# Proposal for a Bird Sound Identification System

Sarthak Sharma
ssharma19@earlham.edu
Earlham College
Richmond, Indiana, USA

## ABSTRACT

This research study aims to identify bird sounds from recordings that contain a mixture of sounds, including those from birds, animals, and insects. Building on the work of the Nobel Research Institute [11], the research has the potential to assist social biologists in determining the presence of a particular bird species at a given location by evaluating the number of bird counts [1]. However, due to the wide range of bird cries and the challenges in recognizing them, there is currently no approach that can provide 100% accuracy in automating bird call recognition from audio recordings. This study uses a machine learning approach, specifically a deep convolutional neural network architecture, to extract bird sounds from mixed recordings of animal and bird sounds and identify the bird species from the extracted recordings. Our aim is to achieve a recognition accuracy of over 70%. The deep learning convolutional neural network has seen the most significant progress in the automatic bird call recognition challenge, as demonstrated by benchmarking challenges such as the LifeCLEF Bird Competition (BirdCLEF) [12].

## 1 IMPORTANCE OF MY PROPOSAL

The identification of bird sounds has numerous applications and is a crucial task in wildlife research. Through identifying bird sounds, researchers can gain insight into how different species interact with their environment and how they are affected by changes in climate or other factors. For instance, recent research focuses on how migrating birds respond when they encounter barriers to movement, such as mountains or oceans [3]. Moreover, bird sound identification has immense educational potential. Teachers can use recordings of different birds to help students learn about these animals and their habitats. As a bird does say a lot more than just make sounds - *Most children only consider one sound, but an observant child with more experience in bird sounds might ask, "What kind of bird?"* With the help of this research, teachers can provide students with access to sounds of different bird species, which can help answer their questions [9]. The field of bird sound identification is vast, and there are many unanswered questions that need to be explored. This topic is of interest to me as it has the potential to provide solutions to pressing environmental concerns, such as deforestation, climate change, and others [5][4].

## 2 PLANNED CONTRIBUTIONS

I will be using a publicly available data set of bird audio recordings found on Kaggle. The Bird CLEF 2021 competition provides the data set and is available to all [2]. The data set includes over 1,000 species of birds from over 150 countries. I will convert the recordings into spectrograms. I will use CNN to identify as many birds as I can available under the folder 'train-short-audio'. These audio files have been downsampled to 32 kHz and converted to the ogg format to match the test set audio. Once my CNN has been successfully trained on the short audios I will move on to identifying the folder 'train-soundscapes' [2]. This folder contains audio files that are similar to the test set. They are all about 10 minutes long and in the ogg format. Using a CNN I created, I will determine the species of bird based on the more complicated patterns in the bird noises. By training the CNN on a large number of recordings (in this case 'train-short-audio'), the algorithm can learn to make accurate predictions on the type of bird based on their call. [13] This is an effective way to quickly identify bird species in large audio recordings (in this case the 'train-soundscapes' recordings will be used).

## 3 BACKGROUND RELATED / WORK SECTION

### 3.1 Recording

For my research study, I will be using recordings and data set of birds available publicly at Kaggle [2]. There is a long-standing practice of monitoring bird populations by conducting point count surveys [7]. At sampling locations, the observer will visually and aurally count every bird in a given time window (3 or 5 minutes). However, this process can be quite time consuming and requires expert knowledge in the identification of birds. With advances in technology however, there are now new ways to monitor bird populations [7]. One of the main objectives for biologists and ecologists is to be able to collect and analyze data at large spatial scales to monitor the status, trends, distribution, and habitat use of wildlife species at various geographic locations.

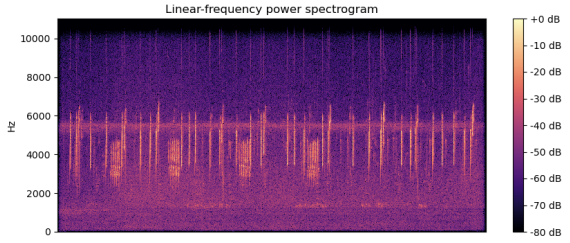


**Figure 1: Audio recording converted to a spectrogram showing many different species over a period of 10 seconds**
*Doi - https://doi.org/10.1371/journal.pone.0179403.g007*

### 3.2 Converting Audio to Visual Graphs

Converting the data set of short bird recordings into spectrograms is important. I converted all the audio files through my Python code which can load audio files using Librosa library, converts it to a spectrogram, and then plot it using the Matplotlib library. The resulting plot is a visualization of the audio signal's frequency content over time.

