2021 International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 25-27, 2021

A Novel Convolutional Neural Network Model for Musical Instruments' Classification: A Deep Signal Processing Approach

Dr. Basavaraj S. Anami
Principal
KLE Institute of Technology
Hubballi, India
anami_basu@hotmail.com

Dr. Kumar Swamy V.
Assistant Professor, Dept. of EEE
KLE Institute of Technology
Hubballi, India
kumarswamy.vadla@kleit.ac.in

2021 International Conference on Intelligent Technologies (CONIT) | 978-1-7281-8583-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CONIT51480.2021.9498541

***Abstract*— Stress management is a challenge in this modern world, and several methods are being practiced such as diet, exercise, sleep, meditation and relaxation. One of the relaxation methods is the Music Therapy, where one listens to Vocal music, Instrumental music, Hindustani classical music, and Western music etc. In instrument music, various types of instruments are used and these instruments vary from country to country, which produce different sounds. Based on the sounds, humans recognize these instruments with certain efficacy that varies with the expertise. In this paper, we propose the classification of 14 categories of musical instruments, namely, bells, cello, clarinet, crotales, double Bass, flute, piano, saxophone, trombone, trumpet, vibraphone, viola, violin, and xylophone etc. based on the sound signal. A total of 8,365 spectrograms, across 14 classes, are used. We have deployed five pre-trained CNN models such as Efficientnet, Googlenet, ResNet, Squeezenet, and Mobilenet for transfer learning. Since the classification accuracies of these pre-trained models are not up to the expectations, hence, a custom designed architecture is proposed, which outperforms the pre-trained models giving 99.81 % classification accuracy. Hyper-parameters are varied during the experimentation and the results are compared with the state-of-the-art methods. The work is helpful for the practicing psychiatrists, in knowing what types of sounds manage, which kinds of stresses.**

Keywords—sound signal processing, signal classification, convolution neural networks, transfer learning

I. INTRODUCTION

World Health Organization (WHO) statistics reveal that around 264+ million people are suffering from mental stresses, which lead to the deadly disease called depression. India has many organizations that render help in stress management, namely, Jivan Aastha helpline, Aasra, Cooj mental helpline Foundation, Vandravela, Sanjivani, sumaitri and many more. There are several ways to manage stress and meditation is one of the most popular and effective way of stress management. Several other methods include diet, exercise, sleep, and relaxation. One of the relaxation methods is the Music Therapy, where one listens to Vocal music, Instrumental music, Hindustani classical music, and Western music etc. India has rich heritage of musicians, who worked miracles, namely, Tansen.

The legendary musician Tansen, working in Akbar's court, sang "Raag Deepak" to lit lamps in the night and "Raag Megha Malhar" for rainfall. This is the power of vocal music. Musicians invariably use musical instruments while singing. There are different types of musical instruments, which vary from country to country, in terms of shape, size, way the sound produced and material used etc. Recognizing the instrument being played from its sound is considered as a trait of human beings. Musical instruments are also categorized into idiophones, membranophones, chordophones, aerophones and electrophones [24] and are shown in Fig. 1. There are instruments, which produce almost similar sounds, such as, violins, guitars, pianos, flute and brass instruments, causing problems in the recognition. Technology intervention would help in automatic recognition of musical instruments from their sound signals. In this work, we have considered 14 types of musical instruments, as given in Table 1. Each instrument produces a definite spectrogram based on the FFT (Fast Fourier Transform) of audio signals.

(a) (b) (c) (d) (e)

Fig. 1. Musical instrument- (a) idiophones (b) membranophones (c) chordophones (d) aerophones (e) electrophones

TABLE I. MUSICAL INSTRUMENTS USED IN SIGNAL CLASSIFICATION

Sl No	Musical Instrument	Sl No	Musical Instrument	Sl No	Musical Instrument
1	Bells	**6**	Flute	**11**	Vibraphone
2	Cello	**7**	Piano	**12**	Viola
3	Clarinet	**8**	Saxophone	**13**	Violin
4	Crotales	**9**	Trombone	**14**	Xylophone
5	Double Bass	**10**	Trumpet		

II. LITERATURE SURVEY

In order to know the state-of-the-art in this area, a literature survey is carried out and the gist of the papers is as given under.

