# Birds Species Audio Data Collection and Classification Using HRNet

Shova Gelal Bhawana Ojha Barosh Manandhar[[1]](#footnote-1)\*

*Department of Artificial Intelligence, School of Engineering, Kathmandu University, Nepal*

*{shovagelal392, bhawanaojha48, barooshmanandhar}@gmail.com*

# *Abstract*

Birds Species Classification using their sound is a very challenging task. In this project we have collected the bird’s species sound recordings from various internet sources of about 62 classes and prepared our own birds species audio classification dataset. We manually listened all the audio files and clean the datasets that were noisy. After that we made the two folders one for training datasets and another for testing datasets. At first, the prepared datasets were converted into their respective spectrograms and then it was fed into HRNET Model for training. We evaluated our model performance using accuracy and loss curve which demonstrated high training accuracy and low validation accuracy leaving it for further other metrices to work upon in future days.

**Keywords:** Artificial Intelligence, HRNET, Spectrograms, Metrices, Accuracy, Audio Classification

**1. Introduction**

Birds plays a crucial role in maintaining ecological balance.[1]Forests are home to nearly two-thirds of all bird species, including more than 70% of globally threatened species. However, forests in Key Biodiversity Areas (KBAs, most of which have been identified for birds) across the world are being lost, fragmented and degraded[1].It influences biodiversity, pollination and pest control. Identifying bird species from their audio recordings is very important for their conservation through their vocalizations but it poses challenges due to overlapping sounds and environmental noise. Traditional methods are labor-intensive and prone to error. In order to solve those issues, an automated solution is required.

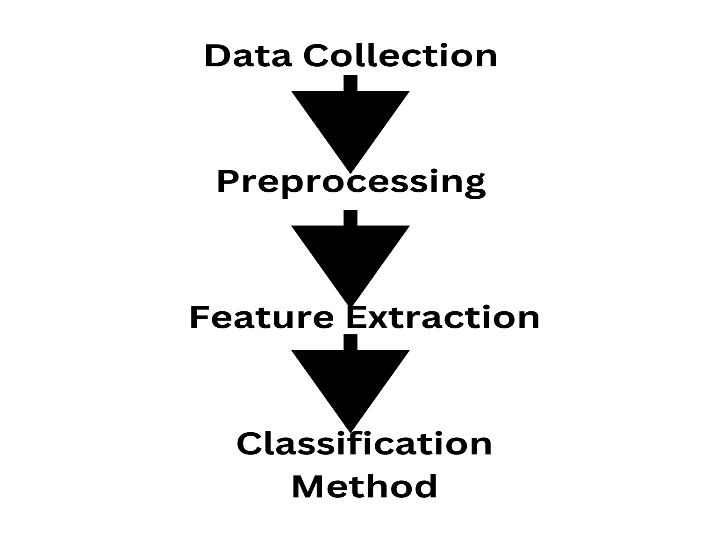
With the arrival of convolutional neural networks (CNNs, ConvNets), automated processing of field recordings made a huge leap forward [1].This project is also focused on creating a Birds Audio Classification System using HRNet , a state-of-the-art deep learning architecture, to classify species from their audio recordings. Mel spectrograms, which capture essential features of bird vocalizations, act as inputs to the HRNet model. By automating bird identification, this system helps ornithologists and conservationists in tracking bird populations efficiently, contributing to biodiversity preservation.

**2. Related Works**

Birds plays a crucial role in maintaining ecological balance…

**3.MATERIALS AND METHODS**

Our bird classification system consists of four different steps.