The method specifically reads a number of audio files provided by their 'path' argument first. Audio files are loaded as a mono or stereo signal depending on the format. The signal is then transformed into a floating point representation. The algorithm then uses the Short-Time Fourier Transform (STFT) to generate the spectrogram, which represents the signal's power spectrum over time. The amplitude of the STFT is converted to decibels using Librosa's 'amplitude_to_dB' function. Finally, the spectrogram is shown using Matplotlib's'specshow' function, with the linear frequency scale shown on the y-axis.

The resulting plot (Fig 1) shows the time on the x-axis, and the frequency on the y-axis. The intensity of each point in the plot represents the power of the signal at that time and frequency. I used the 'colorbar' function to display a color scale that corresponds to the intensity values in decibels.

### 3.3 Preprocessing

The next step is sorting, processing, and species identification from the transformed spectrograms. After we have successfully converted the audio into spectrograms we would be using CNN architecture to identify the sounds of the birds. The aim for my research would be introducing a new CNN structure that will aim in getting an accuracy above the already running softwares, for example the ResNet-50, a deep convolutional neural network architecture for automated bird call recognition. It has a 62 percent accuracy in identifying bird species [12].

### 3.4 Characteristics Of CNN

A convolutional neural network (CNN) is a special class of neural network that is built with the ability to extract unique features from image data. To train a CNN to help identify bird species from their bird sounds, the first step is to provide the CNN with a large data set of bird sounds and their corresponding species labels. For this we will be using data set provide by Bird CLEF 2021 [2] named 'test-soundscapes'. To obtain accurate findings, we will employed CNN to classify the bird species. Once we successfully train the CNN to identify bird species we can collect more data, and further refine the model also if we have enough time in hand by the end of the next semester. CNNs can notice a variety of patterns in spectrograms. These patterns include frequency and amplitude, as well as rhythm and timing. Additionally, CNNs can also detect features such as formants, which are frequencies that appear in bird songs and calls. [14] Finally, CNNs can also detect harmonic structures, which are patterns of frequencies that occur in bird sounds.

### 3.5 How a CNN Operates

A CNN is a type of artificial neural network (ANN) that is used for image and sound recognition. The CNN is composed of multiple layers, with each layer processing the data in a different way. The layers are connected together, and they can be adjusted to improve the accuracy of the model. The CNN takes an input such as an image or sound recording, and it processes the data to identify patterns. CNNs have been used by researchers to identify bird sounds from audio recordings, with an accuracy of up to 70%. In a paper published in 2019 [15], researchers used a two-stage CNN-based model to identify bird species from audio recordings. The first stage used a CNN to extract features from a given audio recording, and the second stage used a CNN to classify the audio recording based on the extracted features. The researchers evaluated their model on two avian sound datasets, and achieved an accuracy of 70.4% and 70.7%, respectively. This demonstrates the potential of CNNs for identifying bird sounds from audio recordings.

The BirdCLEF 2021 dataset consists of bird vocalization recordings from around the world. The dataset is organized into training and testing sets, each containing over 40,000 recordings of various bird species. The dataset also includes metadata such as the location and date of the recordings.

The program I will be discussing is a Python program that uses the Keras library to build and train a CNN for image classification. The program can be found in the following GitHub repository: https://github.com/Sarthak6929/Bird-Voice-Detection.git. We will go through the main parts of the program and understand how my CNN operates.

## 4 DATA PROCESSING

Preparing the data is the initial stage in creating a CNN for bird voice detection. Once the audio files were converted to spectrogram. The photos' pixel values were rescaled to be between 0 and 1 using the Keras ImageDataGenerator class, which preprocessed the data. After that, the data was divided into training and validation sets.

The CNN used in this project consisted of several layers. The first layer was a Conv2D layer with 32 filters and a kernel size of 3x3. This layer used the ReLU activation function and had an input shape of 100x100x1. The second layer was also a Conv2D layer with 64 filters and a kernel size of 3x3. This layer was followed by a MaxPooling2D layer with a pool size of 2x2 and a Dropout layer with a rate of 0.25.

The next set of layers consisted of two Max pooling layers with a pool size of 2x2 and a Dropout layer with a rate of 0.25.

The CNN's final layers included a Flatten layer to transform the previous layers' output into a 1D vector, a Dense layer with 1024 neurons and a ReLU activation function, a Dropout layer with a rate of 0.5, and a Dense layer with 7 neurons and a softmax activation function.

The CNN was compiled using the categorical cross-entropy loss function and the Adam optimizer with a learning rate of 0.0001 and a decay of 1e-6. The model was trained for 85 epochs with a batch size of 64 and a steps per epoch of 6.