Musical instrument sounds are classified on the basis of the sound onset, which is a sequence of spikes built through a gammatone filter bank. McGill dataset has 2085

musical tones, which are divided into five categories, giving classification accuracy of 75% [1]. Analysis of sound signals processed by cochlear implants is developed. Event-related potential waveforms are processed by measuring the fundamental frequency, duration and temporal gap, specific to children [2]. Features are extracted from the rhythmic sequence of music using graphs. Parameters like modularity, average degree and density of graphs are measured [3]. A study on the effect of musical practice using gestures is carried, which investigates an enhancement of audio-visual processing, using formal music training. Pitch sound duration and intensity are used as matrices [4]. The temporal patterns of fMRI are used for processing sound signals, which are cut into two hemispheres, showing clear asymmetry in fMRI signals [5]. Biological bases are identified to perceive musical timbre. Neurological computation framework using spectro-temporal fields is proposed. The model achieves 98.7 % accuracy to distinguish different sounds irrespective of pitch and playing style [6]. A work on identification of musical style based on modulating neural and behavioral response is carried. Identification of Jazz, classical and rock played by musician and non-musician is carried using mismatch negativity [7].

Researchers have also worked on anthropology as well. The study states that people near Magdalenian observe that Paleolithic shells produce music, enacted as wind instruments [9]. Temporal and dynamics of acoustics to categorize different sounds in experiments reveals that sound signal is identified reliably if duration is in between 12.5 to 200 milliseconds [11]. Sound stimulus is used to extract speech from sound signals. The study is conducted using cortical evoked potential analyzer at a level of 65 dB [12]. Apart from signal identification, multichannel compression is also carried. Attributes like alignment delay, compression speed and gain is used to test the pleasantness of the signal produced [13]. The influence of Syntax and semantics is used to extract texture in musical sequence. The study revealed that rhythms help perceptual and cognitive sequencing for syntax processing [14]. Evaluation of temporal processing on singers playing with and without musical instruments is carried. Tests of frequency and noise gaps are used as features. The study reveals that singers who play musical instruments are better than who only sings [16].

From the literature survey, it is observed that musical instruments' recognition from spectrograms is not being carried out and hence the present paper.

III. METHODOLOGY

The proposed methodology consists of two stages, namely, pre-processing and transfer learning as shown in Fig. 2.

The audio signals are obtained from University of Lowa Electronic Music Studios (ULEMS) [18]. There exist various musical instruments' audio files, which are in .wav format, and only 14 classes are considered in this work. Spectrograms of audio files are generated in the form of 2D images. Spectrogram provides the signal strength over the time at different frequencies present. Audio signal is passed through hamming window and Fast Fourier Transform (FFT) is applied to get spectrogram. Sample images of spectrograms of different musical instruments are shown in Fig. 3.

A. Preprocessing

In this stage, two preprocessing techniques are applied, namely normalization and data augmentation.

1) Normalization:

It is one of the important components in deep learning and makes CNN learn faster and to make gradient descent, activation function, more stable. The values of the pixels in the input images i.e., spectrogram, are normalized in the range [0, 1].

2) Data augmentation:

The number of original spectrograms is around 2,000+ and is found to be less in order to train different CNN models. Data augmentation involves rotation by 50, scaling by 15%, horizontal flipping and adding Gaussian noise. Images of a spectrogram post data augmentation is shown in Fig. 4.

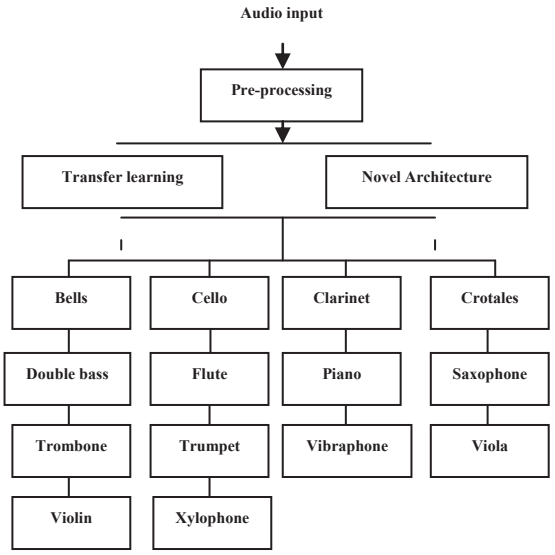

Fig. 2. Block schematic of the proposed study

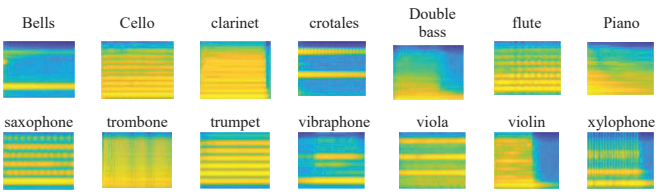

Fig. 3. Sample spectrogram of 14 musical instruments considered.