**Figure 1.** Workflow of the Project

**3.1. Data Collection**

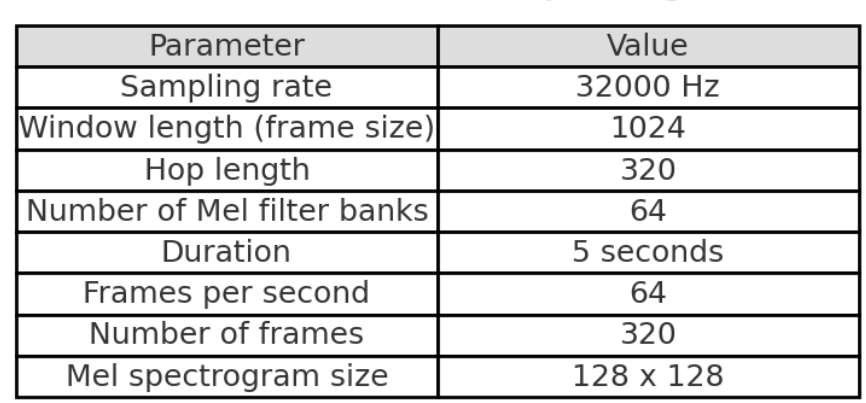
The main work we have done is the preparation of datasets from the different internet sources. We collected datasets of bird’s audio of about 62 classes. These datasets contain bird’s species ranging from hundred data audio data in each species to thousand bird audio data in each class of about 5 second. The name of bird species data that we include are given below:

1. Alexandrine\_Parakeet
2. Bar\_headed\_Goose
3. Baya\_Weaver
4. Black\_breasted\_Weaver
5. Black\_Bulbul
6. Black\_Crowned\_Night\_Heron
7. Black\_Drongo
8. Black\_Kite
9. Black\_Throated\_Thrush
10. Blue\_Tailed\_Bee\_Eater
11. Bronze\_Winged\_Jacana
12. Buff\_Barred\_Warbler
13. Cattle\_Egret
14. Chestnut\_Headed\_Bee\_Eater
15. Chestnut\_Tailed\_Starling
16. Citrine\_Wagtail
17. Common\_Myna
18. Common\_Pochard
19. Demoiselle\_Crane
20. Eurasian\_Coot
21. Eurasian\_Crag\_Martin
22. Eurasian\_Moorhen
23. Eurasian\_Wigeon
24. Fire\_Fronted\_Serin
25. Gadwall
26. Garganey
27. Gray\_Headed\_Lapwing
28. Gray\_Headed\_Swamphen
29. Gray\_Throated\_Martin
30. Great\_Cormorant
31. Green\_Winged\_Teal
32. Hair\_Crested\_Drongo
33. House\_Crow
34. House\_Sparrow
35. House\_Swift
36. Indian\_Pied\_Starling
37. Indian\_Pond\_Heron
38. Jungle\_Myna
39. Kentish\_Plover
40. Large\_Billed\_Crow
41. Lesser\_Kestrel
42. Lesser\_Whistling\_Duck
43. Little\_Cormorant
44. Little\_Egret
45. Mallard
46. Northern\_Lapwing
47. Northern\_Pintail
48. Northern\_Shoveler
49. Pacific\_Golden\_Plover
50. Paddyfield\_Pipit
51. Plum\_Headed\_Parakeet
52. Red\_Billed\_Chough
53. Red\_Breasted\_Parakeet
54. Red\_Rumped\_Swallow
55. Richard\_Pipit
56. Ruddy\_Shelduck
57. Rufous\_Sibia
58. Rufous\_Vented\_Yuhina
59. Scaly\_Breasted\_Munia
60. Slaty\_Headed\_Parakeet
61. Small\_Pratincole
62. Western\_Yellow\_Wagtail

**3.2 Pre-Processing**

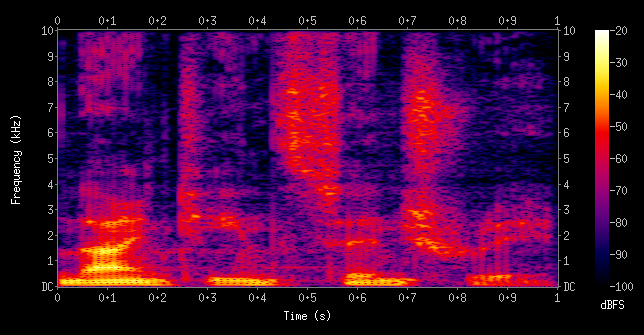
After that our project moved to the crucial phase of audio processing after we had the raw audio data. We manually listened all the audio files and clean the audio files that were noisy and don’t contain any data of the respective class. We found out that few of the audio files had a lot of environmental noise and sounds of humans too.After first, using the librosa library the audio files were read and resampled with a uniform sampling rate of 32 kHz. Stereo audio was converted to mono by averaging channels. To standardize the input length, each audio file is padded with zeros or truncated to a fixed duration of 5 seconds, corresponding to 160,000 samples. The preprocessing also includes normalization to center the audio and scale its amplitude.