## 5 RELATED WORKS

ResNet-50, a deep convolutional neural network architecture for automated bird call recognition. The Res-Net 50 has a 62% accuracy in identifying bird species [12]. I will be using a similar approach and developing a CNN which would help identify the bird species. Although many other research papers discuss different techniques however, it has been determined that deep learning-based technique CNN with fully convolutional learning calls gets more accurate results because it eliminates the possible future modelling error caused by an imprecise knowledge of bird species [6]. I have never worked with CNN before and hence would like to start my research in this field and see how it follows. Similar technique,

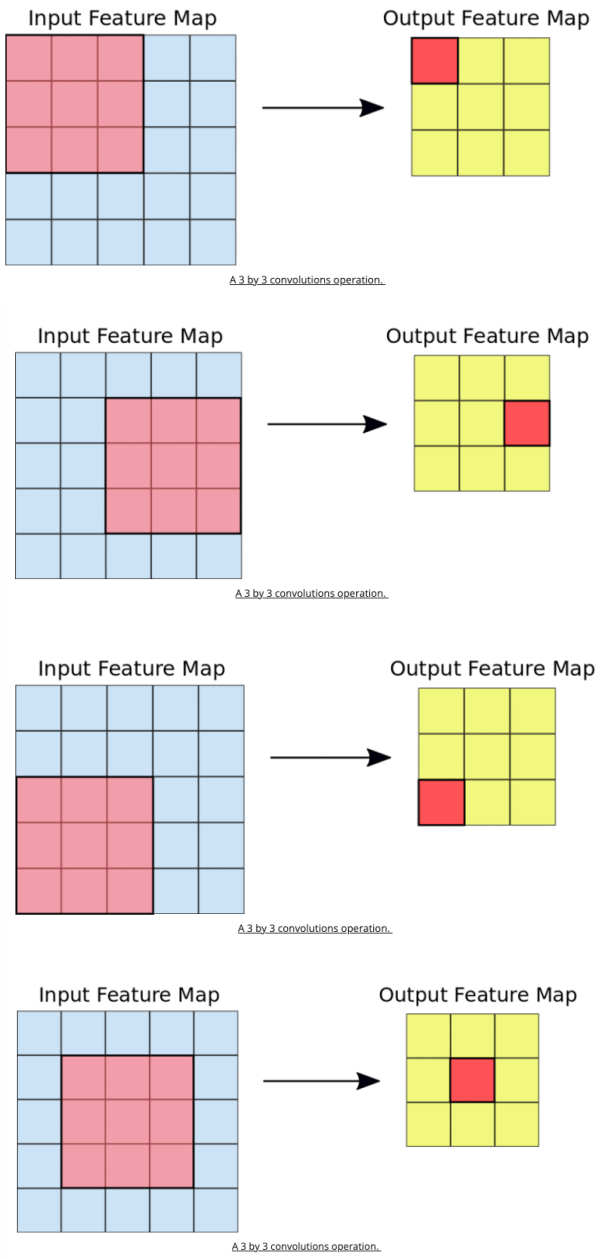


**Figure 2: A 3 by 3 convolutions operation.**
[10]

where passive acoustic monitors are used to record birds and animal sounds and then processed through bio acoustic analysis to get a spectrogram output of the recordings so as to identify bird sounds are mentioned in the research paper [7]. The paper describes how the spectrograms are derived from the recordings, and the process of manually scanning, each bird's audio to classify it in its particular species using the Xeno-Canto Foundation's online bio-acoustic library[7].

