNN. The fully connected layer is very similar to that of the Multi-Layer Perceptron (MLP) [22], which performs classification of the previously extracted features through the use of trained weighted connections.
5. Softmax Layer: The softmax layer is fitted at the output of the Fully Connected Layer [1, 17]. Its role is to consolidate and present the final CNN output to the user [6, 17]. Softmax activation function has been widely adopted in CNN due to its simplicity and probabilistic interpretation [20].

The possible combinations of convolution, RELU and Pooling layers are theoretically limitless and repeatable (however, the practical limitation is the processing power available in the GPU). Through a process called weight tying/sharing, each output from the previous layer is sent as an inputs to the next layer [2, 17].

In this work, a single-convolution structure for CNN parameter was used to test its ability in classifying sound feature. The CNN structure [21] used is shown in Fig.2. Two properties were adjusted during classification process namely the number of filters and filter diameter. These properties were used in CNN convolution part as a filter to extract the information in dataset. Adjusting this two value may affect the classification performance. The possible number of filter for selection [23] in this work is between 1 to 10, while the diameter is between 5 to 50 with 5 increment. The dataset were divided by the ratio of 70:30 for training and testing.

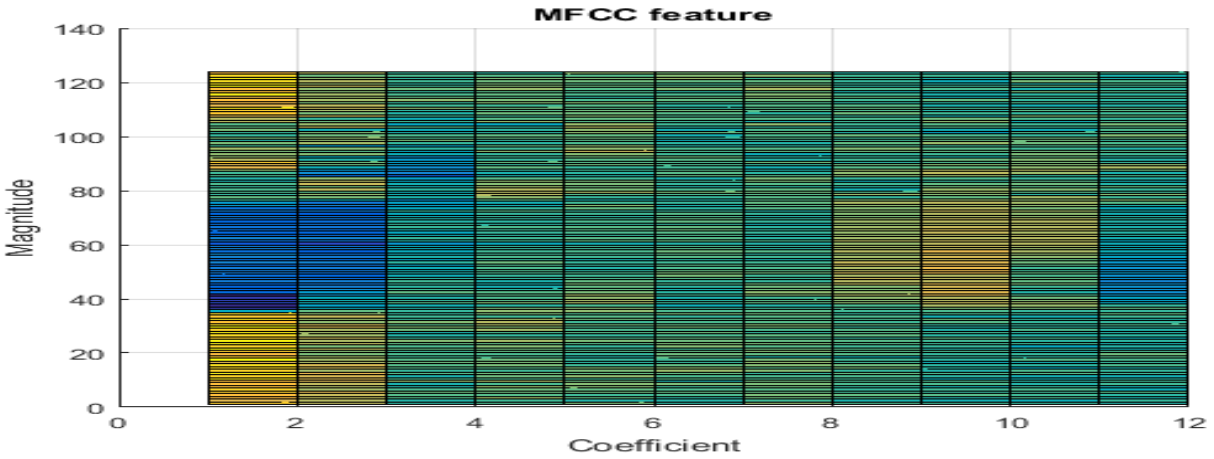

Fig.1. Sample two-dimensional output from MFCC

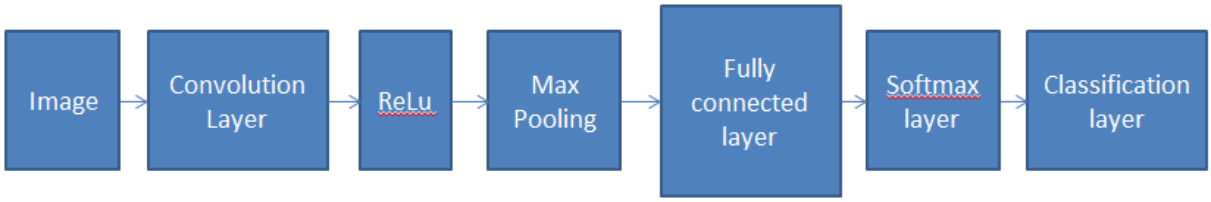

Fig.2. CNN layers to classify MFCC feature

3. RESULTS AND DISCUSSION

Fig.3 shows the result acquired from all possible combination number of filter and its diameter in the convolution layer. From the upwards trend of the graph, we can see that increasing the number of filters and their diameter size increases the CNN accuracy until a certain saturation point (at six filters). As can be seen, when the number of filters were increased to six, there were no noticeable improvement in the training accuracy while showing a decrease in testing accuracy. This may be caused by the low sampling frequency used for collecting data (we used the minimal sampling frequency of 8,000 during MFCC feature extraction). Because of this, the MFCC resolution became lower. With this resolution, we infer that a small number of filters is sufficient to extract the MFCC features. Based on the best number of filters, we proceeded with analyzing the optimal filter diameter size to use with this number of filters.

