e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

A SURVEY: IMAGE AND AUDIO BASED RECOGNITION OF BIRD SPECIES

Prof. Tanuja Mulla A*1, Nagaonkar Danish*2, Toshniwal Varun*3, Ghodke Vrushabh*4, Shinde Tushar*5

*1,2,3,4,5Research Scholar, Department Of Computer Engineering, (SKNCOE- Vadgaon), SPPU Pune, India.

DOI: https://www.doi.org/10.56726/IRJMETS63893

ABSTRACT:

This study explores the integration of audio and image-based recognition techniques to identify bird species through a comprehensive workflow utilizing spectrograms and Convolutional Neural Networks (CNNs). As biodiversity continues to decline, effective monitoring of avian populations is crucial for conservation efforts. Our approach leverages audio recordings of bird calls, which are transformed into spectrograms—visual representations of sound—enabling the extraction of temporal and frequency features. Simultaneously, images of birds captured in their natural habitats are analyzed to complement the audio data.The workflow begins with the collection of diverse datasets comprising both audio and image samples of various bird species. Audio data is processed to generate spectrograms using Short-Time Fourier Transform (STFT), while images are preprocessed for uniformity. A CNN model is then designed to process these spectrograms and images, with a dual-input architecture that allows simultaneous learning from both modalities. This model is trained using transfer learning techniques on pre-trained networks to enhance performance and reduce computational requirements.

Keywords: Bird Species Identification, Spectrogram Analysis, Convolutional Neural Network.

I. INTRODUCTION

The decline in global biodiversity, particularly among avian species, has raised significant concerns among conservationists, ecologists, and researchers. Birds play crucial roles in ecosystems as pollinators, seed dispersers, and indicators of environmental health. However, habitat destruction, climate change, and human encroachment have led to an alarming decrease in bird populations worldwide. Effective monitoring and identification of bird species are essential for conservation efforts and ecological studies, prompting the need for innovative and efficient recognition methods.

Traditional bird identification methods often rely on expert knowledge and field surveys, which can be time-consuming and labor-intensive. Recent advancements in technology, particularly in audio and image processing, present new opportunities for automating bird species recognition. Audio recordings of bird calls and songs are valuable resources, as they capture not only the presence of species but also their behavior and interactions. The transformation of audio signals into spectrograms—visual representations that display frequency over time—enables the extraction of rich acoustic features that can be analyzed using machine learning techniques.

Simultaneously, visual data obtained from photographs or videos can provide complementary information about bird species, such as morphology and coloration. By integrating both audio and image data, researchers can leverage the strengths of each modality, improving classification accuracy and robustness. Convolutional Neural Networks (CNNs), a powerful class of deep learning algorithms, have shown remarkable success in image recognition tasks and can be adapted for analyzing spectrograms as well.

This study aims to develop a comprehensive workflow for the simultaneous recognition of bird species using audio and image data. By employing a dual-input CNN architecture, we explore how the integration of acoustic and visual features enhances the model's ability to accurately classify bird species. The findings of this research are expected to contribute to the fields of wildlife conservation and ecological monitoring, providing tools that facilitate the tracking of avian populations and support conservation strategies in the face of ongoing environmental challenges.

By harnessing the power of advanced machine learning techniques and audio analysis, this project seeks to empower researchers, conservationists, and citizen scientists alike. Our goal is to provide a scalable and efficient tool for bird species-identification that not only enhances the understanding of avian biodiversity but also fosters greater engagement in conservation efforts. Through this work, we aspire to bridge the gap between technology and ecology, ultimately contributing to the reservation of our planet's rich avian diversity.

II. METHODOLOGY

Traditional bird identification methods often rely on expert ornithologists and extensive field surveys,which can be labor-intensive and limited by geographical and temporal constraints. The increasing volume of acoustic data collected through citizen science initiatives and automated recording devices highlights the necessity for automated systems capable of processing and analyzing these recordings. By automating the identification process, we can significantly reduce the time and effort required, enabling broader participation in biodiversity monitoring.

A. Spectrograms

Spectrograms are visual representations of the frequency content in an audio signal over time. They show how the energy of different frequencies changes over a given period, allowing us to analyze patterns and characteristics of sound that might be less apparent in the raw waveform.