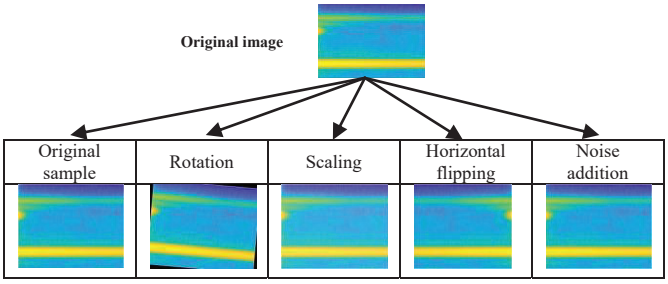

Fig. 4. Spectrograms after post augmentation

B. Transfer learning

We have used five pre-trained models, namely, Alexnet, Googlenet, Resnet, Mobilenet and Squeezenet and experimented for the classification of spectrograms of 14 types of musical instruments.

1) Pre-trained models

The pre-trained models are deployed for classification tasks, instead of developing a new architecture. Initially, we deployed five pre-trained models, namely, Alexnet [19], Googlenet [20], Squeezenet [21], Resnet-50 [22], and Mobilenet [23] and the details of the models are given in Table 2. Classification accuracies, so obtained are given in detail in section 4. Amongst, the considered models, the deepest is Mobilenet, having 53 layers. We experimented by changing the hyper parameters.

TABLE II. DETAILS OF PRE-TRAINED MODELS

Sl no	Model	Depth	Image size
1	Alexnet	08	227x227x3
2	Googlenet	22	224x224x3
3	Squuezenet	18	227x227x3
4	Resnet-50	50	224x224x3
5	Mobilenet	53	224x224x3
6	Proposed	06	100x100x3

2) Novel architecture

Since the performance of models depend on the depth, larger time and higher dimension of input images are required. Hence, a novel architecture with a smaller number of layers and smaller dimension of input images is proposed. In this proposed architecture for musical instruments sound classification, we have used a small number of layers and input layers accept images of size 100 x 100 x 3. Block diagram of the proposed architecture is shown in Fig. 5 and respective parameters considered are given in Table 3. Two layers of convolution, followed by down sampling, max pooling, a fully connected layer, Softmax layer and finally classification layer.

TABLE III. NOVEL CNN ARCHITECTURE PARAMETERS

	Name	Number	Description
Layers	Input	1	100x100x3
	2-D Convolution	6	Filter size- 16, 32 and 64
	Batch Normalization	6	--
	Activation function	6	Relu
	2-D Max Pooling	4	Pool size-2
	Fully Connected	1	No of classes-14
	Softmax	1	--
	Classification	1	Cross entropy
Hyper Parameters			
	No of epochs	30	Set based on performance of the network
	Batch size	300	
	Learning rate	0.0001	

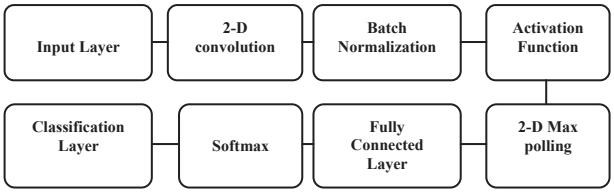

Fig. 5. Block diagram of proposed CNN architecture

IV. EXPERIMENTS AND RESULTS

Initially, we conducted experiments on pre-trained models, namely Alexnet, Googlenet, Squeezenet, Resnet and Mobilenet. The hyper-parameters of these pre-trained models are varied to derive their better efficiency. The models are being implemented using Matlab 2020b version. A total of 1,675 spectrograms are created from the University Iowa Music Studio dataset (UIMSD). A total 8,375 images of spectrograms are obtained after data augmentation, as discussed in section 3.1. With this available data, 70% is used for training and 30% is used for validation keeping in mind avoidance of overfitting.