Table 1 shows the ten best results from all the possible combinations of filter number and its diameter in the convolution layer. The best result achieved was using four filters with a diameter of 30. This parameter combination produced 94.29% accuracy for training set and 92.78% accuracy for testing set.

Another observation was that four filters was the majority choice for the top ten results. From this information, we analyzed the CNN further using four filters, and the result is shown in Fig.4. It can be seen clearly there was a dramatic improvement in CNN training and testing accuracy when the filter diameters were varied between 5 and 10. When the filter diameter was further increased, there were minor improvements to the classification accuracy. With resolution at 8,000 sampling frequency, diameter size of 30 give the balance result between training and testing accuracy. The resulting confusion matrix for classification is shown in Fig. 5 and Fig. 6.

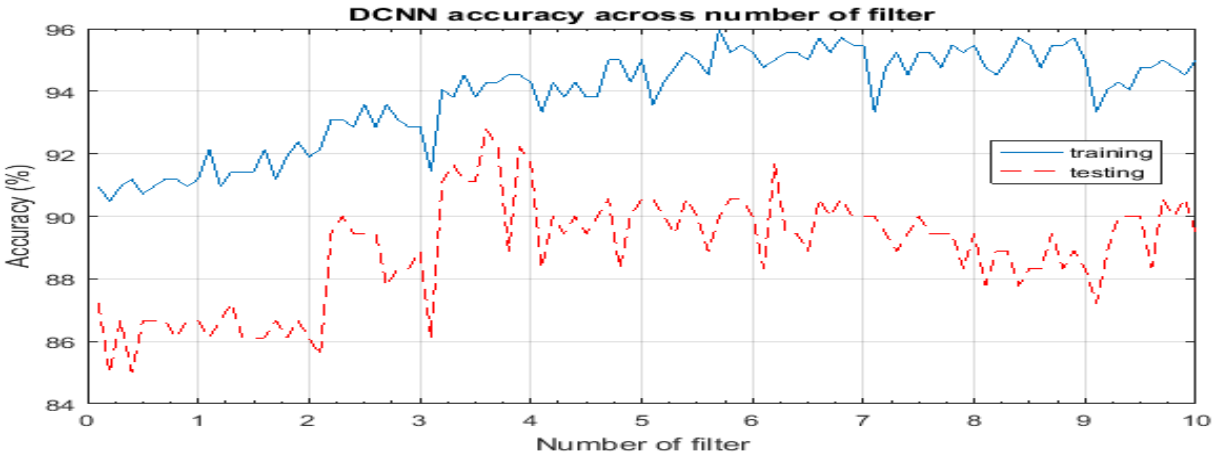

Fig.3. CNN accuracy versus number of filters

Table 1. Best ten result achieve from the combination of number of filters and filter diameter

No. of Filters	Diameter of Filters	Training Accuracy	Testing Accuracy
4	30	94.29	92.78
4	45	94.52	92.22
4	35	94.29	92.22
7	10	95.00	91.67
4	50	94.29	91.67
4	15	93.81	91.67
4	25	93.81	91.11
4	20	94.52	91.11
4	10	94.05	91.11
10	45	94.52	90.56

