Christopher Yang
DATA 5322
May 11, 2025

Sound of Seattle Birds: Convolutional Neural Networks to Classify Bird Species Using Audio Spectrograms

Abstract

This report presents a deep learning based approach for classifying bird species based on audio recordings spectrograms of their calls using convolutional neural networks (CNNs). Two custom built CNNs were designed, interpreted, and evaluated for two classification tasks. One for binary classification between just 2 of 12 bird species common in the Seattle area, and another for multiclass classification across all 12 species. The spectrograms used in this analysis are representation of audio clips that were processed into 128 x 517 Mel spectrogram images, which served as our CNN inputs. To optimize the performance of each model, hyperparameters were tuned and complexities (layers of models) were experimented with. Our analysis also included 3 external test audio mp3 files which were converted to spectrograms and predicted on. Despite class imbalance and the computational restraint of a 2019 macbook, the best performing model for binary classification achieved 84% accuracy with 0.9 overall AUC, and the best performing model for multiclass classification achieved ~24% accuracy across the 12 species.

Intro

The goal of this project was to develop and evaluate custom CNN models capable of identifying bird species from spectrogram representations of audio recordings of their calls provided in a preprocessed HDF5 file. The dataset used for training the models are a subset of a Kaggle Bird Calling Competition dataset originating from Xeno-Canto [1], which is a crowd-sourced bird sounds archive containing over 264 species' bird calls. As mentioned before, both binary and multiclass classification models were developed and evaluated using just 2 species of birds for binary (House Sparrow and Song Sparrow) and all 12 species for multiclass. For binary classification, a more deep and complex network of 5 layers yielded the best results and captured more insight within our data while the multiclass classification struggled overall.

Theoretical Background

Convolutional Neural Networks (CNNs) are a type of deep learning model that has become a standard approach for image classification tasks due to their ability to learn hierarchical features from said images, like audio spectrograms. A CNN is typically made up of multiple layers that extract and combine different spatial patterns:

- **Convolutional Layers** - These layers apply learnable filters that extract local level features. Each filter detects a specific feature like frequency modulations or time-based chirps in our case. These layers typically need the number of filters tuned properly so that we do not use too many or too little parameters in our model
- **Activation Functions** - These introduce non-linearity in our model, which allows the model to learn and capture more complex patterns in our data
- **Pooling Layers** - These layers reduce spatial resolution to help generalize the learned features. Typically 2x2 windows are used, but this can be changed to 3x3 or something else depending on how your model is performing
- **Dropout Layers** - These layers randomly disable a subset of neurons in our model during training to prevent overfitting. These layers need tuned rates, so that we do not drop too many neurons or keep too many either. It is good to scale the dropout rate with the complexity of the model or even filter sizes.
- **Fully Connected (Dense) Layers -** These layers combine the features that we learned for the final predictions our model is going to make.
- **Global Average/Max Pooling** - This layer aggregates spatial features and reduce the parameter count while preserving the key activations to make the model more efficient. This is used often in place of flatten since flatten results in a very high number of parameters which make our model take longer to run or inefficiently overall.
- **Batch Normalization** - These layers stabilize and accelerate the training of our models by normalizing the layer's outputs

For binary classification, sigmoid activation functions are used with binary cross entropy because we are classifying between 2 classes. For multiclass classification, softmax activation functions are used with categorical cross entropy because we are classifying across many different classes (more than 2 classes).

To combat overfitting and instability in our models when training, there are many different regularization techniques:

- **L2 regularization (weight decay)** is often used to penalize large weights to allow the model to prevent one class/feature from essentially taking over the model
- **Dropout** with variable rates are used (scaled by layer) to prevent overfitting
- **Early Stopping** is used to stop training once the model stops learning effectively. This is usually indicated with the validation loss of the model when it stops improving
- **Reduce Learning Rate on Plateau** is used to dynamically lower the learning rate of the model when the model stops improving meaningfully

Evaluation metrics include:

- **Accuracy**: This checks the proportion of correct classifications/predictions
- **Confusion matrix**: This helps us visualize the correct/incorrect classifications of our model especially for multiclass
- **Precision/Recall/F1-Score**: Accuracy can be misleading especially when the data contains class imbalance, so looking at precision, recall, and f1-score provides a more

meaningful understanding of false positives, true positives, and the balance between the two

- **AUC**: Used mainly for binary classification, but can technically be used for multiclass as well even though it is much more complicated with multiclass. For binary at least, it measures the ability to separate between the two classes.