**Table 1**. Parameters of Spectrograms



**3.3 Feature Extraction**

Following preprocessing, the script extracts features by computing Log-Mel spectrograms. This involves performing a Short-Time Fourier Transform (STFT) with a Hanning window, mapping the spectrogram to the Mel scale using librosa filters, and converting the result to decibel units to enhance perceptual relevance. The spectrograms are resized to a uniform shape of 128x128 pixels using interpolation, ensuring compatibility with machine learning models.



**Figure 2.** Spectrogram of the spoken words "nineteenth century". Frequencies are shown increasing up the vertical axis, and time on the horizontal axis. The legend to the right shows that the color intensity increases with the density.

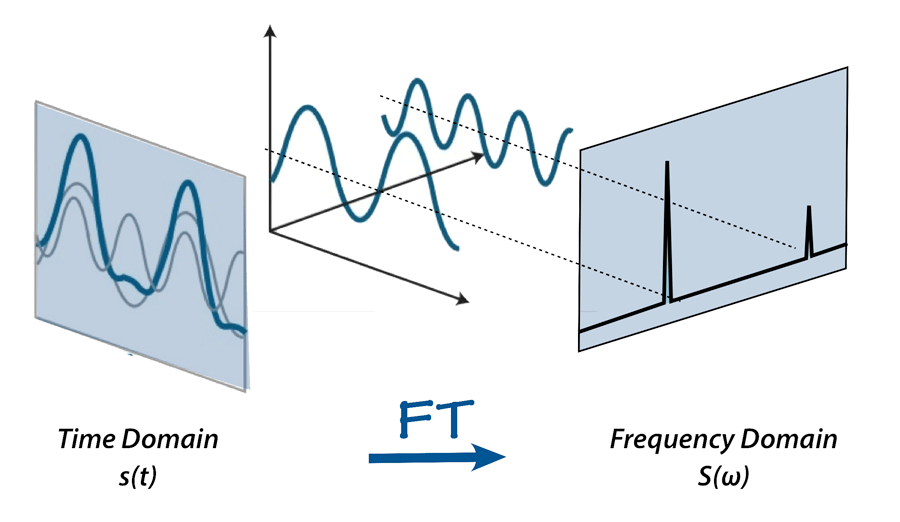
Two essential ideas are combined in a log-Mel spectrogram, which is a representation of an audio signal: the logarithmic transformation and the Mel scale. Its capacity to record perceptually significant aspects of sound makes it popular for use in audio processing tasks, especially in the field of machine learning.

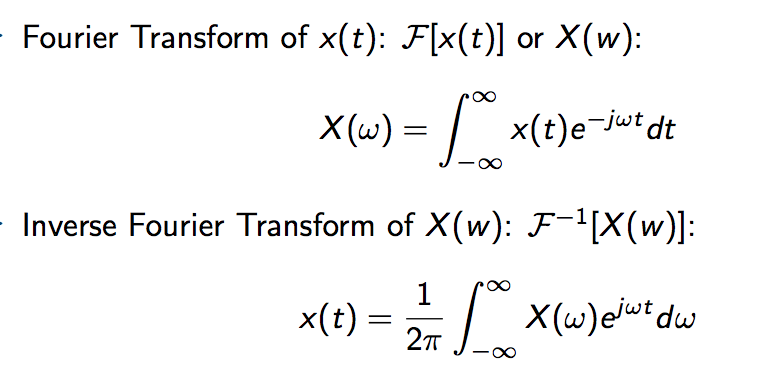
**Components of Log-Mel Spectrograms**

**Audio Signal**

Time-Domain Signal: The raw audio signal is a time-domain representation where the amplitude of sound waves is plotted against time.