Here's a breakdown of how they work:

Decomposing the Audio Signal

Audio signals are complex waveforms that can be broken down into individual sine waves of different frequencies using a mathematical process called the **Fourier Transform**. In practice, a variant called the **Short-Time Fourier Transform (STFT)** is often used, which splits the audio into small, overlapping time windows and calculates the frequency spectrum for each.

Frequency and Time

In a spectrogram, the **horizontal axis** represents time, and the **vertical axis** represents frequency. For each time window, the STFT computes how much energy exists at each frequency. This produces a matrix of values: time windows on one axis, frequencies on the other.

Amplitude (Energy) as Color or Intensity

Each cell in this matrix represents the amplitude or intensity of a particular frequency at a specific time. This amplitude is shown as color (or brightness) in the spectrogram. High-intensity (loud) frequencies might be represented by brighter colors, while lower-intensity (quiet) frequencies appear darker.

Logarithmic Scaling

To capture the full range of frequencies that humans can hear, frequencies are often displayed on a **logarithmic scale**. This helps emphasize lower frequencies, which are typically more relevant in audio analysis.

Interpreting Spectrograms

Spectrograms are valuable for identifying distinct features in audio. For example, you can see patterns that correspond to speech phonemes, musical notes, or environmental sounds. They are widely used in audio processing applications like music analysis, speech recognition, and even in medical fields like analyzing heart and lung sounds.

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are highly effective for image detection and classification tasks because they can automatically learn features in images through a series of convolutional and pooling operations. CNNs mimic the way humans process visual information, breaking down images into hierarchical layers of features to detect edges, shapes, and complex patterns. Here's how they work:

Convolution Layer: Learning Local Patterns

The convolution layer is the core of a CNN. In this layer, small filters (kernels) scan over the input image, learning small features like edges, textures, and patterns. Each filter slides across the image, performing a mathematical operation called **convolution**, which multiplies the filter values with the pixel values and sums the results. This process produces a **feature map** that highlights where specific features occur in the image.

Early layers detect basic features like edges and corners, while deeper layers capture complex features like textures and shapes. Multiple filters are used per layer to capture different features.

Activation Function: Adding Non-Linearity

After the convolution operation, an **activation function**—usually the **ReLU (Rectified Linear Unit)**—is applied to introduce non-linearity. ReLU replaces negative values with zero, allowing the network to model more complex relationships. Without ReLU (or similar activation functions), the CNN would act as a linear model, unable to learn complex patterns essential for image recognition.

Pooling Layer: Reducing Spatial Dimensions

Pooling layers reduce the size of feature maps, which decreases computational requirements and makes the model more efficient. The most common pooling technique is **max pooling**, which divides the feature map into regions and retains only the highest value in each region. This process reduces the spatial dimensions while preserving important features, making the network more robust to small changes and distortions in the image.

Stacking Convolution and Pooling Layers

Convolutional and pooling layers are stacked to progressively detect higher-level features. For example: First layers might detect simple edges and textures. Intermediate layers detect parts of objects, like corners, circles, or other shapes. Deeper layers detect complex, task-specific features, like faces or objects.

Fully Connected Layers: Learning Non-Local Patterns

After several convolutional and pooling layers, the output is "flattened" into a one-dimensional vector.This flattened vector is fed into **fully connected layers**—traditional neural network layers where each neuron is connected to every neuron in the previous layer. These layers combine the learned features from previous layers to make final predictions. The final layer typically uses a **softmax** activation function for classification, assigning probabilities to each class.

Training and Backpropagation

During training, CNNs use **backpropagation** to adjust weights in the network based on the error of their predictions.

A loss function calculates the difference between predicted and actual outputs, and backpropagation adjusts weights to minimize this loss.

Application to Object Detection

In object detection, CNNs not only classify images but also identify object locations. Techniques like **Region-Based CNNs (R-CNN)**, **YOLO (You Only Look Once)**, and **SSD (Single Shot Multibox Detector)** extend CNNs to localize objects by:

Proposing regions in the image that might contain objects.Applying CNNs to each proposed region to classify and localize objects.Using bounding boxes to indicate object locations, combined with probability scores for each class.

III. RECENT WORKS

1. In year 2018 IEEE published a paper named Automatic acoustic detection of birds through deep learning : The First Bird Audio Detection Challenge.

