



# **iJRASET**

International Journal For Research in  
Applied Science and Engineering Technology



---

# **INTERNATIONAL JOURNAL FOR RESEARCH**

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume: 7      Issue: XI      Month of publication: November 2019**

**DOI: <http://doi.org/10.22214/ijraset.2019.11058>**

**[www.ijraset.com](http://www.ijraset.com)**

**Call:  08813907089**

**E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)**

# Audio based Species Identification for Monosyllabic Call Birds using Convolutional Neural Networks

Rigved Kulkarni<sup>1</sup>, Bhavin Chandarana<sup>2</sup>

<sup>1</sup>Pune University

<sup>2</sup>IIT Madras

**Abstract:** In this paper we present 2 simple approaches to identify the species of a bird based on its call, using deep learning technique of CONVOLUTIONAL Neural Network (CNN). We present a 1D CNN approach that takes direct signals as inputs to the Neural Network. We extract features from the sound signals and input them to a 2D CNN as part of the second approach. This approach was tried out on 2 data sets of bird calls for 39 Indian species and 35 European species

**Keywords:** CNN, 1D-CNN, 2D-CNN, Bird species identification, MFCC

## I. INTRODUCTION

### A. Overview

Vocalization is a vital mode of communication for birds. Bird calls and songs have distinct spectral features<sup>[1][2][3]</sup>. Spectral Analysis of bird speech is gaining attention in recent times as a means of identifying the bird species from its call or song<sup>[4][5][6][7][8][9][10]</sup>. Properties of vocalization vary based on various physiological as well as external factors. While we ignore the latter in this study, differences in the former can be exploited for identifying a bird species from its call. When done with the help of machines, this can potentially become a valuable tool for ecological surveys, ornithological studies as well as hobbyist birdwatching.

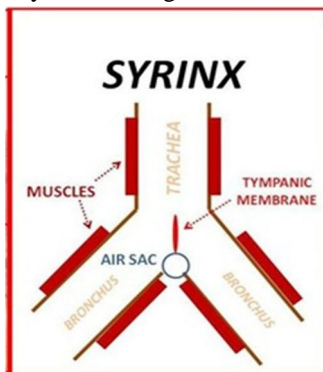


Fig. 1. Simplified model of a bird Syrinx<sup>[11]</sup>

### B. Biology

Anatomy of bird speech differs from that of humans in some key areas. Birds don't have vocal cords. Sound is instead produced via an organ called the Syrinx (shown in Fig 1) which is located near the junction of the trachea and the bronchi<sup>[12]</sup>. Its sound source is the tympani form membrane which faces to the bronchus on one side and the air sac on the other. When some of the syringeal muscles contract, the lumen of the bronchus is throttled and produces vibration in the membrane. When stretched along one dimension, the membrane produces harmonic sounds, but when stretched along two dimensions, the sound contains non-harmonic components. Syrinx can feature multiple simultaneous oscillation modes. In many cases, the sound production system of birds operates in a highly nonlinear or chaotic mode which gives rise to rapid changes in the operation of the organ. As a result, temporal variations in bird sound are typically found to be much faster than that of human sound. Hence a higher temporal resolution compared to that used in human speech processing is recommended in the bird sound analysis<sup>[13]</sup>

Like human speech, bird speech can be modeled as a collection of syllables. If the same syllable is repeated, it's called a call. If it's a repeating pattern of different syllables it's called a song. Despite these differences it was found that analytical techniques used for human speech analysis can be used for bird speech analysis as well. The current study is restricted to monosyllabic calls only. We will ignore song birds. Special variations like mating calls, gender and seasonal variations are also ignored.