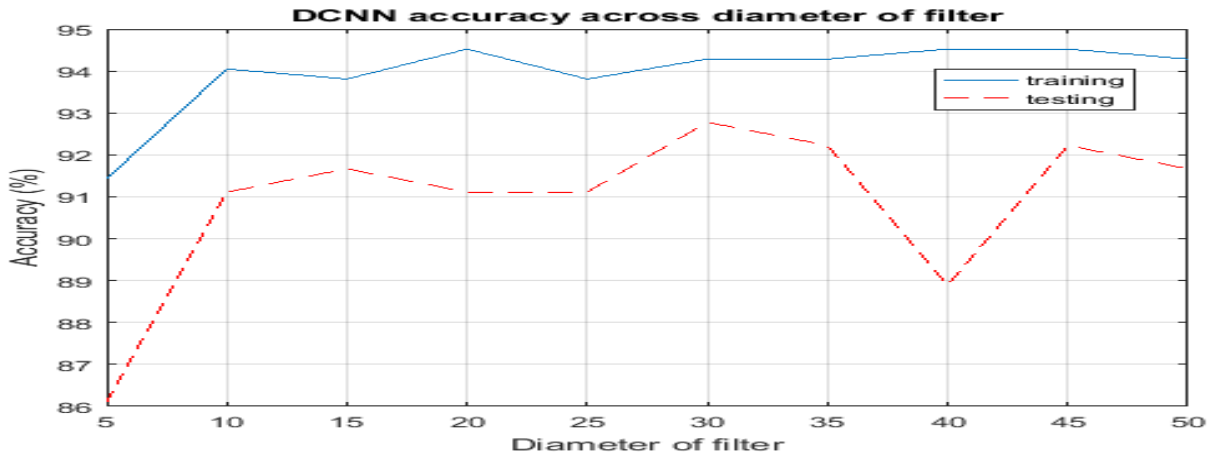

Fig.4.CNN accuracy versus filter size (optimal number of filters set at four)

Fig.5. Confusion plot for training set

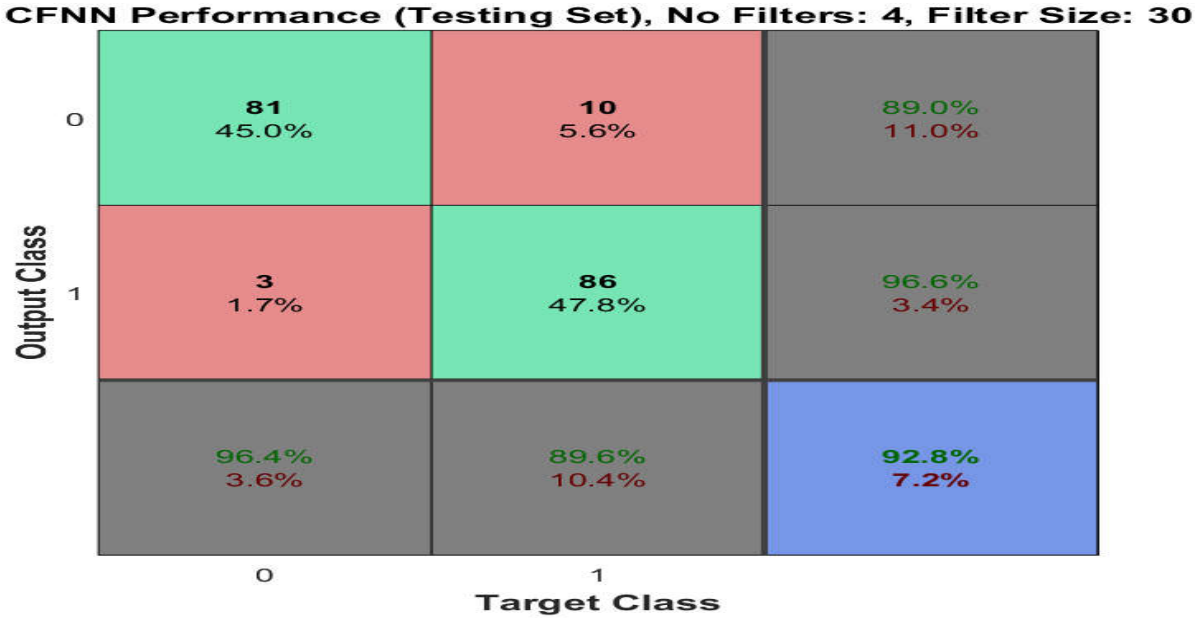

Fig.6. Confusion plot for testing set

4. CONCLUSION

This paper has presented a CNN trained on MFCC features to diagnose asphyxia in babies. A single convolution CNN was trained with MFCC features of normal and asphyxiated baby cry signals. The results show that the proposed method yielded very high accuracy, proving that the proposed method is very suitable for newborn asphyxia diagnosis based on non-invasive data acquisition method.

5. ACKNOWLEDGEMENTS

The authors would like to graciously acknowledge the Ministry of Higher Education and UniversitiTeknologi MARA for supporting this research work through Grant No: FRGS/1/2015/ICT03/UiTM/02/4.