**Fourier Transform**

Short-Time Fourier Transform (STFT): To convert the time-domain signal into the frequency domain, the audio signal is divided into short overlapping windows, and a Fourier Transform is applied to each window. This results in a spectrogram, which is a representation of how the frequency content of the signal changes over time.



**Spectrogram**: A plot of frequency (y-axis) vs. time (x-axis), where the color intensity represents the amplitude (magnitude) of the frequency components.

**Mel Scale:**

**Mel Filter Banks**: The Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. It approximates the human ear's response to different frequencies. To create a Mel spectrogram, the linear frequency scale of the spectrogram is mapped to the Mel scale using a set of triangular filters known as Mel filter banks.

**Mel Spectrogram:** The result of applying the Mel filter banks to the power spectrogram. It has fewer frequency bins than the original spectrogram, and the bins are spaced according to the Mel scale.

**Logarithmic Transformation:**

**Log Scale:** The amplitudes in the Mel spectrogram are converted to a logarithmic scale. This transformation is done because the human ear perceives loudness on a logarithmic scale, meaning a change in amplitude at higher volumes is perceived less significantly than the same change at lower volumes.

**Log-Mel Spectrogram:** The final representation where both the frequency axis is on the Mel scale, and the amplitude is on a logarithmic scale. It captures the perceptual aspects of sound more effectively than the linear spectrogram.

Class labels, derived from the folder structure of the audio files, are encoded as one-hot vectors for training purposes. Finally, the processed features, along with their class labels, are saved in .npz format for use in model training and testing. This automated pipeline, implemented in the main function, systematically processes all .mp3 files in a specified directory, saving the output in a structured format. The workflow ensures consistency, scalability, and readiness for downstream machine learning tasks

**3.4 Classification Method**

The categorization of bird species from sound recordings employs a deep learning framework focused on the High-Resolution Network (HR-Net). Pre-extracted Log-Mel spectrograms, which capture both temporal and spectral characteristics of audio signals, act as inputs for the HR-Net. This architecture is specifically crafted to sustain high-resolution representations across the network, allowing accurate feature extraction and strong classification

HR-Net is a convolutional neural network that functions on high-resolution input data representations, improving positional sensitivity and spatial localization in the resulting feature representations. In contrast to conventional convolutional networks that reduce input data to lower resolutions, HR-Net maintains high-resolution features across its entire architecture. This method reduces information loss, rendering HR-Net appropriate for tasks demanding detailed spatial information, like human pose estimation and bird species classification.The structure of HRNet is unique in its approach to maintaining high-resolution representations throughout the entire network. The key characteristics of HRNet are:

1.Parallel multi-resolution subnetworks

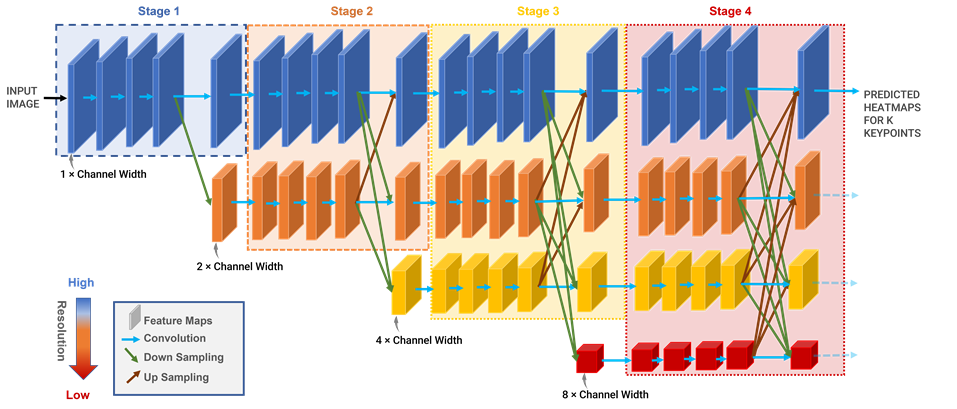
2.Feature fusion across multiple scales utilizing exchange units