## II. METHODOLOGY

### A. Overview

As shown in Fig 2, we start with preprocessing where we segment the sound files of each dataset ( $D_i$ ) into a corresponding dataset of numerous smaller files containing a single syllable each ( $S_i$ ). After that, 4 models were trained using  $S_i$ 's. A 1D CNN and a 2D CNN each for Indian and European birds

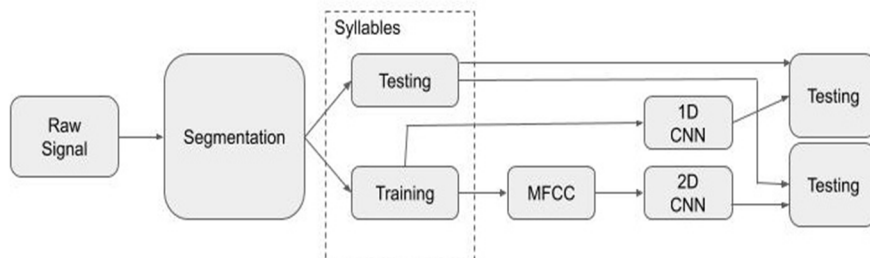


Fig. 2. Block diagram of the approach

( $D_1, D_2, D_3$ : Original sound file datasets

$S_1, S_2, S_3$ : Extracted corresponding syllables

$M1_{1D}, M1_{2D}, M2_{1D}, M2_{2D}$ : CNN Models

$MFCC_1, MFCC_{12}$ : Respective Mel Cepstrum Coefficients)

For 1D CNN, we take the raw signal directly as an input where as for 2D CNN, we extract cepstral features that are then used as input.

### B. Data Acquisition

- 1) Dataset 1: On field recordings were collected for 41 different bird species totaling to 163 files. These were bought from a local ornithologist<sup>[14]</sup>. These are all Indian birds. Hereafter referred to as **D1**
- 2) Dataset 2: Training dataset of The ICML 2013 Bird Challenge, hosted on Kaggle<sup>[15]</sup>, containing recordings for 35 species with 2801 individual data points, was procured. Based on the literature from the Kaggle website, we find that these are European birds<sup>[15]</sup>. Hereafter referred to as **D2**

For each dataset, 20% was removed at the start for testing. From the remaining 80%, 64% was used for training and 16% for validation

### C. Segmentation

For auto segmentation, [17] was largely followed. Give below is a step by step procedure for the same

- 1) Audio files from each dataset were resampled to 22050 Hz. A spectrogram was obtained with Hanning window, column width of 512 and 75% overlap
  - 2) This spectrogram normalized to the [0,1] interval and bottom 5 and top 30 frequency bins are removed as they predominantly contain noise
  - 3) The resulting image was converted to a binary mask using Median Clipping. Here, each pixel is set to 1 if its value is greater than 3 times the median of its row and also its column. Else its set to 0
  - 4) The resulting image is cleaned using standard morphological image processing techniques of closing, dilation and median filtering
  - 5) Smaller regions (i.e. groups of connected pixels not exceeding a threshold) in the resulting are removed to get the final mask. 2 Approaches were tried here:
    - a) A fixed threshold was set for all files. While this worked well for files with high SNR, those with low SNR gave both false positives and false negatives
    - b) To avoid these cases, a dynamic threshold was selected to only choose a given number of regions with the highest area
  - 6) Manual inspection was done to remove any remaining false positives and false negatives. While (5.2) prevented all noise packets from being selected in all scenarios, some false positives were encountered in cases where the syllables were close enough to become a continuous area, making room for a noise packet to be selected instead. (5.1) had instance of both
- The outcome of each of these steps can be seen in Fig 3.



