International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-13 Issue-1, December 2023

Common Bird Sound Recognition at Vietnam Based on CNN

Phan Thi Ha, Trinh Thi Van Anh

Abstract: ***This article about developing a software extracting bird sound from a website [13], that has sounds of different bird species in Vietnam, explores the CNN model to develop a bird sound recognition system. The process includes conducting methodological experiments on self-collected datasets, providing assessments based on obtained results and building a bird sound recognition application.***

Keyword: ***CNN, RNN, CRNN, Bird CLEF, Recognition, Classification, Bird Sound, Mel Spectrogram, Xception, Mobile Net, Efficient Net.***

I. INTRODUCTION

Bird songs and sounds play a major role in their interactions and communication with other species, and are an important part of their behavior. Bird sounds can vary greatly in their frequency, structure, timbre and complexity and the classification of bird sounds and songs is highly practical. For example, in primeval forests, ecologists can identify birds' territories and their living conditions as well as the current forest condition by the way birds communicate and their sounds.

A common method for identifying bird sounds is to use spectroscopy, a visual representation of the frequency and amplitude of sound waves over time, such as the mel-frequency cepstral coefficients MFCC [11], Mel Spectrograms [8]… using K-nearest neighbors algorithm (KNN) or HMM, convolutional neural network (CNN) [6,7,10,12], Recurrent Neural Network (RNN) [2,3] or Convolutional Recurrent Neural Network (CRNN) [4].

All data for training machine learning models is mostly very raw, recorded in noisy outdoor conditions, with cross-stitching noises and other sounds from wind, insects and animals. Currently, there is a raw data set used in the data science competition on bird sound recognition called BirdCLEF [14] which includes the sounds of many bird species from different regions of the world.

Manuscript received on 03 November 2023 | Revised Manuscript received on 12 November 2023 | Manuscript Accepted on 15 December 2023 | Manuscript published on 30 December 2023

**Correspondence Author(s)*
Dr. Phan Thi Ha*, lecturer, Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT), Ha Noi, Vietnam, and Computing Fundamental Department, FPT University, Hanoi, Viet Nam. Email: hapt@ptit.edu.vn, hapt27@fe.edu.vn and hathiphan@yahhoo.com, ORCID ID: 0009-0009-5421-1990
Trinh Thi Van Anh, lecturer, Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT), Ha Noi, Vietnam, and Computing Fundamental Department, FPT University, Hanoi, Viet Nam. Email: vanh22@yahoo.com, anhttv@ptit.edu.vn and anhttv20@fe.edu.vn, ORCID ID: 0009-0005-8062-2014

However, it has little to no data on common bird species of countries from tropical climates in general and from Vietnam in particular such as storks, white-breasted waterhens, etc. Therefore, we have built a separate data set of bird sounds of these common bird species in Vietnam for this research article. Within the scope of this article, we would like to present the research process with applying deep learning and convolutional neural networks [1][16][17][18][19] to identify common bird sounds in Vietnam. The project is divided into four parts as follows.

II. BUILDING A MODEL OF BIRD SOUND RECOGNITION PROBLEM BASED ON CNN MODEL

A. Building a General Model for the Bird Sound Recognition Problem

We have built model Figure 1 as an overview of bird sound recognition using CNN:

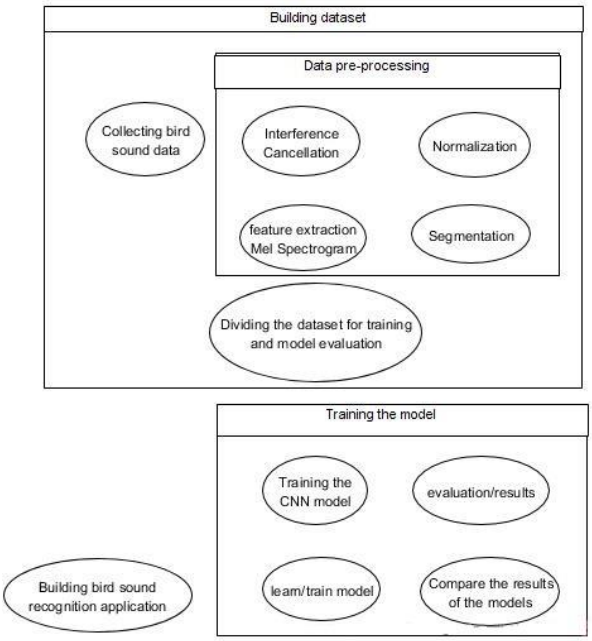


Figure 1. General Model for Bird Sound Recognition Problem

- Step 1. Building dataset
  - Collecting bird sound data
  - Data preprocessing
  - Dividing the dataset for training and model evaluation
- Step 2: Training the model
  - Training the CNN model
  - Evaluation and results