Study done in that paper was assessing the presence and abundance of birds is important for monitoring specific species as well as overall ecosystem health. Yet acoustic monitoring is often held back by practical limitations such as the need for manual configuration, reliance on example sound libraries, low accuracy, low robustness and acoustic conditions.

2. In year 2016-IEEE published a paper named Bird Detection in Audio: A Survey and A Challenge.

Study done in that paper was many biological monitoring projects rely on acoustic detection of birds. Despite increasingly large datasets, this detection is often manual or semi-automatic, requiring tuning/postprocessing. We review the state of the art in automatic bird sound detection, and identify a widespread need

for tuning free and species agnostic approaches.

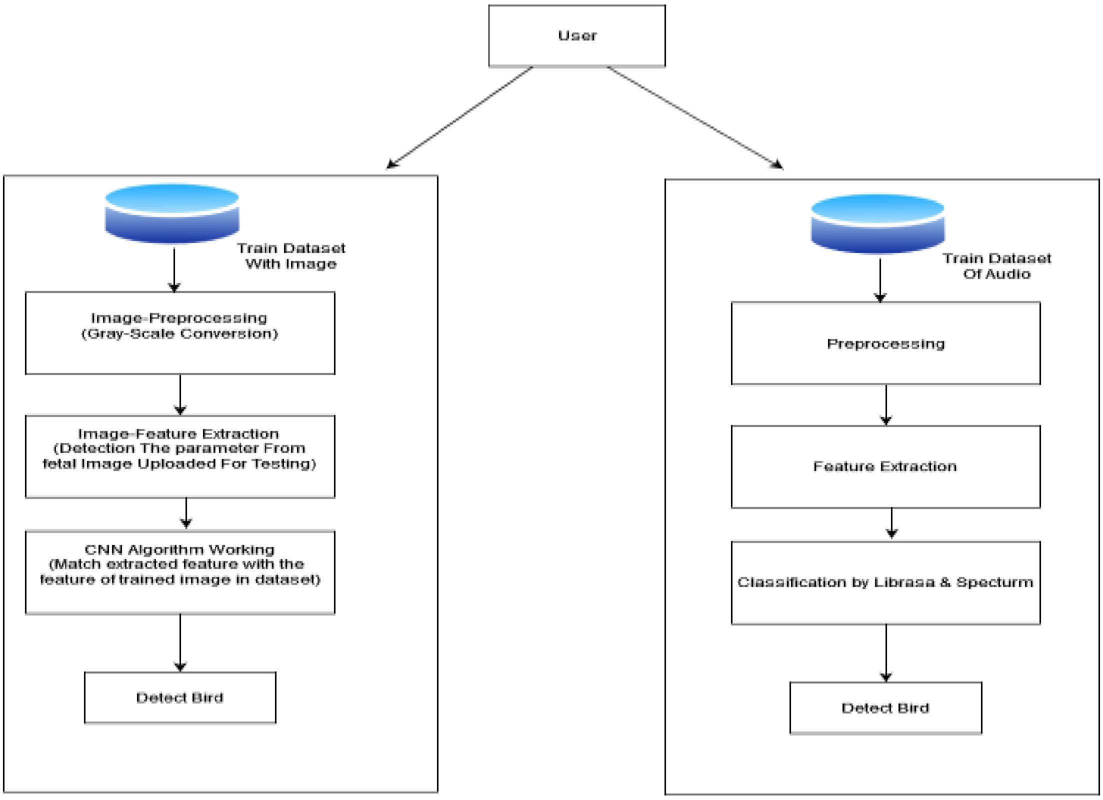
3. In year 2024 there was a paper published named Bird Sound Identification using Deep Learning.

This paper identifies the model for classifying the bird sounds using the Convolutional Neural Network (CNN) algorithm. CNN are considered as one of the powerful toolkits developed in the machine learning and that have shown to be more efficient particularly in the field of image processing and sound recognition.

4. In year 2023 there was a paper published named Automatic acoustic detection of birds through deep learning: The First Bird Audio Detection Challenge.

This paper identifies the model for classifying the bird sounds using the Convolutional Neural Network (CNN) algorithm. CNN are considered as one of the powerful toolkits developed in the machine learning and that have shown to be more efficient particularly in the field of image processing and sound recognition.

