Deep Learning Approaches to Classify Seattle Bird Calls

Abstract

This study investigates how well convolutional neural network (CNN) models are able to classify bird species based on audio clips of their calls. The study will focus on twelve bird species commonly found in Seattle, using data from Xeno-Canto which is a crowd sourced bird sound archive. Two classification tasks will be explored, including a binary model to distinguish between an American Crow and White-crowned Sparrow, as well as a multi-class model that can identify a bird call between all twelve species. Both models perform significantly better than random guessing, showing that CNN models pick up general patterns of different bird calls and can predict on new data. These results provide evidence for feasibility of using deep learning models to classify species from real world data to improve current issues like climate change and promote biodiversity.

Introduction

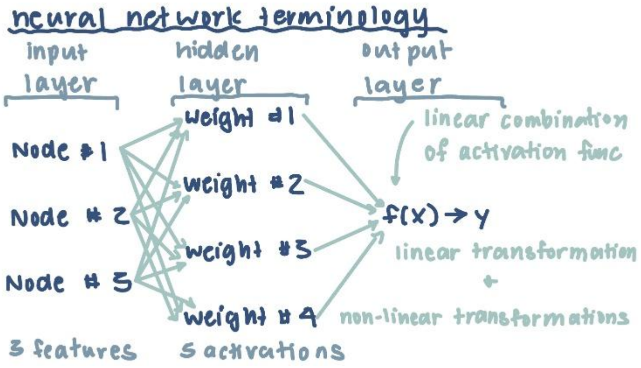
Xeno-Canto is a crowd sourced bird sounds archive where researchers and scientists worldwide contribute their resources to compile data on various bird species (2). This study will pull bird call audio files from this dataset, focusing specifically on twelve bird species commonly observed in Seattle, including the American Crow, American Robin, Bewick's Wren, Black-capped Chickadee, Dark-eyed Junco, House Finch, House Sparrow, Northern Flicker, Red-winged Blackbird, Song Sparrow, Spotted Towhee, and White-crowned Sparrow (1). The original audio recordings of these birds are processed into spectrograms, which are visual representations of sound frequency over time. Spectrograms capture the distinctive frequency patterns of each species calls and are used as the input data to train and evaluate the following models.

This report explores two primary classification tasks. The first model is a binary classification that distinguishes between two acoustically similar species: the American Crow and White-crowned Sparrow. The second model is a multi-class classification problem, where the model will distinguish between the twelve bird species that are common to the Seattle area. Deep learning models, mainly convolutional neural networks, will be applied to spectrogram data to investigate how well these models classify real world data. CNN models are particularly useful at classifying spectrogram data because these models are designed to recognize visual patterns, for example how a spectrogram might change in frequency, harmonics, or volume. The findings of this study will help inform future implementations of using deep learning to approach key challenges in today's world like how climate change affects endangered species and how we might promote biodiversity.

Theoretical Background

Neural networks are machine learning models that follow a decision-making process that weighs various input options and arrives at a conclusion. Neural network models are particularly effective at modeling complex, non-linear relationships in high-dimensional data like images or audio files. A neural network model contains an input layer, one or more hidden layers, and an output layer. Each layer contains nodes that compute weighted sums of inputs, applies an activation function, and then passes the results to the next layer.

Figure 1. Neural Network Diagram



To implement a neural network model, the structure is set up to identify the number of hidden layers. Activation functions, like ReLU, are also defined here to introduce non-linearity into the model and determine whether a neuron's output should be passed to the next layer. Next, the model is trained with an optimizer parameter that specifies how a model's weights are updated. RMSprop is an example of an optimizer parameter that adjusts the learning rate for each weight individually using a moving average of the squared gradients. The optimizer aims to minimize loss functions, such as binary cross-entropy for multi-class tasks or categorical cross-entropy for multi-class tasks, which is also a training parameter that needs to be specified. Finally, the epoch parameter specifies the number of full passes through the training data and the batch size identifies the number of training samples used in each step of the gradient computation.

Convolutional Neural Networks are a specific type of deep learning neural network that learns features by sliding a filter across data to detect patterns and create a summary. CNN models use filters that convolve over the input to detect localized features. Pooling layers reduce dimensionality and computational power, an example is the max pooling method, which takes the maximum value within a given block and condenses the data. CNN models are particularly effective at analyzing spectrograms because they are designed to recognize visual patterns. Since spectrograms are 2D visual representations of audio signals, CNN models are able to learn spectral features like volume, harmonics, and vocal changes to classify various audio clips.