TABLE IV. PERFORMANCE OF PRE-TRAINED MODEL AND PROPOSED ARCHITECTURE

Model	Epoch	Train loss	Valid loss	Valid accuracy (%)
Alexnet	**1**	**0.7324**	**0.3721**	**83.12**
	…	…	…	…
	…	…	…	…
	29	**0.0142**	**0.0043**	**94.36**
	30	**0.0123**	**0.0032**	**98.09**
Googlenet	**1**	**0.8123**	**0.3214**	**82.31**
	…	…	…	
	…	…	…	
	29	**0.1638**	**0.1752**	**98.26**
	30	**0.1612**	**0.1874**	**98.64**
Squeezenet	**1**	**0.7546**	**0.3412**	**91.02**
	…	…	…	…
	…	…	…	…
	29	**0.0054**	**0.0214**	**97.32**
	30	**0.0067**	**0.0311**	**97.87**
ResNet-50	**1**	**0.5412**	**0.3462**	**85.32**
	…	…	…	…
	…	…	…	…
	29	**0.0049**	**0.0124**	**98.21**
	30	**0.0036**	**0.0187**	**98.54**
Mobilenet	**1**	**0.7874**	**0.6412**	**78.25**
	…	…	…	…
	…	…	…	…
	29	**0.3142**	**0.1235**	**97.62**
	30	**0.3214**	**0.1249**	**97.97**
Proposed	**1**	**0.7865**	**0.5984**	**83.59**
	…	…	…	…
	…	…	…	…
	29	**0.0023**	**0.0122**	**99.32**
	30	**0.0019**	**0.0132**	**99.81**

A. Performance of Pre-trained Model

Performances of pre-trained models and the proposed Novel architecture are given in Table 4. The models are tested against parameters such as number of epochs, training loss, validation loss and accuracy. From Table 4, we infer that the proposed novel architecture out performs the

considered pre-trained model. In case of pre-trained models, the parameters, namely batch size, number of epochs and learning rate are set to 300, 30 and 0.0001 respectively by trial and error. The performance of the model is evaluated using performance metrics like accuracy, precision, sensitivity, specificity, F1-score [24], and the values are shown in Fig. 6.

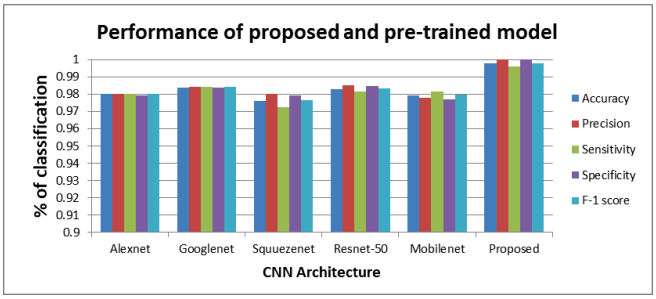

Fig. 6. Performance of proposed and pre-trained model.

B. *Performance Analysis with Different Optimizer:*

Experiment with different optimizers is carried out and the values of performance parameters are given in Table 5. Optimizers like Stochastic Gradient Descent with momentum (sgdm), rmsprop, adam are being deployed. Table 3 gives clearly that the use of sgdm optimizer helps in classification.

TABLE V. PERFORMANCE OF PRE-TRAINED MODEL AND PROPOSED ARCHITECTURE WITH DIFFERENT OPTIMIZERS

Model	Optimizer	Accuracy	Precision	Sensitivity	Specificity	F-1 score
Alexnet	Sgdm	0.9800	0.9803	0.9803	0.9795	0.9803
	Rmsprop	0.9782	0.9803	0.9765	0.9795	0.9784
	Adam	0.9761	0.9803	0.9727	0.9794	0.9765
Googlenet	Sgdm	0.9840	0.9842	0.9842	0.9837	0.9842
	Rmsprop	0.9825	0.9842	0.9803	0.9836	0.9823
	Adam	0.9852	0.9842	0.9765	0.9836	0.9803
Squuezenet	Sgdm	0.9761	0.9803	0.9727	0.9794	0.9765
	Rmsprop	0.9721	0.9803	0.9652	0.9792	0.9727
	Adam	0.9712	0.9765	0.9652	0.9751	0.9708
Resnet-50	Sgdm	0.9830	0.9851	0.9814	0.9846	0.9833
	Rmsprop	0.9792	0.9851	0.9742	0.9845	0.9796
	Adam	0.9755	0.9778	0.9742	0.9768	0.9760
Mobilenet	Sgdm	0.9792	0.9778	0.9814	0.9769	0.9796
	Rmsprop	0.9735	0.9778	0.9706	0.9766	0.9742
	Adam	0.9698	0.9706	0.9706	0.9688	0.9706
Proposed	**Sgdm**	**0.9981**	**1**	**0.9962**	**1**	**0.9981**
	Rmsprop	0.9962	0.9962	0.9962	0.9962	0.9962
	Adam	0.9924	0.9888	0.9962	0.9886	0.9925

C. *Performance Analysis with different Batch size*

The Performance is evaluated for different batch sizes, namely, 100, 200 and 300. The corresponding results are plotted, as shown in Fig. 7.