IV. OBSERVATIONS AND FINDINGS


User
Train Dataset With Image
Image-Preprocessing (Gray-Scale Conversion)
Image-Feature Extraction (Detection The parameter From fetal Image Uploaded For Testing)
CNN Algorithm Working (Match extracted feature with the feature of trained image in dataset)
Detect Bird
Train Dataset Of Audio
Preprocessing
Feature Extraction
Classification by Librasa & Specturm
Detect Bird

As per the observation of the recent works, findings are:

A] Key Issues & Insights

Data Quality and Quantity:

Current State: Limited dataset, covering only a few bird species or limited geographical regions.

Gap: Need for a larger and more diverse dataset that includes rare species, different geographical regions, and more diverse environments.

Algorithm Sophistication:

Current State: Basic algorithms with limited accuracy.

Gap: Lack of use of deep learning and neural networks. Incorporating models like CNNs improve performance.

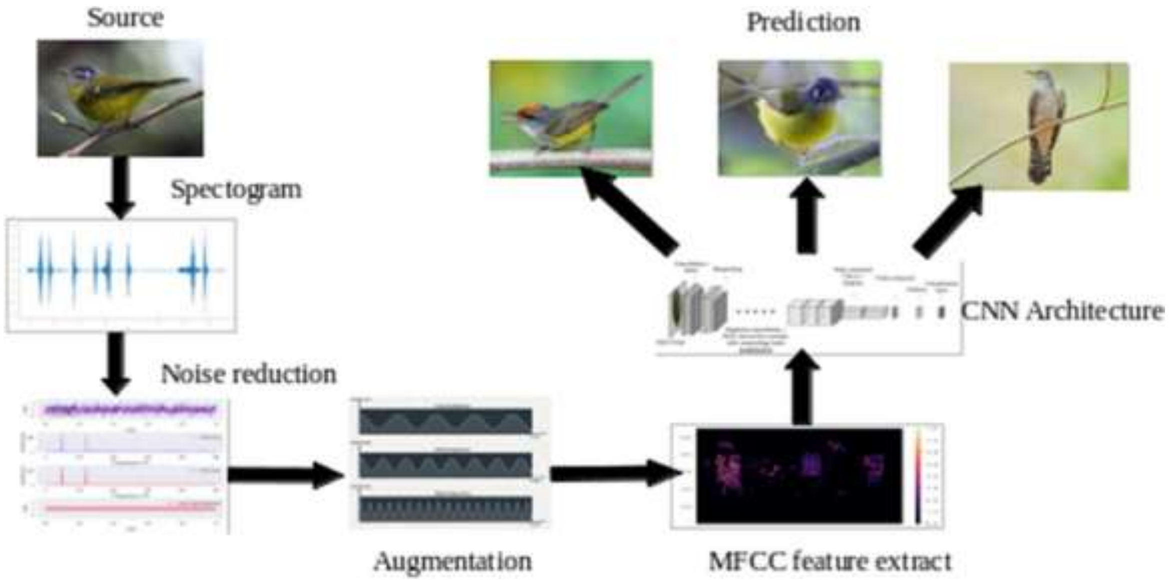
We can eliminate or reduce all these issues by using spectrogram and CNN models.We can use diverse datasets from python libraries to tackle the issue of limited dataset. Python provides numerous libraries which we can use.

As we know, a **spectrogram** is a visual representation of the frequencies within a sound signal over time. It breaks down an audio signal into its frequency components, allowing us to see how different frequencies vary in intensity as time progresses. This makes it incredibly useful in fields like music analysis, speech recognition, and bioacoustics (studying bird sounds).