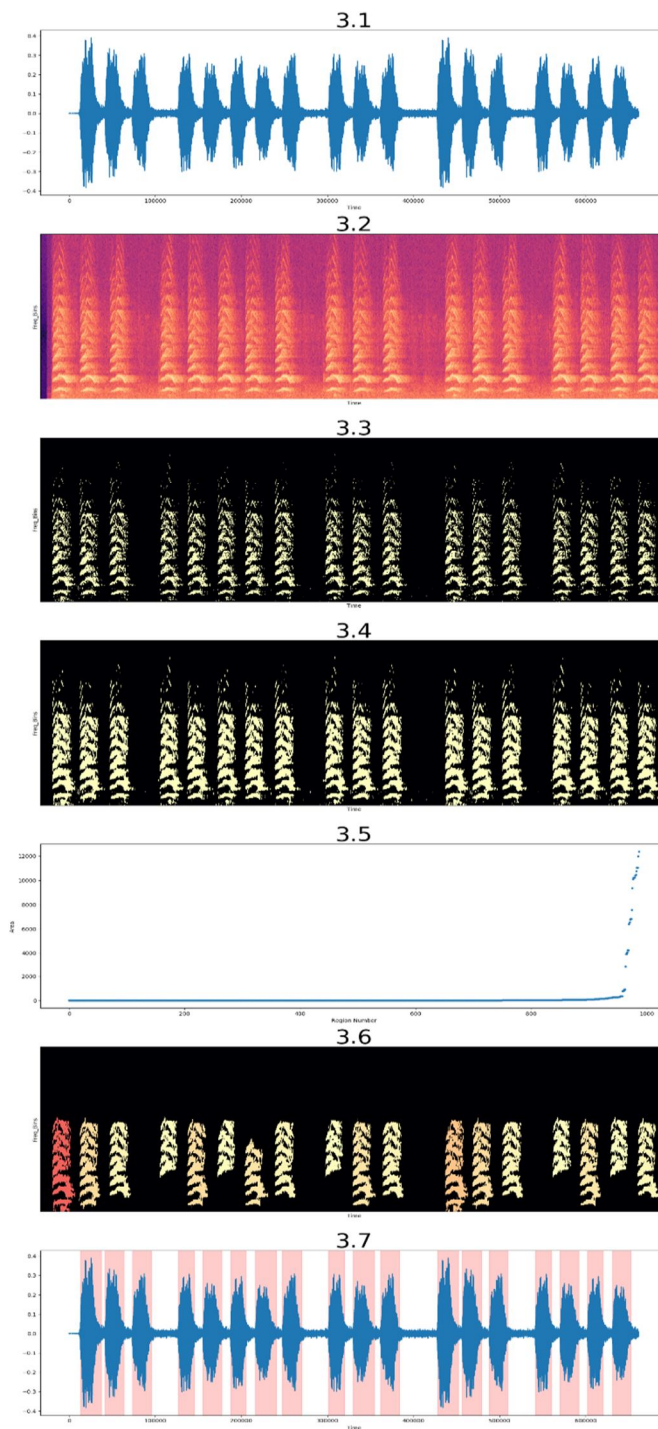


Fig 3. Segmentation Outcomes

3.1. Raw Signal (amplitude vs time)

3.2. Spectrogram (freq. bins vs time)

3.3. Median Clipped Mask (freq. bins vs time)

3.4. Post applying closing dilation & median filtering (freq. bins vs time)

3.5. Areas of regions (area vs region number)

3.6. Mask with syllables selected (freq. bins vs time)

3.7. Segmentation Output (amplitude vs time)

The above auto segmentation algorithm was able to correctly identify close to 80% of the syllables correctly as can be seen in Table 1 below.

TABLE I. AUTOSEGMENTATION ACCURACY

Dataset	No. of files	Syllables (Auto-segmentation)	Syllables (manual inspection)	Accuracy
D1	163	1317	1700	77.47%
D2	35	2338	2801	83.47%

#### D. Data Augmentation

Post segmentation, it was found that there was a disparity in the number of syllable samples per species in D2. To fix this, a minimum threshold of 50 was selected and data augmentation was done for all classes with less than the threshold number of samples. The following methods were employed for this:

- 1) *Adding Noise*: A random noise, 3-10% the RMS value of the signal was added to a randomly chosen sample. This helps in reducing generalization errors
- 2) *Pitch Shift*: Small amounts Pitch Shifts can help reduce classification errors<sup>[7][18][19]</sup>. 5% pitch shift was also applied to the above chosen sample

Samples for chosen for data augmentation till the total number of samples reached the threshold number

#### E. Feature Extraction

Cepstrum belongs to the class of homomorphic representations<sup>[18][20]</sup>, that has been found to be useful for various recognition tasks. The Mel cepstrum<sup>[21]</sup>, particularly, is popular for its robustness and simplicity. It also has the advantage of not requiring any parameter tuning to compute<sup>[22]</sup>. While Mel cepstrum is useful for analysis of human speech, it has been found to be effective for avian speech analysis as well<sup>[22][23][24]</sup>. Mel Frequency Cepstral Coefficients (MFCC's) are calculated from the logarithmic Mel cepstrum as follows:

$$MFCC_i = \sum_{k=1}^K X_k \cos\left(\frac{i(k-0.5)\pi}{K}\right) \quad (1)$$

$MFCC_i$  is the  $i$ th MFCC,  $X_k$  is the logarithmic energy of the  $k^{\text{th}}$  Mel-spectrum band, and  $K$  is the total number of the Mel-spectrum bands. For the 2D CNN model, first 40 MFCC coefficients of the sample are used as input