Neural networks have many limitations and are not always the optimal model to implement. Neural networks are slow and computationally expansive. These models need lots of training data, and will often overfit to existing patterns. Dropout is a parameter that can be added when defining the structure of a neural network; this is a regularization technique that randomly deactivates a specified percentage of neurons during training to prevent overfitting. Data augmentation improves the size and diversity of training data by applying transformations to add noise to the original data, which helps models better generalize to a wider variety of data and prevent overfitting. Neural Networks are also not easily interpretable, like a linear regression, so there is a loss of model application in many real-world problems and applications. It is always a good path to try simple and more easily explained models first, however there are some situations where neural networks might be the only model choice. Neural Networks are flexible and often the only choice to analyze sound and image data; in specific cases these models are strong and robust options.

Methodology

The original Kaggle dataset on birdcall data contains 264 species. The first step of the data preprocessing is narrowing down the dataset to the following twelve bird species that can be found in Seattle: the American Crow, American Robin, Bewick's Wren, Black-capped Chickadee, Dark-eyed Junco, House Finch, House Sparrow, Northern Flicker, Red-winged Blackbird, Song Sparrow, Spotted Towhee, and White-crowned Sparrow. Sound clips were selected from each species and subsampled to 22050HZ. The first three seconds of the sound clip are selected and a spectrogram, or an image of the bird call, is produced for each 2-second window. This results in a 128 frequency x 517 time image of the bird call, and 38-630 samples are produced for each of the selected bird species. Since there is a significant class imbalance, where some bird species have 30 samples while others have 630 samples, these class imbalances will be addressed appropriately for each model in this study.

Binary CNN Classification: American Crow and White-crowned Sparrow

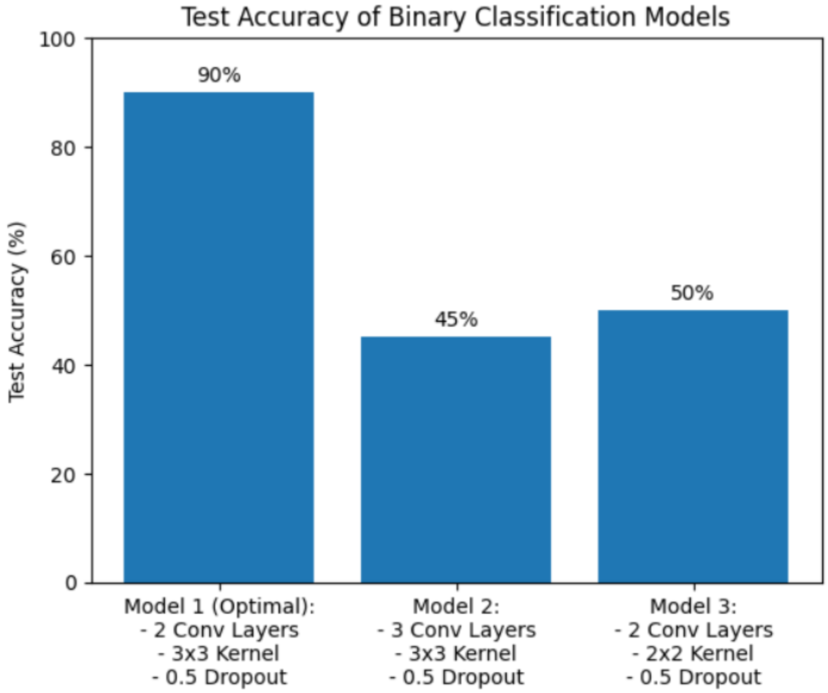
This binary classification will use a convolutional neural network to identify the call of an American Crow or White-crowned Sparrow. These two species were chosen due to their relatively similar class sizes in the overall data set, as well as challenging a CNN to learn patterns among similar bird calls. Each spectrogram captures the unique acoustics of either the crow or sparrow, which is combined to create the input (X data frame) or predictor variables for this model. The target variable (y data frame) is a binary label indicating the bird species, where 1 represents the American Crow and 0 is the White-crowned Sparow.

The dataset for this binary classification includes 66 American Crow samples and 91 White-crowned Sparrow samples, yielding a moderately balanced dataset. Hence, all spectrogram data for these two species will be used as part of the training, test, and validation splits. These splits were chosen to ensure enough data is included to train the model while preserving enough samples for unbiased evaluation and parameter tuning.