Step 3: Building bird sound recognition application

B. Building Dataset

a. Collecting Data

We built software to automatically collect raw data from the site [13] including bird sounds in Vietnam. The software will collect a list of web paths for birds as demonstrated in Figure 2 and then collect web paths for bird sounds as demonstrated in Figure 3.

Table 1. Description of Data Used in the Article

Species	Percentage	Species	Percentage
Coucal	4.48%	Pigeon	0.65%
Black-collared starling	0.58%	Slaty-breasted rail	0.81%
Red-whiskered bulbul	8.85%	Drongo	1.59%
Wagtail	0.61%	Magpie robin	5.34%
Chóc quạch	0.65%	Stork	0.92%
Cồng cộc	0.67%	Spotted dove	1.80%
Cuckoo	3.22%	Cúm núm	0.89%
Quail	4.96%	Snipe	3.56%
Swallow	1.32%	Héc xoan	1.15%
Nightingale	3.65%	Canary	1.14%
Olive-winged Bulbul	0.89%	Blackbird	1.82%
Orange-headed thrush	1.24%	Sunbird	1.67%
Huýt cô	1.13%	Babbler	6.32%
White-eye	3.46%	Lesser whistling duck	0.98%
Red-eyed	1.13%	Mắt xéo	1.15%
Curlew	1.24%	Ngũ sắc	0.82%
Robin	1.18%	Hornbill	0.98%
Quế lâm	0.65%	White-breasted waterhens	3.44%
Bald coot	1.16%	Brown starling	1.83%
Worm bird	2.94%	Sparrow	5.58%
Asian Fairy-bluebird	0.98%	Thanh tước	1.19%
Trao trảo	2.88%	Trĩ	0.82%
Swamphen	1.16%	Trích ré	3.60%
Black-crowned night heron	1.13%	Parrot	0.65%
Xanh tím	1.15%		

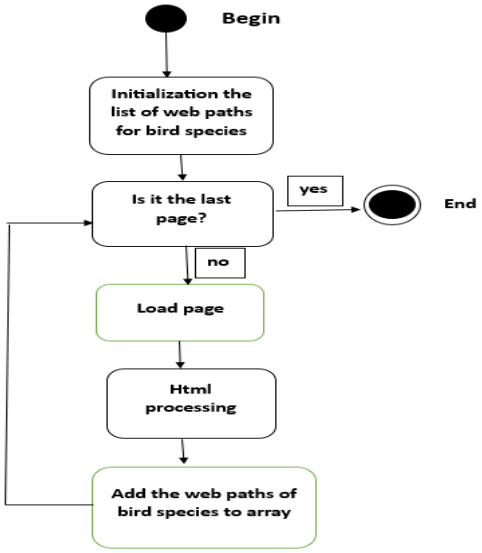


Figure 2. Map of Web Paths for Collecting Bird Species from Tiengdong.Com

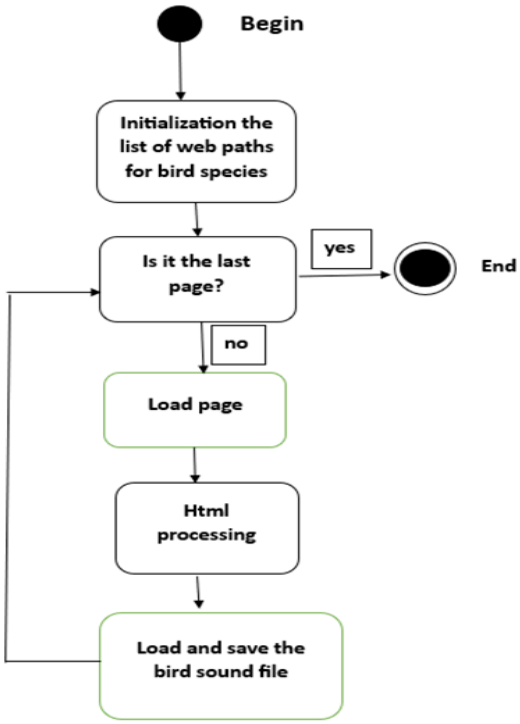


Figure 3. Map of Web Paths for Collecting Bird Calls from List of Web Paths in Figure 2