#### F. Modelling

While RNNs/LSTMs might typically be used for classifying time series data, there is increasing application of CNN techniques in signal processing<sup>[25][26][27]</sup>. Here, we have tried two network architectures; a 1-D CNN which works directly with the raw audio signal and a 2-D CNN which takes the MFCC sequence of the audio signal as the input

The motivation behind using the raw audio signal is the simplicity of the pipeline - once trained, this method requires no pre-processing of the audio and can be used directly. The intuition behind using a 1-D, single channel CNN is that the classifier will try to learn the spatial patterns in the audio signal which are specific to a given bird species. The MFCC sequence captures the unique patterns in each audio signal, which are thus unique for each bird species to give us a single channel 2-D input. The spatial variations in the MFCC sequence are then classified using a 2-D CNN. Details of the architecture of both CNNs are specified in the next section. For both classifiers the input signal is down-sampled to 16 kHz and clipped to a two-second duration. In cases where the audio signal is shorter than two seconds, it is padded with zeros since both CNN architectures can only process inputs of a fixed dimension.

### G. Network Architecture

Both 1-D and 2-D CNN's consist of a sequence of three 'blocks' followed by a dense layer, as shown in Fig 4 and Fig 5. Each block comprises of two convolution layers, followed by a max-pooling layer. Each layer uses ReLU non-linear activation, with the second convolution layer also using batch normalization for efficient training. For regularization, a dropout layer has been added to each block. A dropout probability of 0.2 was chosen for 1D CNN and 0.35 for 2D CNN. The final output layer uses a softmax activation with loss calculated using categorical cross-entropy