- *Training set*: 54% (~84 samples)
- *Test set:* 33% (~52 samples)
- *Validation set*: 13% (~21 samples)

To efficiently identify the optimal convolutional neural network architecture while minimizing computational power, several CNN models were first evaluated on a subset of the data. Specifically, 30 samples were randomly selected from each class using the same training, test, and validation splits as discussed. This exploratory method allowed for testing multiple architectures and tuning key parameters before training on the full dataset. The main parameters explored includes the number of convolutional layers, kernel (filter) sizes, and dropout rates, with the goal of balancing model complexity and reducing overfitting. The following architectures were tested (see test_models.ipynb for full implementation details):

- Model 1 (Optimal): 2 Cov Layers, 3x3 Kernel Size, 0.5 Dropout
- Model 2: 3 Cov Layers, 3x3 Kernel Size, 0.5 Dropout
- Model 3: 2 Cov Layers, 2x2 Kernel Size, 0.5 Dropout



Model 1 yielded the best test accuracy of about 90%, balancing model depth and generalization to unseen data. Since this model architecture significantly out-performed the other tested models by 45% more accuracy, the final CNN model trained on the full dataset will be based off of these parameters and model architecture.

The CNN model that is trained on the full dataset follows a similar convolutional neural network architecture to test model 1. The two convolutional layers includes 32 filters in the first layer and 64 filters in the second layer, both utilizing a ReLU activation function with 3x3 kernels to capture acoustic features. After each convolutional layer, max pooling layers helped reduce dimensionality and computational power. Finally, a flatten layer converts 2D features into a 1D vector, and a dense layer with 64 units helps learn complex features. A 50% dropout regularization also helps prevent the model from overfitting. The model is compiled using binary cross entropy as the loss function to compare predicted and actual values, while using accuracy to evaluate the binary classification. Finally, the model is trained on 50 iterations of the full training data using an early stopping method that tracks the validation accuracy, with a batch size of 16.

Multi-class CNN Classification: All 12 Common Bird Species in Seattle

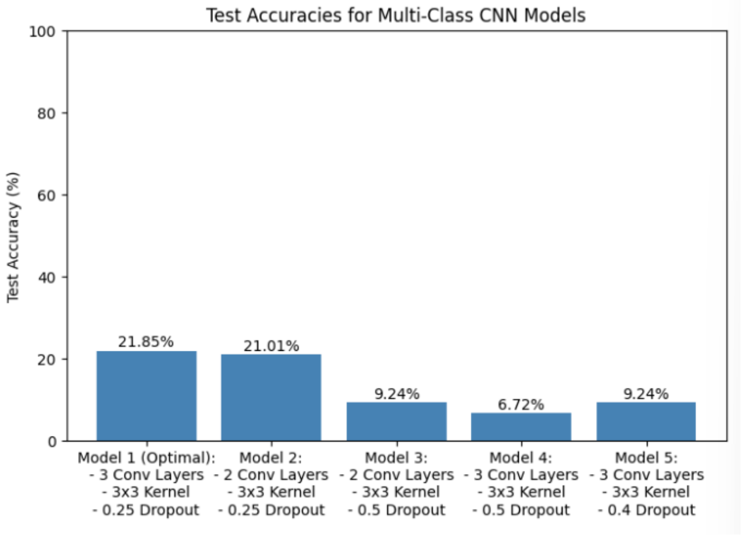
This multi-class classification will utilize a convolutional neural network to identify any one of the twelve bird species from the dataset that can be found in Seattle. Similarly to the binary classification model, the input data will include spectrograms that each represent a unique bird call from one of the twelve species. The target variable consists of numerical indices from 0-11, where each index corresponds to one of the twelve bird species.

The dataset used for this multi-class model is more complex than the binary model, as class sizes vary significantly across the twelve bird species. To ensure the model can generalize the data well enough across all classes, the class imbalance was addressed by simply reducing the significantly larger outlier class size. The house sparrow was the only bird with an extremely large class size of 630, which is randomly reduced to 200 samples. The rest of the class sizes are relatively more balanced, ranging from 37 to 200 samples. The same training, test, and validation splits were used as the binary classification model:

- *Training set*: 54% (~831 samples)
- *Test set:* 33% (~512 samples)
- *Validation set*: 13% (~208 samples)

A similar method was used to find the optimal convolutional neural network architecture for the multi class model, by sub-setting the data and tuning parameters on smaller scale models first. For the multi-class model, 30 samples per class were randomly selected and all models used the same training, test, and validation splits as discussed prior. The main parameters explored also included the number of convolutional layers, kernel (filter) sizes, and dropout rates. The following architectures were tested (see test_models.ipynb for full implementation details):