3.Output with high resolution

The HRNet architecture is made up of four stages, *Stage 1* to *Stage 4*. The network begins with a high-resolution subnetwork that includes a stem block made up of several convolutional and batch normalization layers to handle the input image. The stem block's output is input into multiple convolutional layers that work at a high resolution. Beginning from Stage 2, the network splits into several parallel subnetworks, with each functioning at a distinct resolution. The initial subnetwork maintains the high-resolution trajectory from Stage 1, whereas other subnetworks handle the features at increasingly lower resolutions.

*Parallel multi-resolution subnetworks*

Each resolution's subnetwork branch contains 4 residual units. Every unit executes two 3×3 convolutions at the corresponding resolution. Stage 1 includes solely the feature maps with the highest resolution, at the standard channel width (the channel count). Every stage following Stage 1 includes the feature maps from all earlier stages, along with feature maps at half the resolution and double the channel width of the lowest resolution from the preceding stage.

The diagram below depicts the advanced HRNet architecture. The horizontal and vertical orientations relate to the network's depth and the feature maps' scale, respectively. The transition from one stage of the network to the subsequent stage preserves high-resolution representations during the entire feature learning process. This results in an accurately localized heatmap estimation that maintains spatial information and reduces time complexity.

**Figure 3.** HRNet Architecture

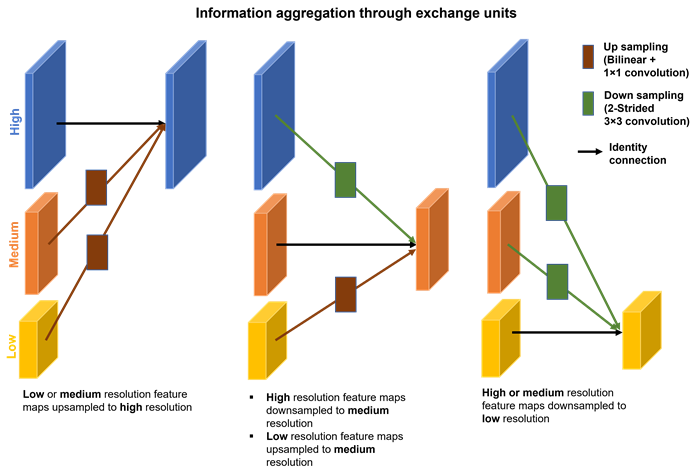
*Exchange units*

HRNet continuously integrates multi-resolution representations at each stage of the process. The network achieves this multiscale feature fusion via multiresolution group convolution utilizing exchange units.

Repeated multiscale feature integration among parallel branches enhances high-resolution representations by utilizing low-resolution representations of the same depth and level. and low-resolution representations enriched by their corresponding high-resolution counterparts.

The high-resolution subnetwork gathers contextual data from lower resolutions, whereas the low-resolution subnetwork obtains detailed spatial data from the high-resolution subnetwork. This interplay results in a highly accurate predicted feature map, preserving both spatial precision and contextual relevance throughout the network

.



**Figure 5.** Information aggregation through exchange units

The The HRNet utilizes the heatmap output from the high-resolution representations of the last exchange unit in Stage 4 for keypoint localization. For instance, to detect KKK keypoints in an object, HRNet generates k heatmap predictions. The network is trained by minimizing the Mean Squared Error (MSE) loss between the K predicted and groundtruth heatmaps. Groundtruth heatmaps are generated by applying a 2-D Gaussian filter centered around the keypoint groundtruth, with a standard deviation of one pixel.

The HRNet architecture begins with a stem layer that downsamples the input and initializes feature extraction. This is followed by multiple branches operating at different resolutions. Each branch employs convolutional blocks to extract hierarchical features, which are then fused across resolutions using dedicated fusion layers.

