CS 522 - Birdsong Analysis and Species Identification with Machine Learning

Bryan Burton
Old Dominion University
Norfolk, VA 23529
bburt007@odu.edu

Abstract

In ecology, identification of species is necessary in order to determine the composition of habitats, biodiversity, and the health of ecosystems. However, identification can be a difficult process due to the need for frequent field observations and the existence of visually similar species. In this respect, machine learning offers many benefits for the field of ecology through classification. For the identification of birds, audio analysis can be used to simplify the process by using their unique calls as identifiable features. With the use of citizen science recordings, this experiment shows how analyzing spectrograms of audio data can be used to identify birds from field recordings.

1 Introduction

Species composition is an important aspect of ecology used to measure the health of an ecosystem and how it changes over time. The composition of an ecosystem is determined through routine field observations and cataloging the types of organisms present.

Determining the composition of bird species in a habitat can be done through visual observations, but another option is to use audio recordings and distinguishing a species by its unique song or call. However, the process of identifying a bird by listening to its sounds can be difficult and time-consuming. Despite this, audio identification has some benefits over visual identification in that it can be easier to obtain audio data than visual data, and it has the potential to be automated through audio recordings. Audio records can be obtained from a habitat, which can then be processed and classified without the need for a field observation.

With the use of audio recordings of bird songs and machine learning algorithms, it would be possible to classify and identify a species of bird through its vocalizations. The model could be used to track migration habits, or act as an early-warning system to detect non-native birds entering a habitat. Classification of citizen science recordings of bird calls can also be aided with this model to improve the data available for birds.

Through the use of audio feature extraction from spectrogram analysis, it is shown that this approach can be used for bird identification through field recordings.

2 Dataset

The dataset used for this experiment was the Xeno-Canto wildlife sounds repository hosted by the Xeno-canto Foundation. This dataset was chosen due to its large quantity of bird audio data, as well as the ability to filter data based on its quality. Audio data from Xeno-Canto can also be classified as belonging to multiple species, which makes it more realistic to recording obtained from a field observation.

The scope of the dataset was limited to 100 birds most commonly found in the eastern United States. For training the model, 25 birds were selected from this set, containing 21 Passerine birds, 2 Piciform birds, 1 Strigiform bird, and 1 Gruiform bird. 100 audio samples were selected at random from the Xeno-Canto dataset for each of the 25 birds, for a total of 2500 samples. As a result, there was no class imbalance in this dataset at the species level. However, there is a class imbalance at the family level, as most samples in this dataset belong to Passerine birds.

Data from Xeno-Canto was limited to recordings made in the United States that were at most 90 seconds long, had a quality score of "A" or "B", and had a sample rate of at least 22.05kHz.

The iNaturalist Sounds Dataset compiled by Chasmai et al. [2024] was also used in this project, but data from Xeno-Canto was used primarily, as samples from iNaturalist Sounds were often noisy and the dataset was larger, having data for species outside the scope of the project. This dataset was used for data exploration, producing visualizations of audio file spectrograms.

3 Data Processing

Audio files in the dataset were first loaded using librosa and down sampled to 22.05kHz if the sample rate of the file was too large. After down sampling, a mel-scaled spectrogram was obtained from each file, which was used as a feature for model training. Mel-frequency cepstral coefficients (MFCCs) were derived from the mel spectrogram and included in the feature set. Both of these metrics measure the intensity of frequencies at certain points throughout the audio file.

128 samples from the mel spectrogram and 20 MFCCs were obtained from the audio file and used as features. Each sample for the mel spectrogram and MFCCs produced a variable-length array of values that were averaged into a single value . The length of this array was dependent on the duration of the audio file, and averaging the arrays made it possible to work with audio files of varying durations.

The processed features from each audio file were saved to a CSV where each row contained the bird's common name, binomial name, name of the file, and the averaged mel spectrogram and MFCCs.

A similar data processing pipeline was implemented by Zenkov [2020]. Likewise, the averaged mel spectrogram and MFCCs were scaled using scikit-learn's `StandardScaler`.

4 Implementation

The experiment was implemented in Python using Jupyter using the libraries `sklearn`, `librosa`, and `pytorch`.

In the experiment, the dataset was used to train the following models from the `sklearn` library: `KNeighborsClassifier`, `LogisticRegression`, `SVC`, `SGDClassifier`, `DecisionTreeClassifier`, `RandomForestClassifier`, `HistGradientBoostingClassifier`, and `MLPClassifier`. Each model was trained on scaled data using five-fold cross validation.

Hyperparameter tuning was done using grid search to determine the best settings for each model.

A 1D convolutional neural network was constructed using `pytorch` according to the model designed by Kiranyaz et al. [2021] with 3 1D CNN layers and 2 MLP layers. This model was chosen to make use of the one-dimensional structure of the dataset with a convolutional neural network.

5 Results

After performing the experiment, it was found that `sklearn`'s `HistGradientBoostingClassifier` performed the best out of the selection of models, as shown in Figure 1.