Methodology

1. Data Preparation

The data was provided in preprocessed HDF5 format [2]. Each sample represents a 2 second spectrogram of a bird call in the shape (128 x 517). For binary classification, the data was subset further into the two species of interest (Song Sparrow and House Sparrow) since these two species contained the most samples and models do better when they have a lot of data to train on. For multiclass classification, the species were one hot encoded across the 12 species, and some data augmentation was attempted via width and height shifts to try and address the large class imbalance in our data. For both tasks, each spectrogram was normalized to [0, 1], the training and test data were split stratified with a ratio of 80/20.

2. Binary Classification Flow

(1) Gather input spectrograms for the two species (Song Sparrow and House Sparrow)
(2) Preprocess by normalizing and reshaping to correct 4D tensors
(3) Model Architecture:
  (a) Model 1: Simple 2 conv layer CNN with flatten and dense layers
  (b) Model 2: Deeper 3 conv layer CNN with dropout, L2 regularization, and global average pooling instead of flatten
  (c) Model 3 & 4: Final 5 conv layer networks with scaled dropout, ‘same’ padding, label smoothing, and differing batch sizes (16 for model 3, 32 for model 4)
(4) Optimization: Adam optimizer with adjusted learning rate (0.0001), early stopping, and class weights to try and handle some class imbalance
(5) Evaluation: test accuracy, AUC, precision, recall, f1-score from classification report

3. Multiclass Classification Flow

(1) Organize spectrograms for all 12 species
(2) Preprocess by label encoding, one hot transformation, and class weights calculation
(3) Model Architecture:
  (a) Model 1: Used model 3 of binary, changed the activation function to softmax
  (b) Model 2: 4 conv layer simpler model with smaller dropout and global max pooling instead of average pooling. Also used reduced L2 and adaptive learning rate
  (c) Model 3: 3 Conv layers, dropouts between 0.3-0.4, and higher learning rate

(4) Optimization: Adam optimizer, label smoothing, early stopping, reduce learning rate on plateau, and class weights
(5) Evaluation: test accuracy, precision, recall, f1-score

4. 3 External Audio Predictions

(1) Raw MP3 clips were converted to spectrograms [3]
(2) Preprocessed spectrograms to grayscale, resized, and normalized
(3) Made final predictions using Model 3 of multiclass classification (not good results, but were the best results from the models) with top 3 classes reported

Results

1. Binary Classification Results

The binary classification models tried to distinguish between our 2 chosen species: Song Sparrow and House Sparrow. Below is a table containing the results for the first, boilerplate model and the best performing model for comparison.

To reiterate:

- Model 1: Simple, boilerplate model with 2 conv layers with only flatten and dense layers
- Model 3: Deep, complex model with 5 conv layers with scaled dropout, ‘same’ padding, and label smoothing

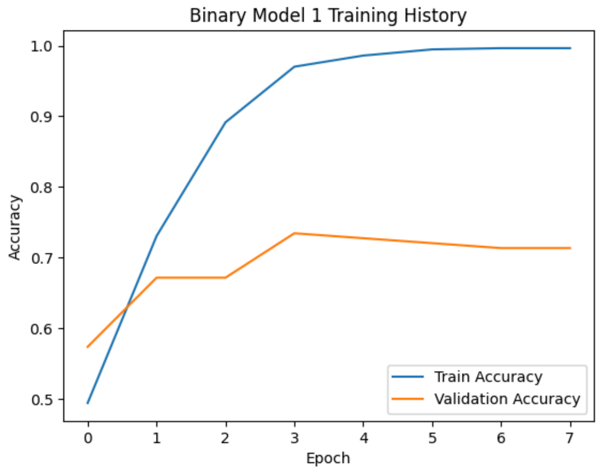
In the table below, we can see:

