Kathmandu University

Department of Artificial Intelligence

Panchkhal, Kavre

A Report on:

"Birds Species Audio Classification Using HRNet Model"

[Code No. AISP 121]

Submitted By:

Shova Gelal(08)

Bhawana Ojha(21)

Baarosh Manandhar(17)

Submitted To

Dr. Yagya Raj Pandeya

Department of Artificial Intelligence

Submission Date: 29 September, 2024

ACKNOWLEDGEMENT

This report for the project entitled *"Birds Species Audio Classification Using HRNet Model"* has been prepared in partial fulfillment of the requirements for the 2nd semester Bachelor of Technology in Artificial Intelligence (B.Tech in A.I.). The project aims to leverage the knowledge gained during this semester to explore the classification model developed for bird's species classification.

We extend our heartfelt gratitude to the Department of Artificial Intelligence (DoAI), School of Engineering (SOE), Kathmandu University (KU), for providing us with the opportunity to embark on this intellectually enriching journey, blending academic theory with practical application.

We want to thank our supervisor Umesh Hengaju Sir for his constant help and guidance during the project. We owe special thanks to Assistant Professor Dr. Yagya Raj Pandey Sir, whose guidance, insights, and encouragement have been invaluable throughout the conceptualization and planning of this project.

ABSTRACT

Birds play a crucial role in balancing the natural ecosystem. Birds Species Audio Classification, while having many beneficial uses, also poses challenges due to its complexity, particularly in noisy environments where sounds from different sources overlap . The automatic classification of bird species from audio recordings is the main goal of this study. Deep learning models are used to recognize different species based on their vocalizations. We convert bird audio data into Mel spectrograms and use HRNet—a network renowned for feature extraction—for classification. The anticipated result is a trustworthy model for classifying bird species, which will support conservation and wildlife monitoring initiatives. Further advancements will focus on improving generality and refining the model.

The main objective of this project is to develop a birds species audio classification model for birds species conservation that enables to balance the ecosystem.

***Keywords**:* HRNet, Birds Audio Classification, Mel Spectograms

TABLE OF CONTENT

ACRONYMS/ ABBREVIATIONS

HRNet - High-Resolution Network

AI - Artificial Intelligence

MFCC - Mel-Frequency Cepstral Coefficients

CHAPTER I

INTRODUCTION

1.1 Background

Birds are essential to the natural ecosystem because they have a direct impact on human health, food supply, ecological balance, and other factors. Although audio classification has many useful applications, its complexity can often be a problem, especially in noisy environments where sounds from several sources overlap. Our study specifically addresses the requirement for a reliable and accurate system that can identify different kinds of birds based just on their sound. There is a need for an automated solution because traditional manual methods of bird identification are frequently laborious and error-prone. With the help of sophisticated machine learning models like HRNet, this study seeks to improve and streamline the process of identifying different bird species based just on their audio recordings.

This project contributes to the growing field of AI-based audio classification by offering a tool that not only simplifies the identification process but also supports efforts in wildlife conservation and ecological studies. The proposed system will use deep learning algorithms to process audio data and extract relevant features, converting the audio into spectrograms that reveal the distinctive characteristics of each bird call. These visual representations of sound are then fed into the HRNet model, which classifies the species.

1.2 Problem Statement

It is essential for ecological study, environmental management, and biodiversity conservation to monitor and identify different bird species by their vocalizations. But manual identification requires a lot of time, labor, and human error—especially when working with big audio files. The inefficiencies of current approaches, especially in remote or densely inhabited locations, make it difficult to monitor bird populations efficiently. These methods also lack automation and scalability.

Using the auditory signals of different bird species, an automated, precise, and scalable system for species classification is the issue. By using mel spectrograms and deep learning models—specifically, HRNet—to classify bird species according to their vocalizations, this study aims to solve this issue and offer a more effective tool for bioacoustic monitoring and conservation initiatives.

1.3 Objectives

1. Classify bird species based on their audio recordings using HRNet Model.
2. Evaluate the effectiveness of HRNet architecture for audio classification tasks.
3. Understand to utilize mel spectrogram features for audio classification tasks.