Finally, the resulting dataset is a collection of audio files saved as mp3 of 49 bird species. There are 115 samples in total, each ranging in duration from under 3 to 16 minutes. The total duration of all samples from all species was 20 hours and 45 minutes.

a. Data Preprocessing

- Noise reduction

The article uses spectral subtraction to eliminate noise. Spectral subtraction is a technique for reducing noise in an audio signal by estimating the noise spectrum and subtracting it from the signal. Here is a general outline of how to perform spectral subtraction:

- ✓ Divide audio signal into overlapping frames.
- ✓ Calculate the power spectrum of each frame using Fast Fourier Transform (FFT).
- ✓ Estimate the noise spectrum by averaging the power spectrum of several noise-only frames.
- ✓ Subtract the estimated noise spectrum from the power spectrum of each frame.
- ✓ Apply amplification coefficient to the acquired spectrum to avoid noise amplification.
- ✓ Use the inverse FFT to convert the modified spectrum back to the time domain.
- ✓ Overlap and add the resulting frames to reconstruct the audio signal.
- ✓ Spectral subtraction can be effective in reducing noise in an audio signal, but it can also create audible noise and decrease signal quality if the noise spectrum estimate is incorrect.
- ✓ To improve the efficiency of spectral subtraction, you can use more advanced techniques such as Wiener filtering or Kalman filtering.

Retrieval Number: 100.1/ijitee.A97561213123
DOI: 10.35940/ijitee.A9756.1213123
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

- Canonicalizing Audio Signal

Most of the original audio files are not canonicalized, due to differences in peak amplitudes, resulting in some audio files being quieter compared to others. The canonicalization process involves dividing the audio signal by its maximum absolute value. This ensures that the maximum value of the audio signal is 1.0, while maintaining the relative proportions of the other values in the signal. Note that the canonicalization process does not change the overall volume of the audio signal.

- Segmentation

We need to segment the audio before putting it to the CNN for training to:

✓ Improve efficacy: By dividing the audio signal into smaller segments, CNN can more easily identify patterns and classify each segment more accurately. This can help improve efficacy of tasks such as speech recognition or speaker recognition.

✓ Better understand the audio signal: Segmentation can help break the audio signal into more manageable parts, making it easier to understand and analyze.

✓ Process more efficiently: Segmenting an audio signal allows CNN to process that signal more efficiently because it will only need to work on a small portion of the signal at a time. This is especially useful for large data sets or when working with limited computational resources.

✓ Increase efficiency: Segmentation can also help improve the efficiency of CNN by allowing it to handle variations in the audio signal, such as changes in ambient noise or speaker characteristics.

Overall, segmentation is a useful preprocessing step that helps improve the performance and efficiency of CNN when processing audio data. Therefore, we divide the collected audio files into separate segments, all of which have the same duration of time accordingly.

- Mel Spectrogram Feature Extraction

Once the samples are divided into equal duration, the data is ready for the feature extraction step. The feature extraction process is performed with multithreading to increase speed.

Segmenting the datasets

To ensure that the machine learning model can generalize new data well, we have divided the data into 3 sets: training, validation (val) and testing according to the 8:1:1 ratio.

The training set is used to train the model, such as to optimize the parameters of the model so that it can make accurate predictions on the data.

The validation set is used to evaluate the model's performance during training, by calculating the model's error rate on a set of data that the model has never worked on before. This helps identify when the model is overfitting, for example, when the model starts memorizing the training data instead of learning more general patterns.

The testing set is used to evaluate the model's performance on completely unseen data, after training is complete. This provides a final estimate of the overall model's error and can be used to compare the performance of different models or configurations.

C. Training and Evaluating Results

a. Training hardware environment

The identification system is programmed with Python language and uses the open source library TensorFlow.