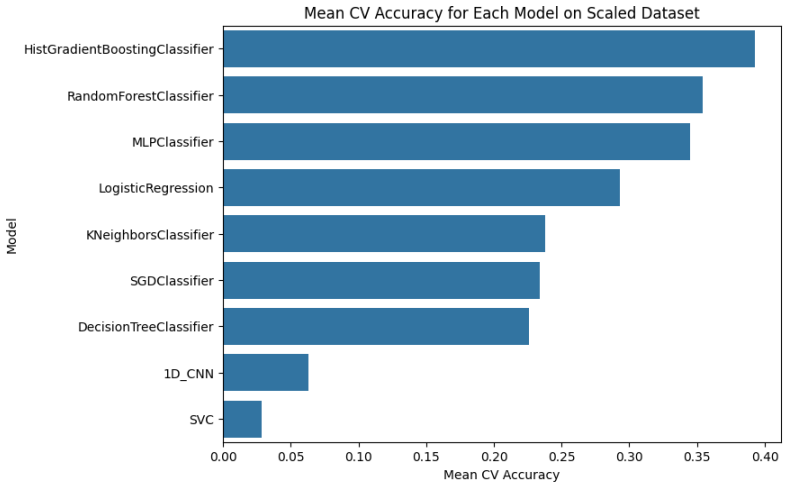

Figure 1: Mean Cross Validation Accuracy for Each Model

While training each model, it was found that the models tended to overfit, which led to accuracy scores that were lower on the test set than on the training set. To resolve this, regularization was done with some improvement to the fitting of the data, although the highest mean accuracy seen was around 40% for the best-scoring model.

For this dataset, the `HistGradientBoostingClassifier` could more accurately predict samples from *Bubo virginianus* and *Antigone canadensis* than the other species, as shown in Figure 2.

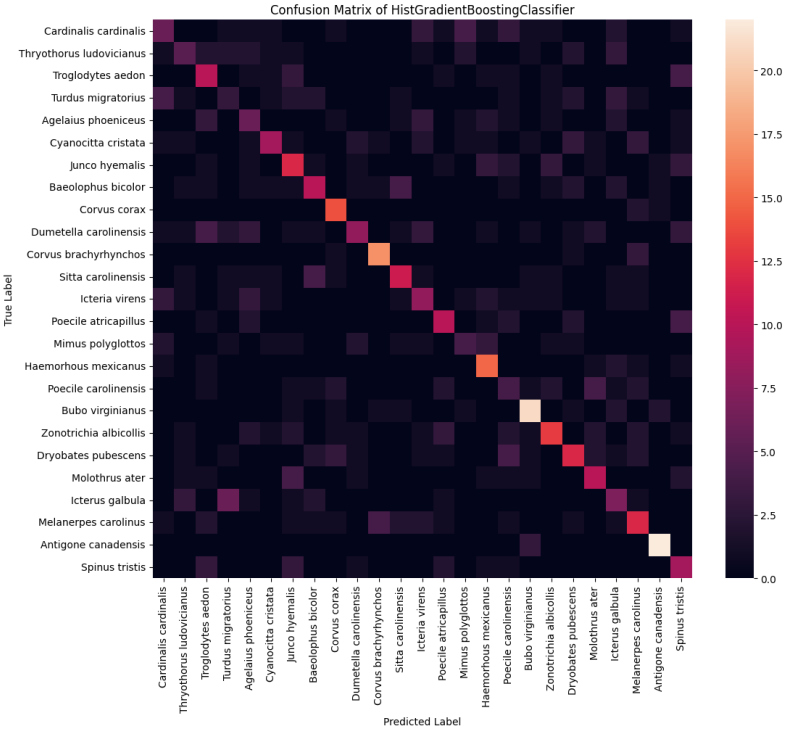

Figure 2: Heatmap of `HistGradientBoostingClassifier` Predictions

6 Discussion and Further Experiments

The results of this experiment show that frequency features in audio files can be used to identify bird species from their songs and calls.

In the results, it was shown that two species, *Bubo virginianus* and *Antigone canadensis*, were more accurately predicted than the other birds in the set. This is likely because these two birds are from different families than most of the birds, who are Passerines. It is likely that the Passerine birds, being from the same family, have similar morphologies and similar-sounding bird calls or songs. The differences of *Bubo virginianus* and *Antigone canadensis* from the rest of the set could be why they were more often identified correctly.

A way to improve the performance of the model may be to limit the set of birds to species in a specific location. In one approach, multiple small models could be trained, and unlabeled data where the location is known can be used for inference on a specific model. Other features such as the location of the recording, time of year, and length of the bird's call may be valuable for training these models. By doing this, the models could be more specialized and produce more accurate predictions on a smaller set of classes.

Another way to improve the performance of the model would be to change the methods for audio processing. In this experiment, much of the data was averaged, allowing audio files of different lengths to be used for training while losing information that could improve accuracy. A different approach would be to clip the audio file to the start and end of the bird's call, and use zero-padding to keep the data uniform in size. Noise reduction could be used as well to isolate the call of the target bird. As well, limiting the frequency range of the audio samples to the range that birds can hear, around 1-4 kHz on average as documented by Beason [2004] may help reduce overall background noise.

References

Robert Beason. What can birds hear? *U.S. Department of Agriculture National Wildlife Research Center-Staff Publications*, 78, 09 2004.

Mustafa Chasmai, Alex Shepard, Subhransu Maji, and Grant Van Horn. The inaturalist sounds dataset. *Advances in Neural Information Processing Systems*, 2024.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021. ISSN 0888-3270. doi: https://doi.org/10.1016/j.ymssp.2020.107398. URL `https://www.sciencedirect.com/science/article/pii/S0888327020307846`.

Xeno-canto Foundation. URL `https://xeno-canto.org/`.

Ilia Zenkov. sklearn-audio-classification. `https://github.com/IliaZenkov/sklearn-audio-classification`, 2020.