- Model 1 (Optimal): 3 Cov Layers, 3x3 Kernel Size, 0.25 Dropout
- Model 2: 2 Cov Layers, 3x3 Kernel Size, 0.25 Dropout
- Model 3: 2 Cov Layers, 3x3 Kernel Size, 0.5 Dropout
- Model 4: 3 Cov Layers, 2x2 Kernel Size, 0.5 Dropout
- Model 5: 3 Cov Layers, 2x2 Kernel Size, 0.4 Dropout



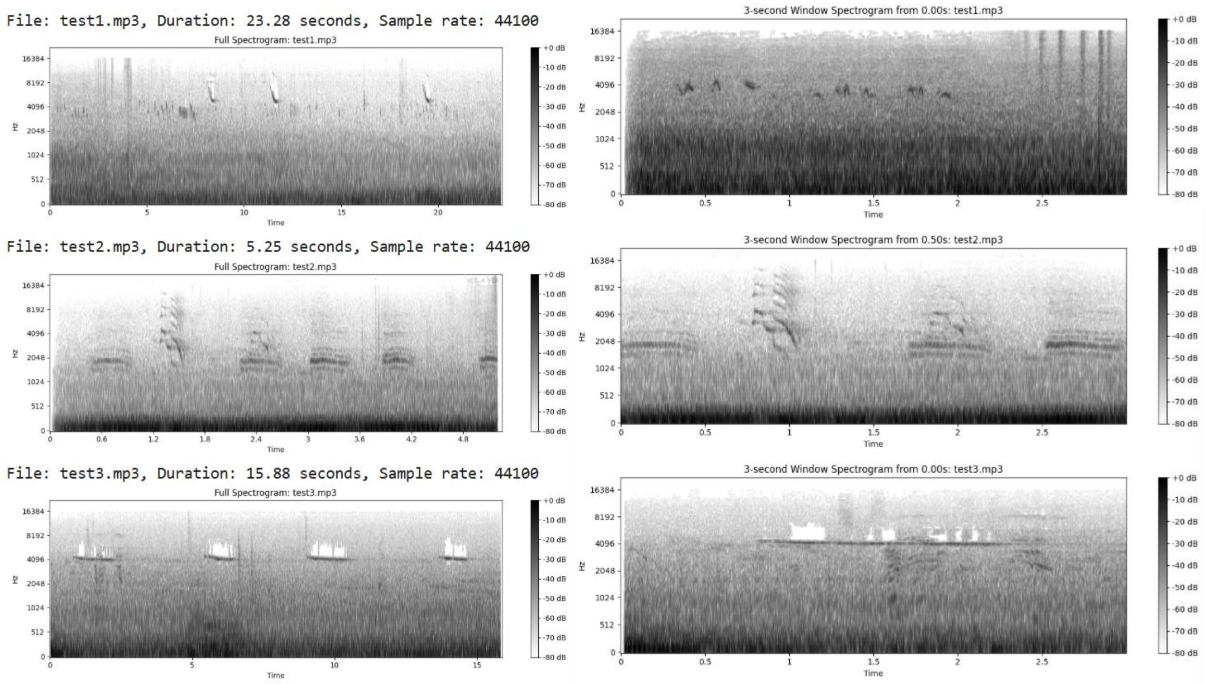
Model 1 yielded the best accuracy of about 31.93%. Although the accuracy is not substantially high, the model predictions prove much better than guessing 1 in 12 bird species (~9%), which shows that this CNN architecture is able to capture some variances among the bird calls. This is also a significantly smaller dataset, so training the model on the full data set should show additional improvement.

The final model follows a similar CNN architecture and tuning parameters as the optimal model found during testing on a subset of the data. The final model includes two convolutional layers with 32 and 64 filters using ReLu activation and a 3x3 filter, max pooling to reduce dimensionality, flatten to pass a one-dimensional vector through 64 dense layers, and a dropout layer that randomly drops 50% of neurons for regularization. The model is compiled using categorical cross entropy as the loss function to compare predicted and actual values, while using accuracy to evaluate the multi-class classification. Finally, the model is trained on 50 iterations using an early stopping method tracking the validation accuracy, with a batch size of 16.