Hardware environment used for experiments:

- Operating system: MacOS Monterey 12.6
- CPU: Intel® CoreTM i5-9400F
- RAM: 16GB
- GPU: AMD Radeon™ RX 570 Graphics Card

b. Deep Learning Model Used

Here we use the Xception [5] architecture based on Depthwise Separable Convolution as a more efficient alternative to traditional convolutions. Depthwise separable convolutions split a standard convolution into a depthwise convolution, which applies a separate convolution to each channel of the input, and a point convolution, which combines the result of depth convolutions. This allows the model to learn more complex features while using fewer parameters and requiring less computation, making it easier and more efficient to train the model. In addition to using deep separable convolution, the Xception architecture also incorporates bypass connections, allowing the model to learn more easily by allowing easier gradients across the network. This helps the model converge faster and achieve better results. Overall, the Xception architecture is a high-performance and efficient model for image classification and has been widely used in many applications MobileNetsV2

MobileNet [1] is a lightweight convolutional neural network architecture developed for image classification on mobile devices. The MobileNet architecture is designed to be efficient and lightweight, making it well-suited for use on mobile devices and other resource-constrained platforms.

It achieves this effect using Depthwise Separable Convolutions, which are a variation of standard convolutions that reduce the amount of computation required. The MobileNet architecture combines bottleneck layers and depth-separable bottleneck layers to achieve high efficiency in image classification. MobileNetsV2 models are based on the MobileNetV1 architecture. MobileNetsV2 models achieve better results with fewer parameters and computational overhead than MobileNetV1 models. EfficientNet [9] is a family of convolutional neural network architectures developed for image classification on mobile devices. The EfficientNet architecture is designed to be efficient and lightweight, making it well-suited for use on mobile devices and other resource-constrained platforms.

III. EXPERIMENT AND EVALUATION

After training the data set with the above 3 neural network architectures, we performed evaluation and obtained the results in Table 2, relatively positive compared to the results obtained in "Sound-based Bird Classification" [10] with the highest accuracy of 87% with the bird call dataset of 27 Polish bird species downloaded from xeno-canto.org [15].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 2. Comparison of Test Results Between Models

Model	Accuracy	macro avg			weighted avg		
		precision	recall	f1-score	precision	recall	f1-score
Xception	**0.93**	**0.93**	**0.95**	**0.93**	**0.94**	**0.93**	**0.93**
MobileNetsV2	0.85	0.84	0.90	0.85	0.87	0.85	0.85
EfficientNetsB3	0.90	0.89	0.91	0.89	0.90	0.90	0.90

All three models gave relatively good results. We chose the Xception model with the best results to build a bird sound recognition application.

Table 3. Test Results When Using the Xception Model

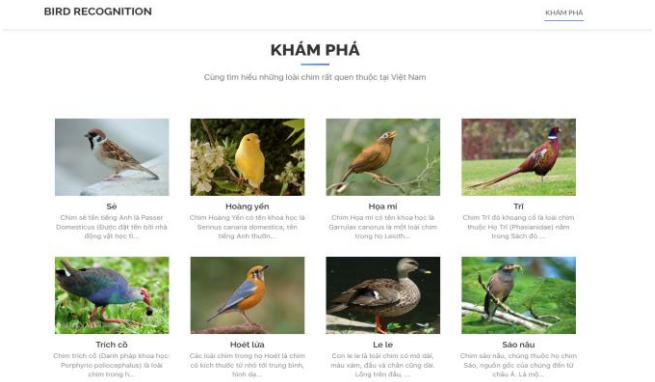
Species	precision	Recall	f1-score
Coucal	1.00	0.88	0.94
Pigeon	1.00	1.00	1.00
Black-collared starling	1.00	1.00	1.00
Slaty-breasted rail	1.00	1.00	1.00
Red-whiskered bulbul	0.98	0.98	0.98
Drongo	1.00	1.00	1.00
Wagtail	1.00	1.00	1.00
Magpie robin	0.93	0.99	0.96
Chóc quạch	0.67	0.91	0.77
Stork	1.00	0.54	0.70
Little Cormorant	0.92	1.00	0.96
Spotted dove	1.00	1.00	1.00
Cuckoo	1.00	0.98	0.99
Gallicrex cinerea	1.00	1.00	1.00
Quail	1.00	1.00	1.00
Snipe	1.00	1.00	1.00
Swallow	0.70	0.28	0.40
Héc xoan	0.95	1.00	0.97
Nightingale	0.96	0.93	0.94
Canary	0.48	0.72	0.58
Olive-winged Bulbul	0.90	0.64	0.75
Blackbird	0.58	1.00	0.74
Orange-headed thrush	0.94	0.79	0.86
Sunbird	1.00	0.96	0.98
Huýt cô	0.75	1.00	0.86
Babbler	0.97	0.97	0.97
White-eye	0.72	0.92	0.81
Lesser whistling duck	1.00	1.00	1.00
Red-eyed	1.00	1.00	1.00
Mắt xéo	1.00	1.00	1.00
Curlew	1.00	1.00	1.00
Five color	1.00	1.00	1.00
Robin	1.00	1.00	1.00
Hornbill	0.94	1.00	0.97
Quế lâm	0.92	1.00	0.96
White-breasted waterhens	1.00	1.00	1.00
Bald coot	1.00	1.00	1.00
Brown starling	1.00	0.89	0.94
Worm bird	1.00	1.00	1.00
Sparrow	1.00	1.00	1.00
Asian Fairy-bluebird	0.94	1.00	0.97
Thanh tước	1.00	1.00	1.00
Yellow-vented bulbul	1.00	1.00	1.00
Phasianus versicolor	0.59	1.00	0.74
Swamphen	0.74	0.94	0.83
Trích ré	0.98	1.00	0.99
Black-crowned night heron	1.00	1.00	1.00
Parrot	1.00	1.00	1.00
Xanh tím	1.00	1.00	1.00