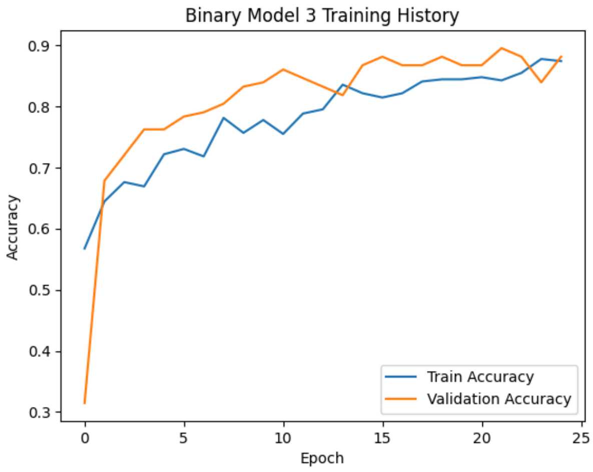
- The simple model (model 1) did relatively well considering how boilerplate it was. 72% accuracy with an AUC of 0.65 is doing better than random guessing and meaningfully learning the patterns of our data. The precision of 0.79 and recall 0f 0.81 for House Sparrow indicates that the model does well in classifying House Sparrows. On the other hand, the precision of 0.52 and recall of 0.49 for Song Sparrow indicate that this model struggles with classifying Song Sparrows. We can also see the class imbalance with the number of samples in each class, which led to the model performing better regarding House Sparrows than Song Sparrows.
- The more deep model (model 3) did much better since it had more tuning and layers to it. 86% accuracy with an AUC of 0.9 indicates that our model can effectively distinguish between the two classes. Despite also having that class imbalance, model 3 had a lot more balance in the precision and recall than model 1 indicated by the macro f1-score of 0.82. One thing to note is that while model 3 did perform much better, it took over 11 times longer to run than the simpler model 1.

Metric	Model 1	Model 3
Accuracy	72%	86%

AUC	0.65	0.90
Precision (Song Sparrow)	0.52	0.82
Precision (House Sparrow)	0.79	0.87
Recall (Song Sparrow)	0.49	0.68
Recall (Song Sparrow)	0.81	0.94
Samples (Song Sparrow)	53	53
Samples (House Sparrow)	126	126
Macro F1-Score	0.65	0.82
Time	~2m	~23m

Below are also the plots of training and validation accuracy over the epochs trained. We can see that the simple model (model 1) had some overfitting seen with the large, increasing gap between the training and validation accuracy as the training progresses, while the deeper model (model 3) had little to no overfitting seen with the similar training and validation accuracy, which is what we want to see. Also note that the simpler model stopped very early on compared to the deeper model.





2. Multiclass Classification Results

The multiclass classification models tried to distinguish between all 12 of the species in our dataset. Below is a table containing the results of the first multiclass model and the best performing model for comparison.

To reiterate:

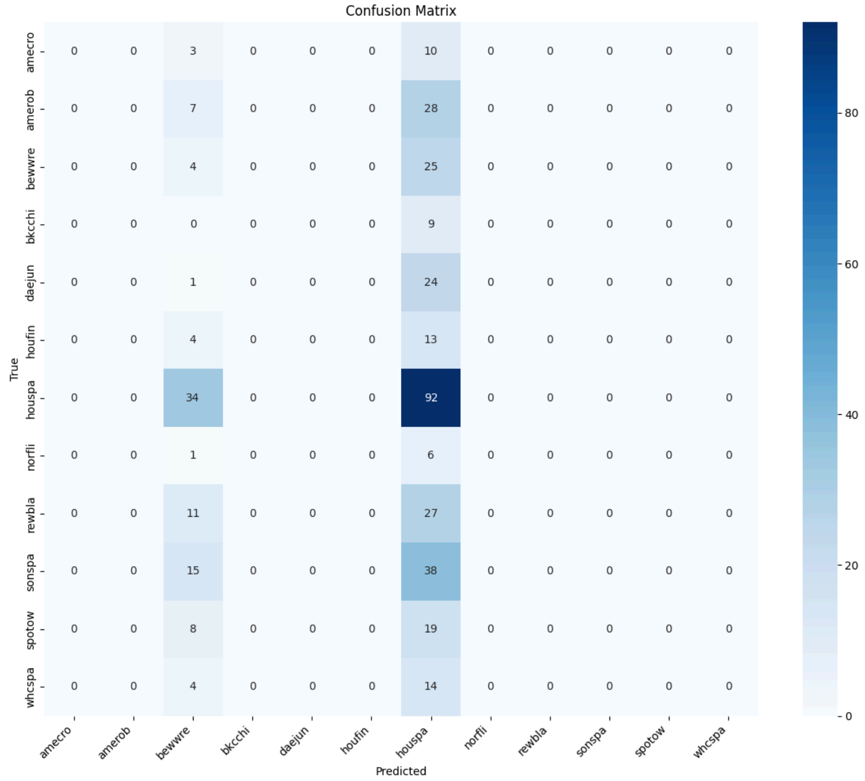
- Model 1: Deep 5 conv layer model with dropout scaling from 0.2-0.5, L2 of 0.0001, learning rate of 0.0005, and global average pooling
- Model 3: Simpler 3 conv layer model with dropout scaling from 0.3-0.4, L2 of 0.0001, learning rate of 0.005, and global max pooling