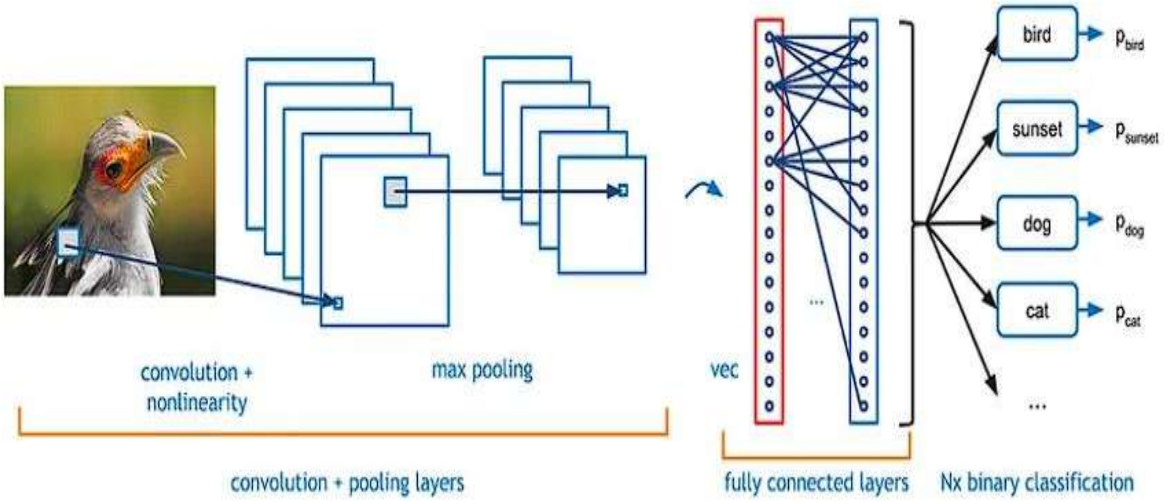
Also, Convolutional Neural Networks (CNNs) are highly effective for image detection and classification tasks because they can automatically learn features in images through a series of convolutional and pooling operations. CNNs mimic the way humans process visual information, breaking down images into hierarchical layers of features to detect edges, shapes, and complex patterns.

Here's how Spectrograms and CNNs work:

A. Working of Spectrogram



B. Working of CNN



V. RESULTS

The integration of audio and image-based recognition systems for identifying bird species using spectrograms and Convolutional Neural Networks (CNNs) marks a significant advancement in wildlife monitoring and conservation. By leveraging both vocalizations and visual characteristics, these systems enhance accuracy and provide a robust tool for researchers and conservationists. The ability to automate species identification facilitates large-scale monitoring, allowing for real-time data collection and analysis, which is crucial for tracking biodiversity and conservation efforts. Looking ahead, there is immense potential for improvement through the development of more sophisticated algorithms, real-time deployment in field studies, and integration with other data types like environmental conditions. Additionally, creating user-friendly applications for citizen scientists can broaden data collection efforts, while localized models can enhance accuracy for specific regions. As ethical considerations regarding data use become increasingly important, establishing responsible guidelines will be essential. Overall, the future of bird species recognition systems promises to play a vital role in ecological research and conservation strategies, ultimately contributing to the protection of avian species and their habitats in a changing world.

VI. FUTURE WORK

Furthermore, the future of audio and image-based bird species recognition systems holds exciting possibilities for innovation and collaboration. Advances in machine learning techniques, such as transfer learning and hybrid models that incorporate both CNNs and recurrent neural networks, can lead to even higher classification accuracy. Real-time monitoring systems can be developed for deployment in natural habitats, enabling continuous observation without human interference. Additionally, by integrating geographic and environmental data, researchers can gain deeper insights into species behavior and habitat preferences. The development of mobile applications aimed at citizen scientists will encourage public engagement and expand data collection efforts, fostering a community-driven approach to biodiversity conservation.

VII. REFERENCES

[1] D Stowell, M Wood , H Pamula, H Glotin , " Automatic acoustic detection of birds through deep learning : The First Bird Audio Detection Challenge ",IEEE , July 2018,vol1

[2] Yannis Stylianou , Mike Wood, Dan S, Herve G, Bird Detection in Audio: A Survey and A Challenge, IEEE, September, 2016, vol1,Italy

[3] A. Thakur, V. Abrol, P. Sharma, and P. Rajan, Bird Sound Identification using Deep Learning, 2024

[4] C.-H. Lee, C.-C. Han, and C.-C. Automatic acoustic detection of birds through deep learning: The First Bird Audio Detection Challenge, 2023

[5] Xeno-canto, https://www.xeno-canto.org/, 2018, [Online; accessed 2018- 06- 20].

[6] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo, Automated species recognition of antbirds in a mexican rainforest using hidden markov models, The Journal of the Acoustical Society of America, vol. 123, no. 4, pp. 24242431, 2008.

[7] C.-H. Lee, C.-C. Han, and C.-C. Chuang, Automatic classification of bird species from their sounds using two-dimensional cepstral coeffi- cients, IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 8, pp. 15411550, 2008

[8] A. Harma and P. Somervuo, Classification of the harmonic structure in bird vocalization, in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP04). IEEE International Conference on, vol. 5. IEEE, 2004, pp. V70