1.4 Motivation and Significance:

Many bird species are at risk of extinction due to the growing threats to biodiversity, such as habitat loss and climate change. Monitoring bird populations is crucial for understanding the health of ecosystems and putting conservation measures into place. Historically, bird species identification has relied on visual or manual audio observation, which is labor-intensive and prone to human error. Automating this process using machine learning, particularly deep learning, can provide faster, more accurate, and scalable solutions. This project uses audio data to help ornithologists and conservationists track bird species more effectively, contributing to the larger goal of biodiversity preservation. It is driven by the need for an effective, automated system that can accurately classify bird species from audio recordings.

Conservation biology and the science of bioacoustics will be greatly impacted by this initiative. This technology can be used for extensive wildlife monitoring, assisting researchers in tracking population trends and identifying uncommon or endangered species in a variety of settings. It does this by creating a deep learning model to categorize bird species based on their vocalizations. In addition to saving human labor, automation of bird species identification improves the precision and breadth of biodiversity research. This research will support global efforts to conserve biodiversity and manage ecosystems by giving ornithologists and conservationists a useful tool.

This project will provide a tool that can reliably and quickly classify bird species through audio recordings, empowering researchers, conservationists, and hobbyists. Additionally, it will improve our knowledge of bird populations and their habits and supply vital information for research on biodiversity. The benefits of this technology go beyond study; it can help safeguard ecosystems and increase public awareness of the value of protecting bird species.

CHAPTER II

RELATED WORKS

“Investigation of Different CNN-Based Models for Improved Bird Sound Classification," in *IEEE Access*, vol. 7, pp. 175353-175361, 2019, doi:” by J. Xie, K. Hu, M. Zhu, J. Yu and Q. Zhu investigate different CNN based models for improved sound classification. A key function of automatic classification of bird sounds is the observation and preservation of biodiversity. A new method for continually monitoring birds is made possible by recent developments in deep learning algorithms and acoustic sensor networks. Several deep learning-based classification frameworks for the identification and categorization of birds have been developed in earlier research. In order to further enhance bird sound classification ability, we compare various classification models in this work and selectively fuse them. To be more precise, we use two distinct deep learning architectures to build the fused model in addition to using the same deep learning architecture with various inputs. In order to define the various acoustic components of birds, three methods of time-frequency representations (TFRs) of bird sounds are studied: mel-spectrograms, harmonic-component based spectrograms, and percussive-component based spectrograms.

“Bird Sounds Classification Using Linear Discriminant Analysis” by M Ramashini, P. E. Abas, U. Grafe and L. C. De Silva. In this study, bird species from the Borneo region are classified and identified based on their sounds. The researchers focused on five local bird species and applied audio signal processing techniques to extract 35 features from bird sounds. These features were reduced using Linear Discriminant Analysis (LDA) and classified using the Nearest Centroid (NC) method. The proposed method achieved a prediction accuracy of 96%, outperforming more complex algorithms such as Support Vector Machines (SVM) and K-Nearest Neighbor (KNN). This research demonstrates an efficient approach to bird sound classification in dense jungle environments.

CHAPTER III

METHODOLOGY

3.1 Data Collection

In our project, we focused on gathering data from various sources including Kaggle and ebird.org. We collected the data of 62 classes of bird species. This dataset contains the bird's species ranging from hundred data in each species to thousand audio data of bird's species.

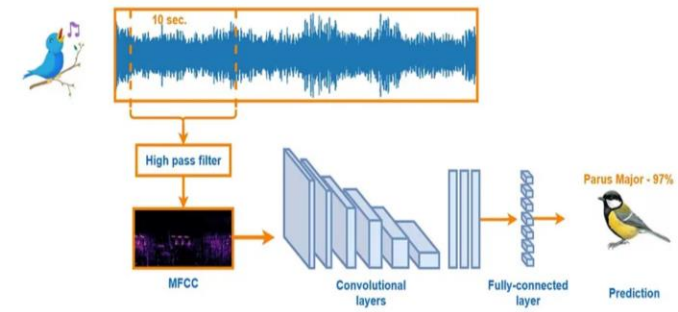
The name of bird species data that we include are given below:

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 Data Preparation



Our project moved on to the crucial phase of audio processing after we had the raw audio data. To ensure uniformity throughout the dataset, we first concentrated on normalizing the audio files to a consistent sampling rate. The preservation of the audio samples' quality and comparability depended heavily on this standardizing process. The audio recordings were then divided into segments of a set length five seconds, in order to prepare them for the analysis that proceeded.