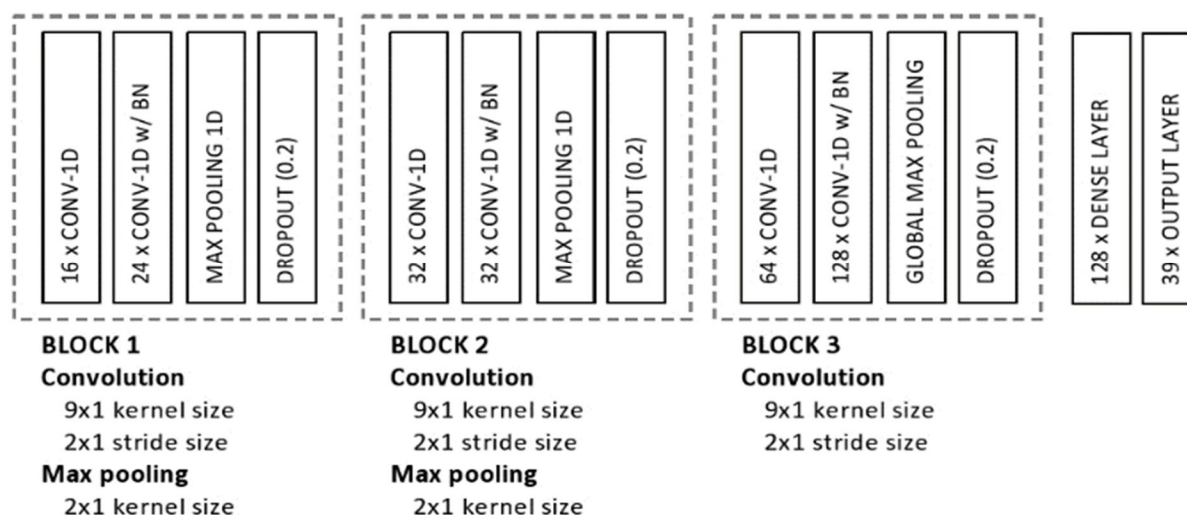


Fig 4. Block diagram for 1D CNN Network Architecture

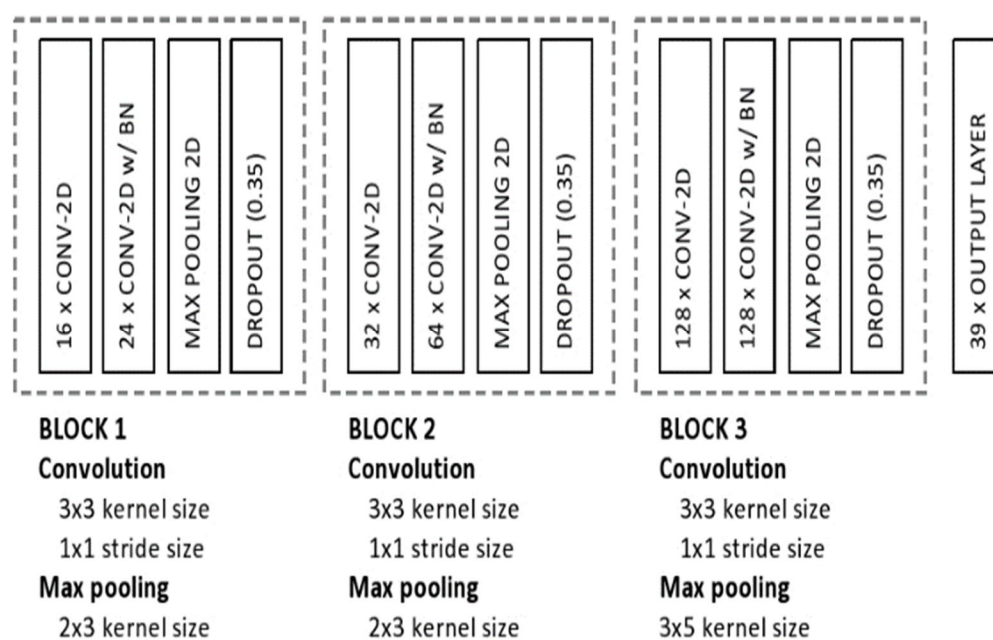


Fig 5. Block diagram for 2D CNN Network Architecture

### H. Training

Both 1-D and 2-D classifiers were trained using a constant learning rate of 0.0001. Both models used a batch size of 32 and Adam optimizer for training. Weights in both networks were initialized with Xavier initialization, and the inputs were normalized for maximal efficiency in training. Both models were trained for 200 epochs

### III. RESULTS

The accuracies for both models are listed in *Table II*. As seen from the results, the 1D CNN classifier is simple and would require little to no pre-processing of the raw audio. However, the 2D CNN based approach results in a better performance

TABLE II. Accuracy

Dataset	Efficiency	
	1D-CNN	2D-CNN
D1 (Validation Set)	73.78%	85.93%
D1 (Test Set)	67.35%	80.47%
D2 (Validation Set)	60.30%	82.54%
D2 (Test Set)	56.33%	74.78%

#### A. Conclusion

Using CNN results in a reasonably better accuracy as compared to linear classifiers or similar machine learning techniques. It is also evident that using feature extraction techniques like MFCC results in a more accurate classification when compared to using raw signal directly. The seeming disparity in the efficiency of the models generated by D1 and D2 can be explained by the fact that D1 was collected using professional equipment and has a very high SNR. Also D1 is a smaller dataset than D2

#### B. Outlook

This study was done with the broad idea that species identification can be automated. The biggest potential hurdle to building a fully automatic solution is the auto-segmentation step and ways to improve or bypass it in the future would be very valuable towards this goal. We would also like to reiterate that song birds as well as specific calls – like seasonal or mating calls we ignored in this study. A real world solution should probably handle those. Also, this study assumes only a single bird species in the foreground. However, it is quite reasonable to assume that there might be multiple birds in the background as well.

### IV. ACKNOWLEDGMENTS

We would like to thank Chinmay Bodhe for help with data gathering, clean up and with technical writing for this paper. We would also like to thank Prashant Ganu and Rahul Kango from Seligent Consulting, Pune, India for providing the computing resources used in this study. The Indian bird dataset was purchased from Sharad Apte – an ornithologist based out of Pune