Predicting on External Test Data

The final multi-class classification CNN model will be used to predict three bird species using their sound from an mp3 file. To preprocess the data, similar steps will be used to convert the mp3 file into a spectrogram. Firstly, each mp3 file will be subsampled to 22050HZ. A spectrogram of the entire file will be printed to examine the optimal 3 second window where bird calls are audible. The following 3-second windows were selected manually by viewing the first visual signs of an apparent bird call as visual frequency waves or changes in sound:

- *Test1.mp3:* 0.0 – 3.0 seconds
- *Test2.mp3*: 0.5 – 3.5 seconds
- *Test3.mp3*: 0.0 – 3.0 seconds



```
File: test1.mp3, Duration: 23.28 seconds, Sample rate: 44100
File: test2.mp3, Duration: 5.25 seconds, Sample rate: 44100
File: test3.mp3, Duration: 15.88 seconds, Sample rate: 44100
```

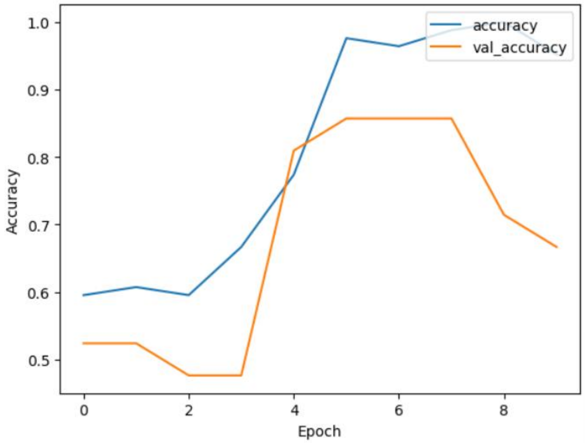
Next, an image of the bird class is produced for the optimal 3-second window which results in a 128 frequency by 517 time image of the bird call. All spetrogram data for the three bird species are saved into one file. The multi-class convolutional neural network model is used to predict on this test data.

Results

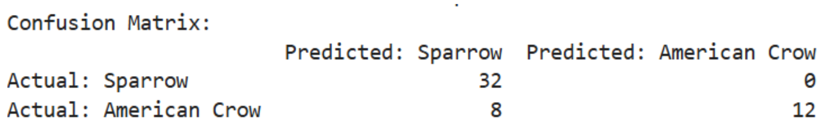
Binary CNN Classification: American Crow and White-crowned Sparrow

The final binary CNN model classifies between an American Crow and White-crowned Sparrow pretty well, clearly learning distinct patterns and features that identify these two bird species. This model returns a relatively high test accuracy, which means that this model correctly classifies 84.62% of the spectrograms it has never seen before. A strong accuracy implies that this model has learned general patterns in the training data and can reliably apply these patterns to data the model has never seen before.

Figure 2. Binary CNN Accuracy Plot



The following plot shows the training and validation accuracy over the number of iterations of the training data or epochs. The blue line represents the change in training accuracy and the orange line represents the change in validation accuracy. Initially, the model does a good job learning patterns in the data and applying them to the test set, as both accuracies rise steadily together. Around the middle of the plot, the training accuracy continues to rise while the validation accuracy dips and diverges from the training accuracy. This divergence shows that the model begins to overfit to the training data for a few epochs, so the early stopping allows the model to stop training before the overfitting gets worse.



```
Confusion Matrix:
                        Predicted: Sparrow  Predicted: American Crow
Actual: Sparrow                         32                         0
Actual: American Crow                    8                        12
```

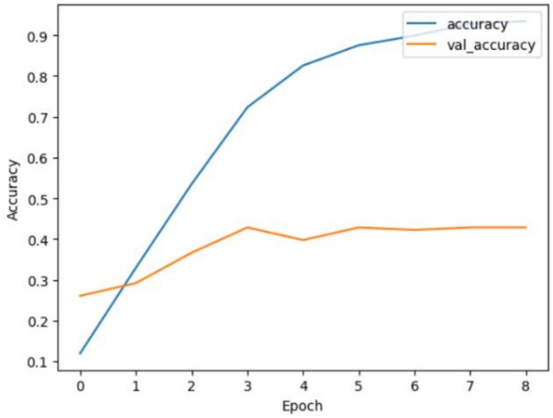
The confusion matrix shows that the model does a great job at identifying sparrow calls, with zero false negative predictions. The model struggles a bit more identifying American crow calls with the model correctly identifying about 60% (recall) of actual crows, but this may be expected due to sparrows having a little bit more training data than crows in the overall dataset. This model has an F1 score of 75%, meaning that overall it does a decent job classifying between the two bird species but evidently could be improved balancing recall and precision for the crow class.