In the table below, we can see:

- Despite one being 3 conv layers and the other being 5 conv layers, both models had similar results across the board. The simpler model (model 3) did do better in each of the metrics, but only by a little bit.
- Overall, both models did terribly. 24% and 26% are not great accuracies at all, however, both models did do better than random guessing (~0.8333%), which means that each model, at the very least, learned something from the data and was able to identify a meaningful trend or pattern.
- Also note that the simpler model (model 3) took longer to train than the deeper model (model 1)

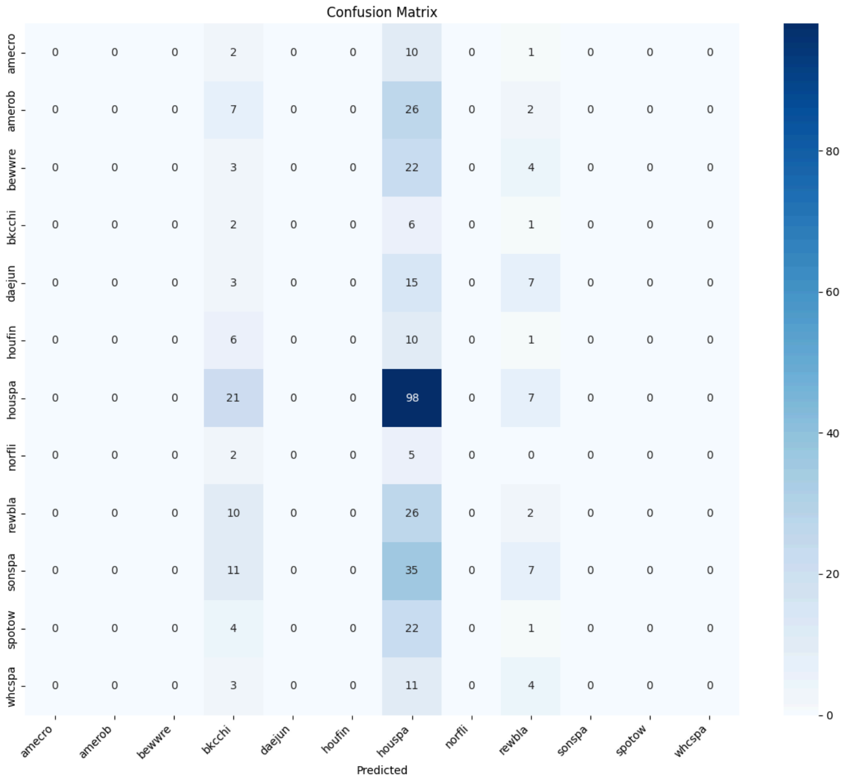
Metric	Model 1	Model 3
Accuracy	24%	26%
Macro Precision (across all species)	0.288	0.0353
Macro Recall (across all species)	0.0723	0.0877
Macro F1-Score (across all species)	0.0411	0.0481
Time	~ 12m	~ 23m

Below is also the confusion matrix of each model. Again, we can see that because of class imbalance, both models mainly classified the species to be House Sparrow. Another thing to note is that both models did not have much diversity in classifying. Due to the class imbalance, each model only classified 2 or 3 species at most where one was guaranteed to be the House Sparrow with the majority of samples. Something interesting was that other than classifying House Sparrows, they disagreed on what the other species were. Again, with class imbalance, the models will likely pick the majority class most of the time, which is what we see. The other classes each model chooses to identify vary depending on the model and its hyperparameters and complexity. **Check the appendix at the end of the report for each species' common name.**