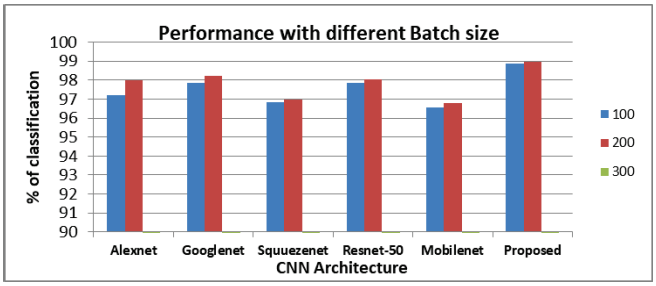

Fig. 7. Performance with change in Batch size

V. COMPARISON WITH EXISTING METHODOLOGY

The results obtained from the proposed architecture are compared with existing methodologies in the literature and the comparison is given in Table. 6. It is observed that the proposed novel architecture outperforms the state-of-the-art methods.

TABLE VI. PERFORMANCE EVALUATION

References	Technique	Classes	Number of samples	Accuracy (%)
[1]	Sound onset	5	2085	75.00
[3]	Graph based	10	1000	72.00
[6]	Cepstral Coefficients	11	1110	98.70
[17]	CNN (AlexNet, VGG16, VGG19, and basic CNN model)	8	--	84.06
Proposed method	**Novel architecture**	**14**	**8365**	**99.81**

VI. CONCLUSIONS

The five pre-trained models are tested with spectrograms obtained from the ULEMS dataset by changing the hyper parameters. The performances of Alexnet, Googlenet, Squeezenet, Resnet-50, and Mobilenet are 98.09%, 98.64%, 97.87%, 98.54% and 97.97% respectively for an epoch size of 30. The proposed novel architecture has given 99.81% accuracy for the same epoch size of 30. The depth of pre-trained models varies from 8 layers, in Alexnet to 53 layers, in Mobilenet and the image sizes are around 227 x 227 x 3. However, the proposed novel architecture has a depth of six layers and the image sizes are 100 x 100 x 3. Overall, the proposed model is considered, as efficient in terms of depth, and input size in classification of 14 musical instruments based on the spectrograms.

REFERENCES

[1] Newton MJ, Smith LS. "A neurally inspired musical instrument classification system based upon the sound onset", The Journal of the Acoustical Society of America, Vol: 131(6), pp. 4785-98. June 2012. doi: 10.1121/1.4707535. PMID: 22712950.

[2] Torppa R, Salo E, Makkonen T, Loimo H, Pykäläinen J, Lipsanen J, Faulkner A, Huotilainen M., "Cortical processing of musical sounds in children with Cochlear Implants", Clinical Neurophysiology, Vol: 123(10), pp. 19966-79. May 2012, doi: 10.1016/j.clinph.2012.03.008, PMID: 22554786.

[3] Melo DFP, Fadigas IS, Pereira HBB, “Graph-based feature extraction: A new proposal to study the classification of music signals outside the time-frequency domain” PLoS One, Vol: 15(11), Nov 2020, doi: 10.1371/journal.pone.0240915. PMID: 33180814.

[4] Proverbio AM, Attardo L, Cozzi M, Zani A, “The effect of musical practice on gesture/sound pairing”. Frontiers in Psychology, Vol: 6, April 2015. doi: 10.3389/fpsyg.2015.00376. PMID: 25883580; PMCID: PMC4382982.

[5] Izumi S, Itoh K, Matsuzawa H, Takahashi S, Kwee IL, Nakada T. Functional asymmetry in primary auditory cortex for processing musical sounds: temporal pattern analysis of fMRI time series. Neuroreport. 2011 Jul 13;22(10):470-3. doi: 0.1097/WNR.0b013e3283475828. PMID: 21642880.