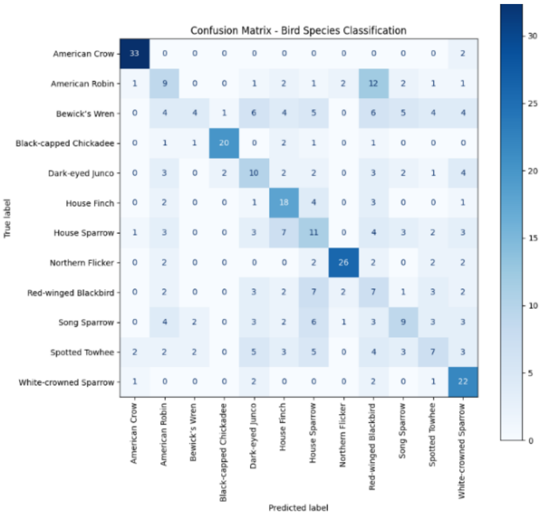
Multi-class CNN Classification: All 12 Common Bird Species in Seattle

The final multi-class classification CNN model identifies some distinct patterns among the twelve bird species that are common to Seattle, but clearly struggles to classify significantly more birds than the binary classification model. This model returns a decent test accuracy, where model is able to correctly classify 44.44% of the spectrograms it has never seen before. This model is significantly better than guessing, where guessing one out twelve species correctly would be about a 9% accuracy, so the model has definitely learned some patterns among the training data.

The following plot shows the training and validation accuracy over the number of iterations of the training data or epochs. The blue line represents the change in training accuracy and the orange line represents the change in validation accuracy. Initially, the model does an okay job learning patterns in the data and applying them to the test set, as both accuracies rise steadily together. However, at around 3 iterations the model starts to overfit to the training data learning nuances of different species but failing to generalize these patterns to new data the model has never seen before.