The network concludes with a global average pooling layer and a fully connected output layer. A softmax activation function predicts class probabilities across 62 bird species, leveraging HRNet's high-resolution representation capability for accurate classification.

This approach utilizes HR-Net's ability for high-resolution representation to categorize bird species based on their distinct audio signals. By maintaining fine-grained spatial and contextual features throughout the network, the method demonstrates strong performance even in challenging acoustic environments.

**4.PERFORMANCE EVALUATION**

The performance of the classification model is evaluated through training and validation accuracy as well as loss metrics. These metrics provide insights into the model's learning process and its ability to generalize to previously unseen data.

**Training and Validation Accuracy**

Accuracy measures the proportion of correctly classified samples during training and validation.Tracking accuracy over epochs helps evaluate the model's learning and improvement over time. A steady increase in training accuracy, accompanied by a similar trend in validation accuracy, indicates successful learning and minimal overfitting.

**Training and Validation Loss**

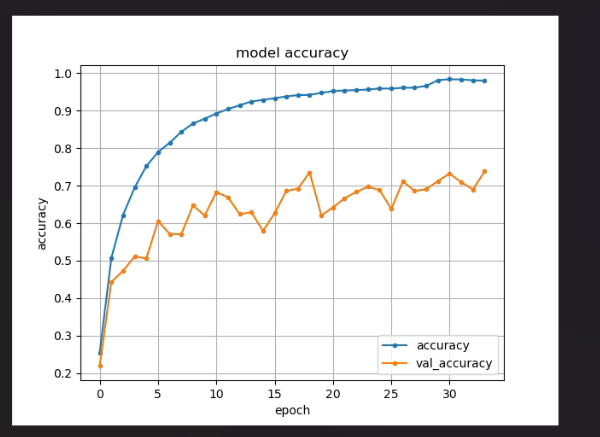
Loss represents the error in the model's predictions. The categorical cross-entropy loss function is used, as it is well-suited for multi-class classification problems. The training loss reflects how well the model adapts to the training data, while validation loss reveals its capacity to generalize to unseen data. A consistent reduction in both training and validation loss over epochs is a positive indicator of model convergence.

**5.RESULTS AND DISCUSSION**

Accuracy, as a measure of evaluation, assesses the ratio of correctly identified bird species compared to the total number of samples. It provides a clear and understandable evaluation of model effectiveness, particularly for multi-class classification problems.

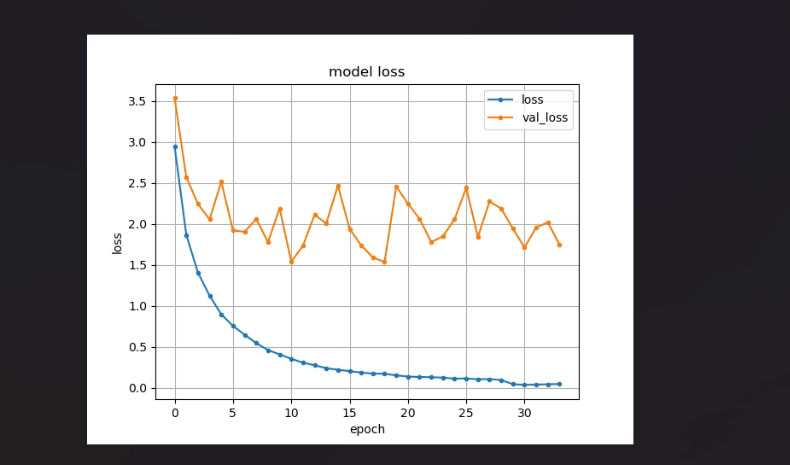
In this research, accuracy functions as a dependable measure of the model's overall performance, guaranteeing that the classification task remains impartial to any particular species. This is especially significant considering the fairly even distribution of bird species within the dataset.