True \ Predicted	amecro	amerob	bewwre	bkcchi	daejun	houfin	houspa	norfli	rewbla	sonspa	spotow	whcspa
amecro	0	0	3	0	0	0	10	0	0	0	0	0
amerob	0	0	7	0	0	0	28	0	0	0	0	0
bewwre	0	0	4	0	0	0	25	0	0	0	0	0
bkcchi	0	0	0	0	0	0	9	0	0	0	0	0
daejun	0	0	1	0	0	0	24	0	0	0	0	0
houfin	0	0	4	0	0	0	13	0	0	0	0	0
houspa	0	0	34	0	0	0	92	0	0	0	0	0
norfli	0	0	1	0	0	0	6	0	0	0	0	0
rewbla	0	0	11	0	0	0	27	0	0	0	0	0
sonspa	0	0	15	0	0	0	38	0	0	0	0	0
spotow	0	0	8	0	0	0	19	0	0	0	0	0
whcspa	0	0	4	0	0	0	14	0	0	0	0	0

Figure 1 (above) shows the confusion matrix of multiclass model 1



True \ Predicted	amecro	amerob	bewwre	bkcchi	daejun	houfin	houspa	norfli	rewbla	sonspa	spotow	whcspa
amecro	0	0	0	2	0	0	10	0	1	0	0	0
amerob	0	0	0	7	0	0	26	0	2	0	0	0
bewwre	0	0	0	3	0	0	22	0	4	0	0	0
bkcchi	0	0	0	2	0	0	6	0	1	0	0	0
daejun	0	0	0	3	0	0	15	0	7	0	0	0
houfin	0	0	0	6	0	0	10	0	1	0	0	0
houspa	0	0	0	21	0	0	98	0	7	0	0	0
norfli	0	0	0	2	0	0	5	0	0	0	0	0
rewbla	0	0	0	10	0	0	26	0	2	0	0	0
sonspa	0	0	0	11	0	0	35	0	7	0	0	0
spotow	0	0	0	4	0	0	22	0	1	0	0	0
whcspa	0	0	0	3	0	0	11	0	4	0	0	0

Figure 2 (above) shows the confusion matrix of multiclass model 3

3. External Clip Predictions

Below is the table containing the predictions that my best multiclass model made (multiclass model 3). As mentioned before, the multiclass models did not do well. As we can see, due to the class imbalance, not only did our model predict House Sparrow (majority class) to be the top prediction, but predictions for all three unknown spectrograms were the same with very slight differences in probability, which is likely from rounding the result of a very small pattern the model did end up finding. Although by listening to the audio clips I can sort of tell maybe the 1st spectrogram has multiple birds, because of the similar results in the prediction data, it is hard to tell if there are multiple birds purely based on the predictions.

Unknown Spectrogram	1st Prediction	2nd Prediction	3rd Prediction	Multiple Birds?
1	House Sparrow - 0.2071	Black-Capped Chickadee - 0.1657	Red-Winged Blackbird - 0.1278	No
2	House Sparrow - 0.2071	Black-Capped Chickadee - 0.1658	Red-Winged Blackbird - 0.1279	No

3	House Sparrow - 0.2071	Black-Capped Chickadee - 0.1657	Red-Winged Blackbird - 0.1278	No

Discussion

This project aimed to classify the species of bird based on spectrograms using custom convolutional neural networks. While the binary classification models yielded great results, the multiclass classification models struggled, showing to be more challenging with the limitations of the imbalance in data and complexity of the overall problem.

1. **Binary Classification (Song Sparrow and House Sparrow)**

The binary classification models showed the benefits of a deeper architecture and meaningful regularization techniques. The simplest, boilerplate model (model 1) was able to achieve 72% accuracy and 0.65 AUC suggesting that even basic CNNs can extract meaningful insights from the spectrograms. However, the simple model did show clear bias towards the majority class (House Sparrow) with the high precision and recall. There was also some overfitting seen with the increasing gap between the training and validation accuracy, which confirmed the need for more complex tuning of the model's parameters.

By model 3, we had multiple improvements, such as scaled dropouts, padding, label smoothing, and reduced learning rate, which led to a more robust model. Model 3 was able to achieve 86% accuracy with an AUC of 0.9 plus much better class balance. The improvement in F1-score and the little to no overfitting in the training curves indicate that the model generalized well even with the imbalance in the data. Although model 3 did take longer to train, the tradeoff was worth it.