Figure 3. Multi-Class CNN Accuracy Plot

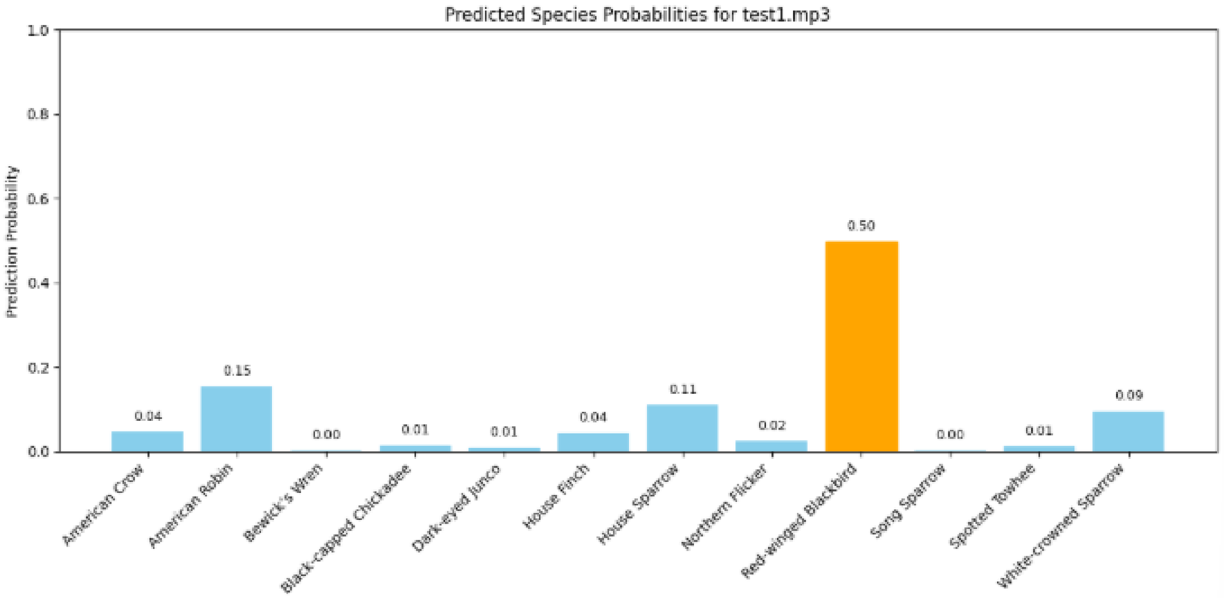




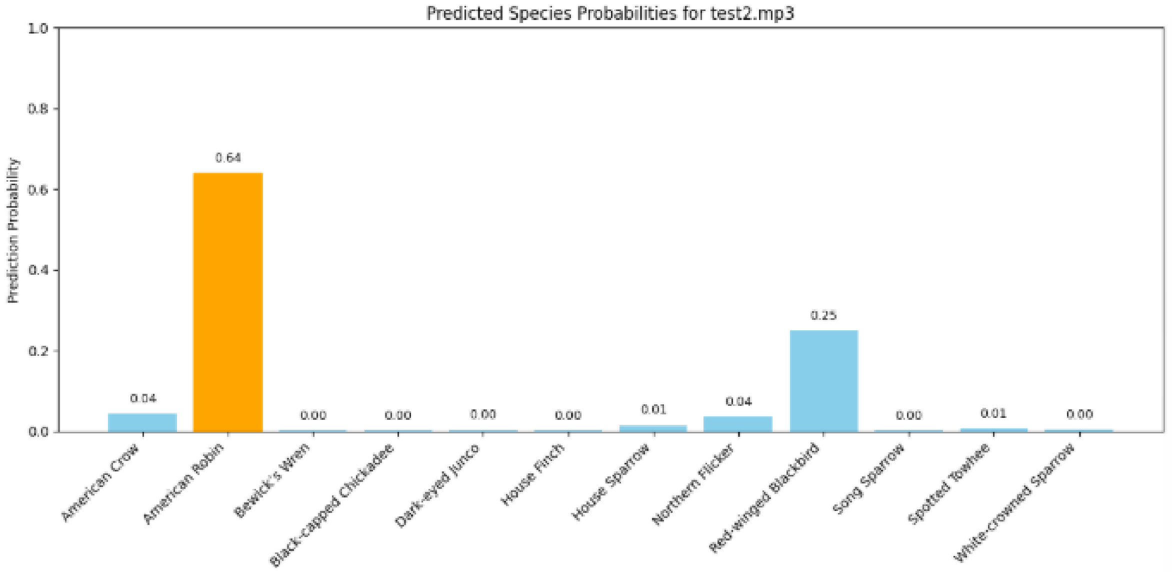
The confusion matrix shows the model generally does a good job at predicting the correct species, as the darker colored diagonal represents accurate predictions. Some species have particularly strong F1 scores above 60%, like the American Crow, Black-capped Chickadee, Northern Flicker, White-crowned Sparrow, and House finch which shows that the model recognizes clear patterns that identify these species bird calls. However, other bird species have F1 scored less than 35% and the model fails to capture variances within these species. Most of the species with lower F1 scores are the under sampled species, which may imply that under sampling these species present too few training samples for the model to capture patterns in these bird calls.

Predicting on External Test Data

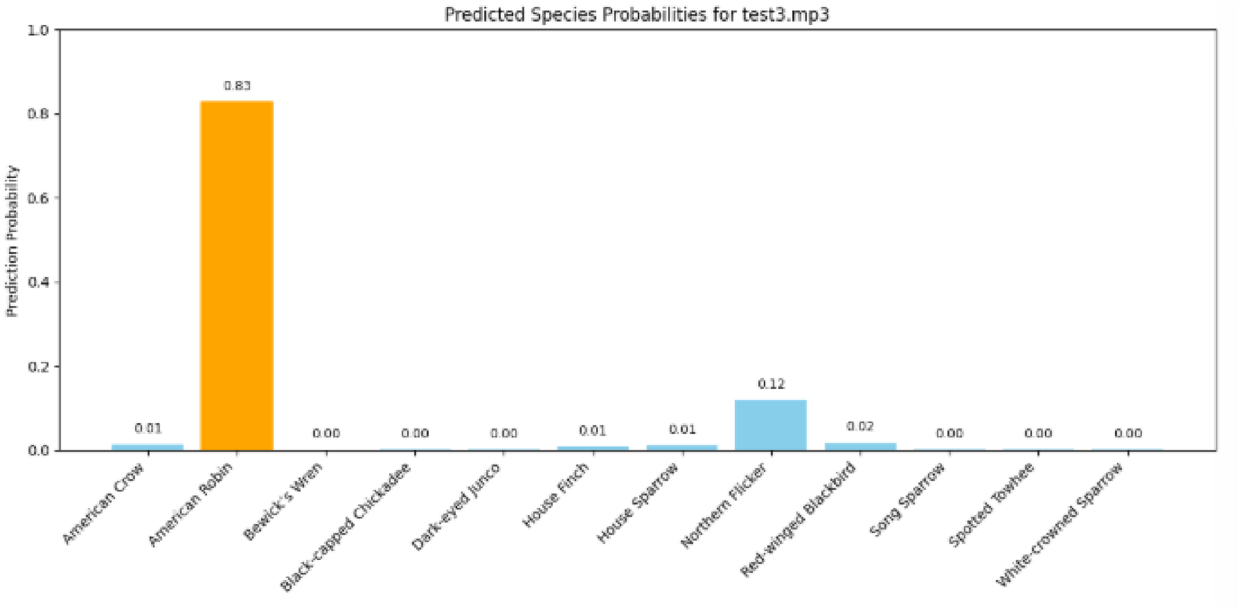
The bird species predictions are Red-winged Blackbird , American Robin, and American Robin for the respective files 1,2 and 3. Graphing the prediction probabilities for each of the twelve bird species will help analyze how confidently the model predicts the specific bird call from each file as well as if there may have been multiple bird calls within the sound clip. Each graph represents the prediction probabilities among all twelve species for each sound file. The blue bars represent the prediction probabilities, and the orange bar represents the highest prediction probability or the species that the model predicts to classify the bird call. The first graph represents the prediction probabilities for the first file, which classifies the bird call as a Red-winged Blackbird. The prediction probability for a blackbird is 49.77%, which is a relatively confident prediction but other probabilities might imply other birds like the American Robin and some sparrows might be present in the background of the clip.