[6] Patil K, Pressnitzer D, Shamma S, Elhilali M. Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol. 2012;8(11):e1002759. doi: 10.1371/journal.pcbi.1002759. Epub 2012 Nov 1. Erratum in: PLoS Comput Biol. 2013 Oct;9(10). doi: 10.1371/annotation/d8b290d3-32b7-4ded-b315-d1e699bb34da. PMID: 23133363; PMCID: PMC3486808.

[7] Vuust P, Brattico E, Seppänen M, Näätänen R, Tervaniemi M. The sound of music: differentiating musicians using a fast, musical multi-feature mismatch negativity paradigm. Neuropsychologia. 2012 Jun;50(7):1432-43. doi: 10.1016/j.neuropsychologia.2012.02.028. Epub 2012 Mar 6. PMID: 22414595.

[8] Timm L, Agrawal D, C Viola F, Sandmann P, Debener S, Büchner A, Dengler R, Wittfoth M. Temporal feature perception in cochlear implant users. PLoS One. 2012;7(9):e45375. doi: 10.1371/journal.pone.0045375. Epub 2012 Sep 21. PMID: 23028971; PMCID: PMC3448664.

[9] Fritz C, Tosello G, Fleury G, Kasarhérou E, Walter P, Duranthon F, Gaillard P, Tardieu J. First record of the sound produced by the oldest Upper Paleolithic seashell horn. Sci Adv. 2021 Feb 10;7(7):eabe9510. doi: 10.1126/sciadv.abe9510. PMID: 33568488; PMCID: PMC7875526.

[10] Vuust P, Brattico E, Seppänen M, Näätänen R, Tervaniemi M. Practiced musical style shapes auditory skills. Ann N Y Acad Sci. 2012 Apr;1252:139-46. doi: 10.1111/j.1749-6632.2011.06409.x. PMID: 22524351.

[11] Ogg M, Slevc LR, Idsardi WJ. The time course of sound category identification: Insights from acoustic features. J Acoust Soc Am. 2017 Dec;142(6):3459. doi: 10.1121/1.5014057. PMID: 29289109.

[12] Polat Z, Ataş A. The investigation of cortical auditory evoked potentials responses in young adults having musical education. Balkan Med J. 2014 Dec;31(4):328-34. doi: 10.5152/balkanmedj.2014.14171. Epub 2014 Dec 1. PMID: 25667787; PMCID: PMC4318404.

[13] Moore BC, Füllgrabe C, Stone MA. Determination of preferred parameters for multichannel compression using individually fitted simulated hearing AIDS and paired comparisons. Ear Hear. 2011 Sep-Oct;32(5):556-68. doi: 10.1097/AUD.0b013e31820b5f4c. PMID: 21285878.

[14] Canette LH, Lalitte P, Bedoin N, Pineau M, Bigand E, Tillmann B. Rhythmic and textural musical sequences differently influence syntax and semantic processing in children. J Exp Child Psychol. 2020 Mar;191:104711. doi: 10.1016/j.jecp.2019.104711. Epub 2019 Nov 23. PMID: 31770684.

[15] Ogg M, Moraczewski D, Kuchinsky SE, Slevc LR. Separable neural representations of sound sources: Speaker identity and musical timbre. Neuroimage. 2019 May 1;191:116-126. doi: 10.1016/j.neuroimage.2019.01.075. Epub 2019 Feb 5. PMID: 30731247.

[16] Ribeiro AC, Scharlach RC, Pinheiro MM. Assessment of temporal aspects in popular singers. Codas. 2015 Nov-Dec;27(6):520-5. English, Portuguese. doi: 10.1590/2317-1782/20152014234. PMID: 26691615.

[17] Raghu, Sriraam N, Temel Y, Rao SV, Kubben PL. A convolutional neural network based framework for classification of seizure types. Annu Int Conf IEEE Eng Med Biol Soc. 2019 Jul;2019:2547-2550. doi: 10.1109/EMBC.2019.8857359. PMID: 31946416.

[18] Electronic music studios, 'University of Lowa Social Electronic music studios', Available: http://theremin.music.uiowa.edu/MIS.html [Accessed: 09- April- 2021].

[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[21] Forrest N. Iandola and Song Han and Matthew W. Moskewicz and Khalid Ashraf and William J. Dally and Kurt Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, 2016.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[24] Dalianis H, “Evaluation Metrics and Evaluation. In: Clinical Text Mining”, Springer, Cham. https://doi.org/10.1007/978-3-319-78503-5_6, (2018), ch. 6, pp. 45-54.