2. **Multi-class Classification (All 12 species)**

The multi-class classification models struggled a lot more due to the class imbalance. Some classes had 10 times more samples than another, which resulted in our heavily biased models towards the dominant House Sparrow class. Further, the overall sample size of the dataset (couple thousand samples) proved to be inefficient in providing strong models, especially across a large number of classes with large class imbalances. The confusion matrices also showed that the models exclusively predicted House Sparrow indicating that the efforts of data augmentation and class weights were not enough to handle the class imbalance.

On the bright side, both the simple 3 conv layer model (model 3) and the deeper, complex 5 conv layer model (model 1) performed better than random guessing, which meant the models did learn something and identified something meaningful in the data.

Something interesting with multi-class was that the simpler model performed better than the deeper model. This could be a one off case, but I think it is worth noting that simple models not just with CNNs can prove to be more effective than more complex ones.

3. External Audio Clip Predictions

Predictions on the 3 external audio clips using the best performing multi-class model gave the same result indicating House Sparrow was the most likely species. The predictions did have very slight differences, so while the model did learn something, it was unable to generalize to a less common species in the data.

The differences in the spectrograms of the clips did lead to subtle differences in prediction probabilities, which suggest that with more tuning and balanced, large dataset, the models might be able to distinguish a lot better.

4. Limitations

A limitation that I ran into was with hardware. Because I was using a 2019 macbook, it took a long time to train and some variations of my model would not end so I had to use less aggressive parameters. The results in the above sections took 30 minutes at most, but the ones that I did not include took over 3 hours to run even with early stopping.

Another obvious limitation was the data size (data size per class as well) and the imbalance in samples. The severe class imbalance led the multi-class models to struggle.

5. Species-Specific Challenges

For binary classification, it was not too challenging to predict the minority class Song Sparrow.

For multi-class classification, it was hard to predict anything other than House Sparrow in terms of the model, so it basically got confused with every other class pretty often. I could sort of tell from just listening to the audio recording that the test1.mp3 had multiple birds, but again, I do not know bird calls. I do think with multiple birds, the spectrogram would only note the more distinguished bird if their frequencies are the same/similar. It might be a reach, but something like Black-Capped Chickadee and Song Sparrow seemed to get mixed up a bit based on our confusion matrix of model 3.

6. Alternatives

We could have tried using pre-trained audio models like YAMNet which is trained on large audio datasets and applied transfer learning to fit our needs.

Another option, although they would be slower to train, could be to use Recurrent Neural Networks (RNNs) or even possibly LSTMs since there are temporal aspects to the data.

Another possibility is to use SVMs or Random Forests (I think they had spectral contrast which could be used here, but not 100% sure).
Regardless, CNNs seem to be the logical choice since our data is images with species specific traits and CNNs are great at detecting and learning local features specifically when spatial locality matters like in our case. Also CNNs are more computationally efficient than RNNs like mentioned above.

Conclusion

This project showed the feasibility and limitations of using CNNs to classify bird species from audio spectrograms. The binary classification models showed that effectively regularized and deep models can perform well even with class imbalance and encourages the use of CNNs in ecological monitoring applications where data for target species is abundant.

At the same time, the multi-class classification models showed that the limitations in data quality and quantity can also limit the performance of the model itself.

These results suggest that for real-world applications using deep learning:

- Balanced and abundant data is essential
- Decent hardware is also beneficial

Despite the hardware and data limitations, this project offers a valuable baseline for future work. It reiterates the importance of data focused model development and shows how even small changes in architecture or preprocessing can lead to significant performance differences.

Works Cited

1. Rao, Rohan. *Xeno-Canto Bird Recordings Extended (A–M)*. Kaggle, 2024, https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-a-m.

2. Mendible, Ariana. *Bird Spectrograms Dataset*. 2025. GitHub, https://github.com/mendible/5322/blob/main/Homework%203/bird_spectrograms.hdf5.

3. Mendible, Ariana. *Test Bird Audio Clips*. 2025. GitHub, https://github.com/mendible/5322/tree/51bd4705bf06d4f46e1848f9261da9b2db3333c0/Homework%203/test_birds.

Appendix

Species Code	Common Name
amecro	American Crow
amerob	American Robin
bewwre	Bewick’s Wren
bkcchi	Black-Capped Chickadee
daejun	Dark-Eyed Junco
housfin	House Finch
houspa	House Sparrow
norfli	Northern Flicker
rewbla	Red-Winged Blackbird
sonspa	Song Sparrow
spotow	Spotted Towhee
whcspa	White-Crowned Sparrow