The following graph represents prediction probabilities for the second file, which classifies the bird call as the American Robin. The multi-class CNN model classifies the robin with 63.92% confidence. It is highly likely the American Robin's call is in this sound clip, with maybe a Red-winged Blackbird in the background as the model predicts their call with 24.97% confidence.



Finally, the last graph represents prediction probabilities for the third file, which classifies the bird call as the American Robin. The multi-class CNN model classifies the robin with 82.82% confidence, which is a relatively high prediction probability. This likely means the American Robin in the main bird call that appears in this audio clip.



Discussion

In this report, two types of convolutional neural network models were implemented to classify between two birds and twelve bird species that are common to the Seattle area. The binary model, one that predicts only between two bird species, performs significantly better than the multi-class model, which can classify between all twelve bird species. The high accuracy of the binary model suggests that simpler classification tasks, one that identifies fewer classes, can learn clearer differences between bird calls and allows the convolutional neural network to learn and generalize patterns more effectively. The multi-class model still performs substantially better than a random guess, which demonstrates that the model was still successful at identifying prominent patterns and characteristics of different bird calls.

Both models present with signs of overfitting, which is indicated by the divergence of training and validation accuracy after an increasing number of epochs. Both models utilized an early stopping method when training the neural networks, which allows the model to reduce the chances of severe overfitting, however the issue is still present. Overfitting is likely caused by the wide range of class sizes, where even implementing different types of class balances from choosing similar class sizes to under sampling large classes/ resampling smaller classes, were not completely effective solutions to overfitting. Future implementations might address this class imbalance by data augmentation to increase the number of training samples or tuning the models further to address model complexity to address the overfitting issue.

Implementing the model on real-world audio clips of bird calls has promising results, where the convolutional neural network was able to moderately classify the first audio as a Red-winged Blackbird with 50% accuracy, while also showing ambiguity as other bird species may also be present in the background of the audio clip. Even though the model was not specifically trained to identify multiple bird species in one sound clip, it does a decent job at analyzing real world audio that might include difference species as well as background noises. The model does an even better job classifying audio clips 2 and 3, predicting the American Robin with 64% and 83% confidence. These high confidence predictions show that cleaner audio files, with less background noises and fewer bird calls, yields more confident results.

Conclusion

This report demonstrates the feasibility of using convolutional neural networks to classify bird species based on audio spectrograms, showing promising results for both binary and multi-class classifications. Even with limited training data, these models are able to extract meaningful patterns and make reasonably accurate predictions. These bird classification models can be implemented in the real world as a means to monitor biodiversity and bird species populations. With climate change and urbanization on the rise, it is important to protect the diversity and populations of species that could be at risk (3). Having a means to automate and track bird species without having to manually listen to live audio or being present at the scene could be the first step towards protecting biodiversity.

References

1. Pyle, P. (2024). Alpha codes for 2022 bird species.

   https://www.birdpop.org/docs/misc/Alpha_codes_eng.pdf

2. Xeno-cano Foundation. (2025.). *Canto*. xeno. https://xeno-canto.org/

3. 2025 National Audubon Society. (2024, January 19). *Survival by degrees: 389 bird species on*

   the Brink. Audubon. https://www.audubon.org/climate/survivalbydegrees

GitHub Link for code: nicolenagata/DATA5321_HW3: Convolutional Neural Networks