The accuracy graph drawn across the training epochs illustrates the model's development in learning. The HRNet reached impressive training accuracy, approaching 1.0, whereas the validation accuracy stabilized near 0.7 after 10 epochs with slight variations. This divergence indicates the occurrence of overfitting, where the model becomes too familiar with the training data but has difficulty adapting to new data.



**Figure 2.** Model Accuracy Curve

**Loss:** The loss curve presented in our project illustrates the model's performance over the course of training, showing both the training and validation loss across epochs.

****

**Figure 3.** Model Loss Curve

Training loss indicates the inaccuracies in the model's predictions. In this case, the categorical cross-entropy loss function is used due to its effectiveness in multi-class classification problems. The loss curve shows the development of training and validation loss throughout the epochs.

Although the training loss consistently reduced, signifying successful learning from the training dataset, the validation loss varied after an early drop and stayed above the training loss. This inconsistency adds to the evidence of overfitting.

The findings illustrate the HRNet model's robust ability to extract features from the training data, as shown by its almost flawless training accuracy. Nonetheless, the difference in performance between training and validation indicates that the model has difficulty generalizing to new data. This overfitting can be addressed in future efforts by using methods like regularization, dropout, data augmentation, or gathering a larger and more varied dataset.

In spite of these difficulties, the model's capability to correctly classify bird species in controlled environments indicates potential for practical use in real-world scenarios. Additional optimization and testing across different acoustic environments will improve the reliability of the suggested classification system.

**6.CONCLUSION**

This research prepared the Birds Audio Datasets and successfully developed an HRNet-based system for the automatic classification of bird species through audio, demonstrating impressive scalability and accuracy with significant potential for environmental applications. In future we will focus on augmenting datasets to enhance the model's generalization capabilities, exploring transfer learning techniques for better representation of underrepresented species, and integrating real-time classification functions to facilitate field operations..

**REFERENCES**

[0] Authors name., paper title ,Published Place (IJEAT), paper number.

[1] Á. Incze, H. -B. Jancsó, Z. Szilágyi, A. Farkas and C. Sulyok, "Bird Sound Recognition Using a Convolutional Neural Network," *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, 2018, pp. 000295-000300, doi: 10.1109/SISY.2018.8524677.

[1] Authors name., paper title ,Published Place (IJEAT), paper number.

[2] Daidai Liu, Hanguang Xiao, Kai Chen, Research progress in bird sounds recognition based on acoustic monitoring technology: A systematic review, Applied Acoustics, 10.1016/j.apacoust.2024.110285, **228**, (110285).

[3] Jaume Segura-Garcia, Sean Sturley, Miguel Arevalillo-Herraez, Jose M. Alcaraz-Calero, Santiago Felici-Castell, Enrique A. Navarro-Camba, 5G AI-IoT System for Bird Species Monitoring and Song Classification, Sensors, 10.3390/s24113687, **24**, 11, (3687), (2024).

[4] Ingrid Molina-Mora, Viviana Ruíz-Gutierrez, Álvaro Vega-Hidalgo, Luis Sandoval, The utility of passive acoustic monitoring for using birds as indicators of sustainable agricultural management practices, Frontiers in Bird Science, 10.3389/fbirs.2024.1386759, **3**, (2024).

[5] Yi Tang, Chenshu Liu, Xiang Yuan, Recognition of bird species with birdsong records using machine learning methods, PLOS ONE, 10.1371/journal.pone.0297988, **19**, 2, (e0297988), (2024).

[6] Joycelyn Longdon, Jennifer Gabrys, Alan F. Blackwell, Taking data science into the forest, Interdisciplinary Science Reviews, 10.1177/03080188241230415, **49**, 1, (82-103), (2024).

[7] Shobha Lolge, Saurabh Deshmukh, Identification and Classification of Singing Birds Using Machine Learning, Proceedings of 27th International Symposium on Frontiers of Research in Speech and Music, 10.1007/978-981-97-1549-7\_28, (415-428), (2024).

<https://datazone.birdlife.org/2024-annual-update>

<https://en.wikipedia.org/wiki/Spectrogram>

1. \* Corresponding Author [↑](#footnote-ref-1)