## 6 ARCHITECTURAL DIAGRAM

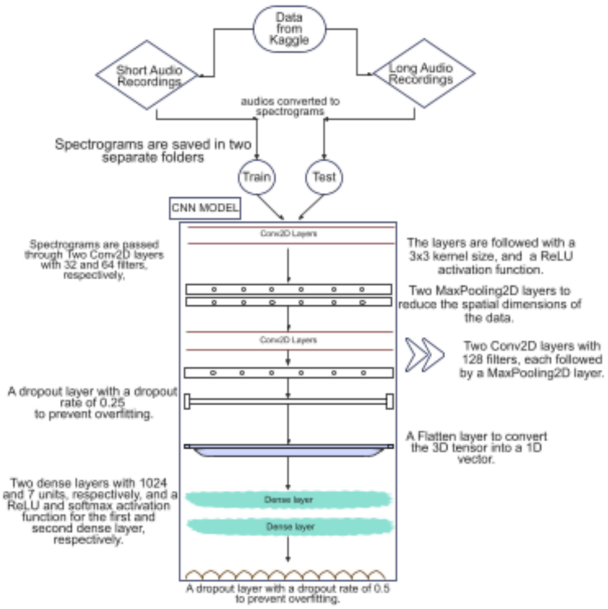


**Figure 3: An Architectural Diagram of my plan**

The input layer will be consist of both short audio recordings as well as the short sudio recordings. Once the CNN is trained on short 10 second videos we will switch the CNN to work on the 10 minute recordings. The Input Layer would consist of, Spectrograms of 10 minutes audio recordings.
Filtering will be done by Convolutional Layer: The first layer in the CNN architecture is a Conv2D layer with 32 filters and a 3x3 kernel size. The Conv2D layer is followed by a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity to the output of the layer. The second Conv2D layer has 64 filters and the same kernel size and is also followed by a ReLU activation function. The Kernel size is the size of the filter used in the convolutional layer. It is usually a square or a rectangle and is usually set to (3,3). All the values I have taken right now have been decided after studying [10] research paper. In the upcoming semester with further research these values might or might not change.
Padding method discussed above is the number of pixels added to the input image before the convolutions are applied. This is done to ensure that the output size of the convolutional layer is the same as the input size.
Pooling Layer: The next layer in the architecture is a MaxPooling2D layer, which reduces the spatial dimensions of the output from the Conv2D layer by a factor of 2 in both the height and width dimensions. Another MaxPooling2D layer follows the second Conv2D layer. The Pooling layer is a subsampling layer that is used to reduce the spatial resolution of the input image. This helps to reduce the

number of parameters in the model and increase its efficiency. It is done by taking the maximum value of a certain region of the image and replacing it with the maximum value of the region.
Two more Conv2D layers with 128 filters and the same kernel size are added to the model, each followed by a MaxPooling2D layer. This helps the model to extract more complex features from the input data. A Dropout layer with a dropout rate of 0.25 is added to the model to prevent overfitting, which occurs when a model becomes too complex and performs well on the training data but poorly on new, unseen data.
The Flatten layer is then used to convert the output of the previous layers into a 1D vector, which can be passed to the dense layers of the model. The first dense layer has 1024 units and is activated by a ReLU activation function, which helps to introduce non-linearity to the output of the layer. Another Dropout layer with a dropout rate of 0.5 is added to this layer to further prevent overfitting. The final dense layer has 7 units, which corresponds to the number of output classes, and is activated by a softmax activation function, which produces a probability distribution over the output classes.
The evaluation of the work will involve measuring the model's performance on a variety of metrics, such as accuracy, precision, recall, and F1-score. Accuracy is a measure of how often the model correctly identifies a bird species from audio recordings. Precision is a measure of how often the model identifies a bird species correctly, out of all the times it made a prediction. Recall is a measure of how often the model identifies a bird species out of all the times that species was actually present in the audio recordings. Finally, F1-score is a measure of the model's overall performance, which takes into account both precision and recall. The model will be evaluated based on its ability to generalize to unseen data, which is a measure of its robustness to changes in conditions such as noise and recording quality. Lastly, the model will also be evaluated based on its performance on other datasets, such as unseen images or sounds, which will give a better indication of its overall capabilities. This evaluation will help determine the model's effectiveness in recognizing different types of bird calls and accurately identifying bird species from audio recordings.

## 7 CONCLUSION

An interesting application of bird sound identification is its potential use in solving various issues. By correctly identifying the sounds made by different species of birds, we can learn more about their behaviour and ecology [8]. This information can then be used to help us manage ecosystems better, prevent the spread of diseases among birds, and even mitigate the effects of climate change on bird populations. My research will contribute towards answering these many important questions. For my research I will be focusing on getting recordings from [2] and then, identifying as many birds as the CNN would allow me to. Numerous grassland bird species, like Northern Bobwhite Quail, are drastically declining. Therefore, it would be advantageous for society as a whole to have a tool that could be used to analyse the recordings in getting an accurate data on different bird species and the potential habitat destruction. The data can be used by biologists and academics to make projections about a species' population, size, and geographic dispersion. All of this data is essential for managing or monitoring a species. Bird Sound Identification has many uses and will solve issues. My research would help answering these problems. Unlike the common belief, bird sounds are not just beautiful music for people to appreciate. They actually have a lot of functions that can be very useful for humans too! For example, by identifying the different types of birds through their unique calls, we can get an understanding about what kind of environment they prefer and whether it is healthy or not. This knowledge can then be used to address some major environmental concerns humanity faces.