Comprehensive cleaning methods were also used in parallel to remove any unwanted frequencies, distortions, or background noise from the audio. In order to improve the audio signals' clarity and eliminate undesirable frequencies, this phase involved using filters.

After the cleaning stage, sophisticated techniques for extracting features were implemented. We produced Mel-frequency cepstral coefficients (MFCCs), a popular feature set in audio processing that is renowned for encapsulating the fundamental aspects of sound. After that, these MFCCs were organized to serve as the foundation of our dataset for machine learning models.

We made sure the audio data was of the highest caliber and prepared for efficient training and analysis by closely adhering to this audio processing pipeline, which laid a solid basis for our project's success.

3.2.1 Log-Mel Spectrograms

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.

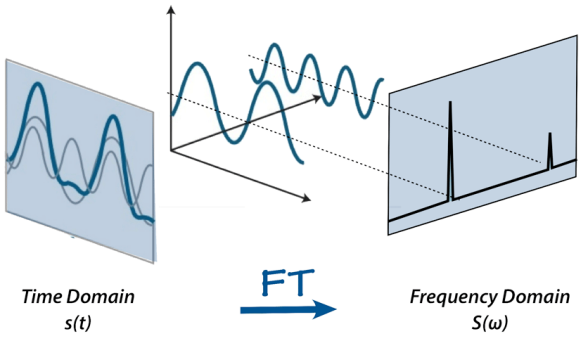
Components of Log-Mel Spectrograms

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.



- Fourier Transform of $x(t)$: $\mathcal{F}[x(t)]$ or $X(w)$:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

- Inverse Fourier Transform of $X(w)$: $\mathcal{F}^{-1}[X(w)]$:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega$$

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.

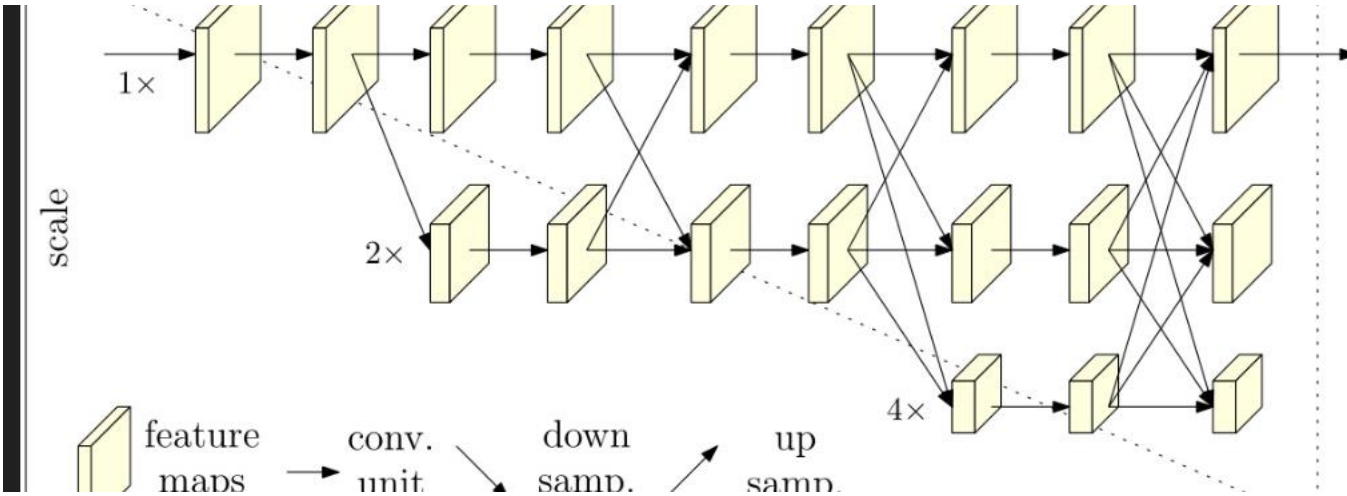
3.3 Training Model: HRNet

These days, deep neural networks are the most widely used machine learning technology. They are now focusing their efforts on environmental sound analysis and classification, having made enormous advancements in speech recognition, image processing, and classification.