The bird sound identification model introduced in this article will be applied on an online website platform. The website's interface is designed to be simple and user-friendly. At the home page interface, the system will provide a list of bird species in the system and a place where users can download audio files to identify bird species.

The system model of the website is simply built on the client-server model. Clients are browsers integrated on desktop or mobile operating systems such as Firefox or Chrome. The server is the component that manages system operations, receives requests from clients, processes them and provides responses within the allowed time frame. The application will be installed as a sub-component of the Server with the function of calculating and identifying bird species with input data being the audio file provided by the Client. If identified, the bird data will be repackaged with response data at the Server and sent to the Client to display to the user.

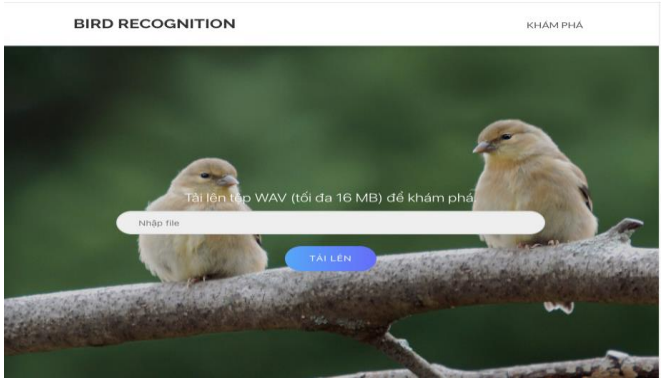
The bird sound recognition website is built using the Python programming language with Django technology. The website's interfaces are programmed with HTML/CSS and Javascript. The process of transactions processing between Client/Server is implemented through the RestFul API. The recommendation system is installed on the Python programming language and built using the open source library Tensorflow.

Users can learn about common bird species in Vietnam available in the system by accessing the Discover section.



Picture 1. Website Interface to View Bird Species List

The system provides bird sound recognition function by allowing users to upload audio files with a maximum wav of 16MB.

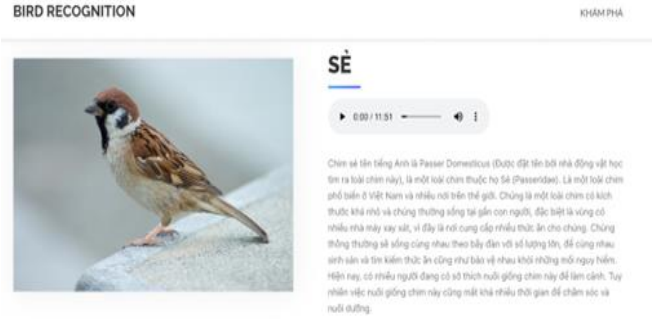


Picture 2. Website Interface for Users to Upload Audio Files

Retrieval Number: 100.1/ijitee.A97561213123
DOI: 10.35940/ijitee.A9756.1213123
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

After user uploads the audio file, the system will record and analyze to find the bird whose sound is most similar to that of the input file. After analyzing and processing, the system will display the results to the user.