## 8 INITIAL RESULT

The program returns an accuracy of between 75% and 80%, which is a promising start for the Python code. This shows that the model has a pretty high degree of accuracy in correctly classifying the input data. Although there may be room for improvement, these findings are encouraging and indicate that the model was successfully trained using the provided data. These first results offer a solid basis for continued development and modification, although additional testing and analysis may be required to confirm the model's dependability and robustness.

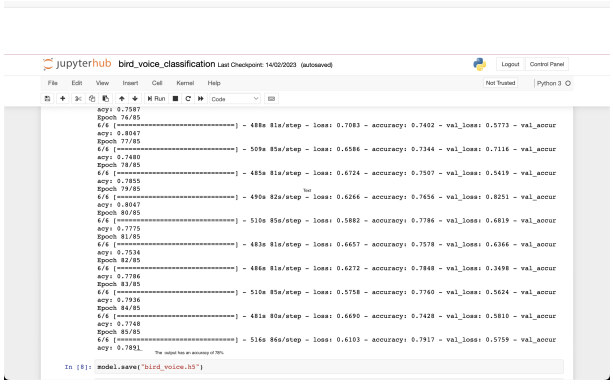


**Figure 4: Initial result**

## REFERENCES


[1] 2022. ARTICLE Why Monitor Birds? *National Park Service* (2022), 4 pages. https://www.nps.gov/articles/000/why-monitor-birds.htm

[2] 2022. BirdCLEF 2021 - Birdcall Identification. (Dec. 2022), 1–4. https://www.kaggle.com/competitions/birdclef-2021/overview

[3] Northern Rocky Mountain Science Center. 2016. The ecology, behavior, and conservation of migratory birds. (2016), 1–4. https://www.usgs.gov/centers/norock/science/ecology-behavior-and-conservation-migratory-birds

[4] The Nature Conservancy. 2022. This Year's Big Wins in Land Conservation. (Dec 2022), e0211970. https://www.nature.org/en-us/

[5] Environment Defence Fund. 2022. We are Environmental Defense Fund, the organization that is all-in on climate — the greatest challenge of our time. Our game-changing solutions put people at the center of all we do. (Dec 2022), e0211970. https://www.edf.org/

[6] Hasan Abdullah Jasim, Saadaldeen R. Ahmed, Abdullahi Abdu Ibrahim, and Adil Deniz Duru. 2022. Classify Bird Species Audio by Augment Convolutional Neural Network. (2022), 1–6. https://doi.org/10.1109/HORA55278.2022.9799968

[7] Kinga Kułaga and Michał Budka. 2019. Bird species detection by an observer and an autonomous sound recorder in two different environments: Forest and farmland. *PLoS One* 14, 2 (2019), e0211970.

[8] Shelby Lawson and Mark E. Hauber. 2022. What Can We Learn from Bird Song? Recent Advances in Functional and Applied Avian Bioacoustic Research. *American Ornithological Society Journals* (2022), 8 pages. https://doi.org/DepartmentofAnimalBiology,SchoolofIntegrativeBiology, UniversityofIllinois,Urbana-Champaign,ILUSA

[9] Debbie Mondale. 2009. Exploring bird sounds with children. (August 2009), e0211970. https://musiconnx.wordpress.com/2009/08/30/exploring-bird-sounds-with-children/#:~:text=Fun%20vocal%20play%20with%20different,enough%20to%20imitate%20it%20effectively.

[10] Derrick Mwiti. 2022. Image Classification with Convolutional Neural Networks (CNNs). (May 2022), 4. https://www.kdnuggets.com/2022/05/image-classification-convolutional-neural-networks-cnns.html

[11] Mike Proctor and Stephen Webb. 2020. Capturing bird calls and other wildlife sounds with bioacoustics. *Capturing Bird Calls and Other Wildlife Sounds With Bioacoustics* 38, 12 (Dec. 2020), 1–4.

[12] Mangalam Sankupellay and Dmitry Konovalov. 2018. Bird call recognition using deep convolutional neural network, ResNet-50. 7, 9 (2018), 1–8.

[13] Jon Schwartz. 2020. Building a Convolutional Neural Network to Classify Birds. *Wildlife Acoustic* 5, 2 (2020). https://blog.jovian.ai/building-a-convolutional-neural-network-to-classify-birds-528794240fa1

[14] Mike Smales. 2019. Sound Classification using Deep Learning. (Feb. 2019), 1–4. https://mikesmales.medium.com/sound-classification-using-deep-learning-8bc2aa1990b7

[15] Stefano Ermon Volodymyr Kuleshov, S. Zayd Enam. 2017. Audio Super Resolution using Neural Networks. (Aug 2017). https://arxiv.org/abs/1708.00853