Considering the HRNet model is the only one that can sustain high-resolution representations across the network, which is essential for capturing hidden trends and complexity in audio data, we chose it over shallower architectures. In contrast to other models, which downsample features before upsampling them, HRNet maintains fine-grained information throughout all of its layers. This makes it especially suitable for jobs like audio classification from Log-Mel spectrograms that need complex feature extraction.

We were able to extract more detailed information from the audio signals by utilizing HRNet's parallel high-resolution streams, which resulted in better classification accuracy and a more thorough examination of the dataset.

3.3.1 Model Architecture



High Resolution Blocks

Our model's high-resolution blocks serve as its basic construction blocks. Every block, which consists of many concurrent convolutional streams with varying resolutions, maintains high-resolution representations throughout. Fusion layers connect these streams, guaranteeing efficient learning of both coarse- and fine-grained characteristics. Four high-resolution layers, known as stages 1 through 4, make up the model. Each stage has a progressively higher number of convolutional streams. In stage 1, there is only one resolution; in stage 2, there is another, lower resolution; in stages 3 and 4, there are three and four resolutions, respectively. The network can generalize across many tasks because of this design, which enables it to capture features at numerous scales simultaneously.

Parallel Learning

HRNet maintains intricate spatial representations by utilizing parallel learning. Unlike conventional CNNs, which downsample to a single low resolution, this CNN keeps high-quality streams and gradually adds lower-resolution equivalents. For instance, to provide a

more comprehensive representation of the input, the input feature map is analyzed at several scales and then fused together at each step. This architecture enhances the model's overall performance on complex tasks by preserving tiny details while learning wider contextual knowledge.

Fusion Operations

HRNet applies fusion operations at each step to improve multi-scale feature extraction. Through layers of upsampling and downsampling, these procedures integrate data from high-resolution and low-resolution streams, guaranteeing that the result is a balanced representation of both detailed and abstracted elements. For tasks requiring spatial accuracy, such audio classification from spectrograms, the network's capacity to catch complex information at diverse scales is facilitated by the fusion of different resolutions.

Shortcuts

Shortcut links between parallel streams are incorporated in HRNet to facilitate information sharing at various resolutions. These shortcuts, which consist of batch normalization and convolutional layers, make sure that the abstracted information learnt by the lower-resolution streams is useful to the higher-resolution streams. By assisting in the mitigation of the vanishing gradient issue, this architecture enhances overall learning efficiency and facilitates the training of deeper networks.

3.4 Evaluation Metrics

As we analyze the accuracy and performance of our music sheet generation model, we will rely on several key metrics such as:

Accuracy: Since accuracy provides a straightforward and understandable assessment of our model's performance, we chose it as one of our main evaluation indicators. The accuracy ratio shows how well our model distinguishes between various bird calls: it computes the ratio of correctly classified bird species to the total cases. Since the bird categorization job is categorical in nature, accuracy is a good measure of how well our predictions performed generally across all species. This metric ensures that performance

is not biased towards any specific species and is especially helpful in situations where the class distribution among bird species is fairly balanced.

CHAPTER IV

SYSTEM REQUIREMENTS AND SPECIFICATIONS

For the compulsions of this project, powerful hardware systems were required as the AI system
demanded complex matrix operations done on higher dimension data of images. A system of
at least 64 GB RAM and over 64 GB of GPU was required to train the supervised model.

About software requirements, any powerful IDE (Integrated Development Environment) was
required.