### REFERENCES

- [1] A. Härmä, P. Somervuo: Classification of the harmonic structure in bird vocalization, "IEEE International Conference on Acoustics, Speech, and Signal Processing" <https://ieeexplore.ieee.org/abstract/document/1327207>, 2004
- [2] P. Somervuo; A. Harma; S. Fagerlund: Classification of the harmonic structure in bird vocalization. "IEEE Transactions on Audio, Speech, and Language Processing", vol. 14, Issue. 6, Nov. 2006). <https://ieeexplore.ieee.org/abstract/document/1709912>
- [3] Hemant Tyagi, Rajesh M. Hegde, Hema A. Murthy, Anil Prabhakar: "Automatic identification of bird calls using Spectral Ensemble Average Voice Prints", 14th European Signal Processing Conference, 2006, <https://ieeexplore.ieee.org/document/7071098>
- [4] Dan Stowell and Mark D. Plumbley, "Audio-only bird classification using unsupervised feature learning", CLEF 2014, <http://ceur-ws.org/Vol-1180/CLEF2014wn-Life-StowellEt2014.pdf>
- [5] Allan Gonçalves de Oliveira et al (2015): "Bird acoustic activity detection based on morphological filtering of the spectrogram" Applied Acoustics
- [6] Juang, C., & Chen, T., "Birdsong recognition using prediction based recurrent neural fuzzy networks", Neuro computing., vol 71, pp. 121-120, 2007.
- [7] Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann, "Audio Based Bird Species Identification using Deep Learning Techniques"
- [8] Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S., et al., "Classification of multiple bird species", Journal of Acoustic Society of America, vol. 131, pp. 4640-4650, 2012.
- [9] Fagerlund, S.: Bird species recognition using support vector machines. In: Journal on Advances in Signal Processing (2007), 7, 64–71
- [10] Kenji Doya, Terrence J. Sejnowski: A Novel Reinforcement Model of Birdsong Vocalization Learning. In: Advances in Neural Information Processing Systems 7 (NIPS 1994)
- [11] <https://echolocation-physiology-ansc3301.weebly.com/birds.html>



- [12] Vicario, D. S. (1991). Neural mechanisms of vocal production in songbirds. *Current Opinion in Neurobiology*, 1 :595-600
- [13] Härmä, Aki. Automatic identification of bird species based on sinusoidal modeling of syllables. In: *Acoustics, Speech, and Signal Processing*, In: 2003 IEEE International Conference on, vol. 5, pp. V-545. IEEE
- [14] <https://www.birdcalls.info/>
- [15] <https://www.kaggle.com/c/the-icml-2013-bird-challenge>
- [16] Deroussen, F., 2001. Oiseaux des jardins de France. Nashvert Production, Charenton, France; Deroussen, F., Jiguet, F., 2006. La sonothèque du Museum: Oiseaux de France, les passereaux. Nashvert production, Charenton, France; <http://naturophonia.fr>.
- [17] Mario Lasseck (2013): Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition
- [18] B. P. Bogert, M. J. R. Healy, and J. W. Tukey (1963): The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking in *Proc. Symp. Time Series Analysis*, Istanbul, Turkey, Jun. 5–9, pp. 209–243
- [19] Schldeluter, J., Grill, T.: Exploring data augmentation for improved singing voice detection with neural networks
- [20] A. Oppenheim and R. Schaffer (1975): *Digital Signal Processing*. New York: Springer, Prentice-Hall
- [21] S. Davis and P. Mermelstein (1980): Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366
- [22] Panu Somervuo, Aki Härmä, and Seppo Fagerlund (2006): Parametric Representations of Bird Sounds for Automatic Species Recognition. *IEEE Transactions on Audio, Speech, and Language processing*, vol. 14, no. 6
- [23] Jiri Stastny, Michal Munk, Lubos Juranek (2018): Automatic bird species recognition based on birds vocalization *EURASIP Journal on Audio, Speech, and Music Processing* 2018:19
- [24] Julien Ricard, Hervé Glotin (2016): Bag of MFCC-based words for bird identification. *LifeCLEF*
- [25] Yunan Wu, Feng Yang, Ying Liu, Xuefan Zha, Shaofeng Yuan (2018): A Comparison of 1-D and 2-D Deep Convolutional Neural Networks in ECG Classification
- [26] Håkon Gimse (2017): Classification of Marine Vessels using Sonar Data and a Neural Network, Norwegian University of Science and Technology
- [27] Thibaut Perola, Michaël Gharbi, Marine A. Denolle (2017): Convolutional Neural Network for Earthquake Detection and Location





10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)