Picture 3. Website Interface to View Recognition Results

IV. CONCLUSION

On the basis of studying and researching deep learning to apply to bird sound recognition based on understanding, analyzing and processing the sounds of birds to solve the recognition problem, the article has achieved the following results:

- An overview of the bird sound recognition problem and approaches to solving the problem
- Learning to build software to extract raw noise from the website https://tiengdong.com, learning the CNN model to build a bird sound recognition system. Conducting methodological experiments on self-collected data sets and evaluation of the obtained results. Building an application to identify bird sounds.

DECLARATION STATEMENT

Funding	No, I did not receive.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017) 2-7
2. A. Goh, E. Tan, and R. Go, "Recurrent neural networks for bird sound classification," in Proceedings of the 17th International Society for Music Information Retrieval Conference, 2016, pp. 518-524.
3. A. Dehghani, M. R. Jahromi, and S. A. Monadjemi, "Automatic bird sound classification using recurrent neural network," in Proceedings of the IEEE 2nd International Conference on Applied Robotics for the Power Industry, 2017, pp. 1-6.
4. Y. Xu, H. Liu, H. Zhang, and Y. Yang, "Bird sound recognition based on CRNN and segment-based data augmentation," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 451-455.
5. Franc¸ois Chollet: Xception: Deep Learning with Depthwise Separable Convolutions (2017). https://doi.org/10.1109/CVPR.2017.195
6. K. J. Piczak, "Environmental sound classification with convolutional neural networks," in Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing, 2015, pp. 1-6. https://doi.org/10.1109/MLSP.2015.7324337
7. K. J. Piczak, "Environmental sound classification with convolutional neural networks," in Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2015, pp. 1-6. https://doi.org/10.1109/MLSP.2015.7324337
8. R. R., & Sengupta, A. (2020). Deep learning for bird species classification and sound localization using mel-spectrogram representations. Applied Acoustics, 165, 107355.
9. MingxingTan, Quoc V. Le: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (2020).
10. R. Tsai, C. Liao, and S. Lee, "Automatic bird sound recognition using convolutional neural networks," Applied Sciences, vol. 7, no. 3, pp. 280-295, 2017.
11. Xuedong Huang, Alex cero, Hsiao_wuen Hon (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall
12. Y. Li, F. Xu, J. Zhu, and H. Li, "Deep convolutional neural networks for bird species classification and detection," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 141-14
13. https://tiengdong.com
14. https://towardsdatascience.com/sound-based-bird-classification-965d0ecacb2b
15. Xeno-canto: https://xeno-canto.org
16. Yogapriya*, J., Dhivya, S., & Suvitha, K. (2020). Convolutional Neural Networks in Image Retrieval System. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 5, pp. 123–129). https://doi.org/10.35940/ijitee.d2001.039520
17. Sundararajan*, R. K., S, S., S, V., & Pandian M, J. (2019). Convolutional Neural Network Based Medical Image Classifier. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 4494–4499). https://doi.org/10.35940/ijrte.c6810.098319
18. Kumar, P., & Rawat, S. (2019). Implementing Convolutional Neural Networks for Simple Image Classification. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 2, pp. 3616–3619). https://doi.org/10.35940/ijeat.b3279.129219
19. Das, S., S, S., M, A., & Jayaram, S. (2021). Deep Learning Convolutional Neural Network for Defect Identification and Classification in Woven Fabric. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 1, Issue 2, pp. 9–13). https://doi.org/10.54105/ijainn.b1011.041221

AUTHORS PROFILE

Dr. Phan Thi Ha is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Vietnam, and Computing Fundamental Department, FPT University, Hanoi 10000, Vietnam as well. She received a B.Sc.in Math & Informatics, a M.Sc. in Mathematic Guarantee for Computer Systems and a PhD. in Information Systems in 1994, 2000 and 2013, respectively. Her research interests include machine learning, natural language processing and mathematics applications.n

Th S. Trinh Thi Van Anh is currently a lecturer and a researcher at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Vietnam and Computing Fundamental Department, FPT University, Hanoi 10000, Vietnam as well. She received a B.Sc. in Electronics-Telecommunications in 1993 (HUST), M.Sc. in Computer Science in 1998 (HUST). Her research interests include machine learning, NLP and opinion mining.

Retrieval Number: 100.1/ijitee.A97561213123
DOI: 10.35940/ijitee.A9756.1213123
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)