4.1 Software Requirements

1. IDE (VS code.)
2. Google Colab

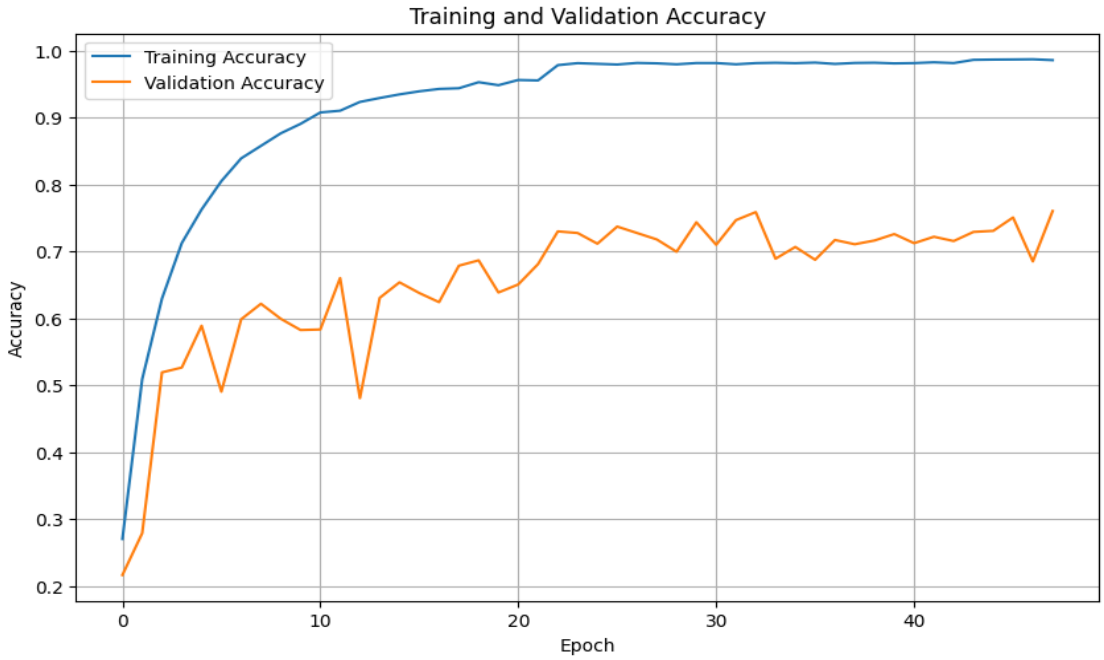
4.2 Hardware Requirements

1. 64 GB of RAM strictly
2. 64 GB of GPU strictly
3. 1 TB External Hard Drive

CHAPTER V

RESULT AND ANALYSIS

Accuracy: The accuracy curve presented in your project illustrates the model's performance over the course of training, showing both the training and validation accuracy across epochs.



Training Accuracy Curve (Blue Line) : The training accuracy starts low and increases steadily as the model learns from the data. By around the 20th epoch, the training accuracy approaches near 100%, indicating that the model is fitting the training data almost perfectly. However, this near-perfect accuracy may signal potential overfitting, especially if the model starts memorizing the training data instead of generalizing well.

Validation Accuracy Curve (Orange Line): In contrast, the validation accuracy rises sharply at first, peaking at 60% by the tenth epoch. From then on, though, it varies between 60% and 70% without exhibiting any discernible improvement. These variations imply that the model is having difficulty generalizing effectively to previously unseen validation data. After the first few epochs, there is a noticeable discrepancy in the accuracy between training and validation runs, which could possibly be an indication of overfitting, a condition in which the model performs well on training data but poorly on fresh, untested data.

CHAPTER VI

CONCLUSION

Our project focuses on developing a strong method for categorizing different bird species according to the audio recordings they make. We extracted useful information from bird sounds and developed an effective classification system by utilizing machine learning models and sophisticated audio processing techniques. By helping scientists track bird populations, this endeavor not only supports the conservation of animals but also plays a crucial role in biodiversity study. Eventually, our efforts contribute to the preservation of bird species and their environments, guaranteeing their survival and deepening our knowledge of the natural world for coming generations.

LIMITATIONS

We must be aware of the various limitations of our model; in spite of the methods and strengths we employed in this project. These restrictions provide light on possible areas for development and future lines of inquiry. The restrictions include high computational and complexity needs, and more.

FUTURE ENHANCEMENTS

There are a few improvements that might be made to our project to further boost its effectiveness and usefulness. These improvements are intended to overcome the existing constraints, investigate fresh avenues for scientific advancement, and increase the project's usefulness. Examining alternative model topologies may improve the model's capacity to produce superior outcomes. Further research could be extended to include additional audio classification areas.

REFERENCES

"Investigation of Different CNN-Based Models for Improved Bird Sound Classification," in IEEE Access, vol. 7, pp. 175353-175361, 2019, doi:" by J. Xie, K. Hu, M. Zhu, J. Yu and Q. Zhu https://arxiv.org/abs/1809.00888

.

"Bird Sounds Classification Using Linear Discriminant Analysis" by M. Ramashini, P. E. Abas, U. Grafe and L. C. De Silva arxiv.org/abs/2303.15823