6. REFERENCES

[1] El Housseini A, Toumi A, Khenchaf A. Deep Learning for target recognition from SAR images. In IEEE Seminar on Detection Systems Architectures and Technologies, 2017, pp. 1-5

[2] Yi H, Shiyu S, Xiusheng D, Zhigang C. A study on deep neural networks framework. In IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2016, pp. 1519-1522

[3] Kang K, Ouyang W, Li H, Wang X. Object detection from video tubelets with convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 817-825

[4] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief Nets. Neural Computation, 2006, 18(7):1527-1554

[5] Fukushima K, Wake N. Handwritten alphanumeric character recognition by the neocognitron. IEEE Transactions on Neural Networks, 1991, 2(3):355-365

[6] Vu T D, Yang H J, Nguyen V Q, Oh A R, Kim M S. Multimodal learning using convolution neural network and sparse autoencoder. In IEEE International Conference on Big Data and Smart Computing, 2017, pp. 309-312

[7] Jigar A D. Swift single image super resolution using deep convolution neural network. In International Conference on Communication and Electronics Systems, 2016, pp. 1-6

[8] Nilanon T, Yao J, Hao J, Purushotham S, Liu Y. Normal/abnormal heart sound recordings classification using convolutional neural network. In IEEE Computing in Cardiology Conference, 2016, pp. 585-588

[9] Lavner Y, Cohen R, Ruinskiy D, IJzerman H. Baby cry detection in domestic environment using deep learning. In IEEE International Conference on the Science of Electrical Engineering, 2016, pp. 1-5

[10] Chang C Y, Li J J. Application of deep learning for recognizing infant cries. In IEEE International Conference on Consumer Electronics-Taiwan, 2016, pp. 1-2

[11] Saha G, Yadhunandan U S. Modified mel-frequency cepstral coefficient. In International Association of Science and Technology for DevelopmentModelling and Simulation, 2004

[12] Pammi SC, Keri V. HTKTrain: A package for automatic segmentation. Hyderabad: International Institute of Information Technology, 2007

[13] Beritelli F, Grasso R. A pattern recognition system for environmental sound classification based on MFCCs and neural networks. In 2nd International Conference on Signal Processing and Communication Systems, 2008, pp. 1-4

[14] Brookes M. VOICEBOX: Speech processing toolbox for MATLAB. London: Imperial College, 1997

[15] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In 30th International Conference on Machine Learning, 2013, pp. 1-6

[16] Saxena S, Verbeek J. Convolutional neural fabrics. In 30th Conference on Neural Information Processing Systems, 2016, pp. 1-9

[17] Nithin D K, Bhagavathi S P. Learning of generic vision features using deep CNN. In 5th International Conference on Advances in Computing and Communications, 2015, pp. 54-57

[18] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(2):107-116

[19] Gulcehre C, Moczulski M, Denil M, Bengio Y. Noisy activation functions. In 33rd International Conference on International Conference on Machine Learning, 2016, pp. 1-10

[20] Liu W, Wen Y, Yu Z, Yang M. Large-margin softmax loss for convolutional neural networks. In 33rd International Conference on Machine Learning, 2016, pp. 507-516

[21] Azlee Z, Nooritawati M T, Ihsan M Y, Zairi I R.The performance of binary artificial bee colony (BABC) in structure selection of polynomial NARX and NARMAX models. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(2):373-379

[22] Ihsan M Y, Azlee Z, Rozita J, Megat S A M A, Rahimi B, Abu H A H, Zairi I R.Comparison between cascade forward and multi-layer perceptron neural networks for NARX functional electrical stimulation (FES)-based muscle model. International Journal on Advanced Science, Engineering and Information Technology, 2017, 7(1):215-221

[23] Ihsan M Y, Azlee Z, Megat S A M A, Nooritawati M T, Hasliza A H, Husna Z A, Zairi I R.Binary particle swarm optimization structure selection of nonlinear autoregressive moving average with exogenous inputs (NARMAX) model of a flexible robot arm. International Journal on Advanced Science, Engineering and Information Technology, 2016, 6(5):630-637

How to cite this article:
Zabidi A, Yassin I M, Hassan H A, Ismail N, Hamzah M M A M, Rizman Z I, Abidin H Z. Detection of asphyxia in infants using deep learning convolutional neural network (cnn) trained on mel frequency cepstrum coefficient (mfcc) features extracted from cry sounds. J. Fundam. Appl. Sci., 2017, *9(3S), 768-778*