MACHINE LEARNING CLASSIFICATION OF NATIONAL PARKS BIRD AUDIO DATA

by

BENJAMIN FLANDERS

(Under the Direction of Adam Goodie)

ABSTRACT

The South East Coast National Parks Inventory and Monitoring Division has designated landbird presence as a vital sign for measuring ecosystem health. Technicians analyze audio recording device output to determine landbird species presence. This thesis proposes a machine-learning-based process for segmenting and classifying bird vocalizations from an internal South East Coast National Parks Landbird Audio dataset. We first generate a binary vocalization problem-set using a modified Democratic-Co-Learning approach to construct a missing binary element. We then form a multiclass bird species dataset. Using these problem-sets, we train various machine learning classifiers and use a DWT-MFCC feature extraction approach that outperforms both DWT and MFCC. The thesis results in a thorough process that is adaptable to other signal datasets.

INDEX WORDS: Random Forest Classifier, Support Vector Classifier, AdaBoost, Democratic Co-Learning, Discrete Wavelet Transform, Mel Frequency Cepstral Coefficient, DWT-MFCC, Soft Thresholding, Class Imbalance Problem, SMOTE, National Parks, Landbird

MACHINE LEARNING CLASSIFICATION OF NATIONAL PARKS BIRD AUDIO DATA

by

BENJAMIN FLANDERS

B.S., University of Georgia, 2018

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

MASTER OF SCIENCE

ATHENS, GEORGIA

2021

MACHINE LEARNING CLASSIFICATION OF NATIONAL PARKS BIRD AUDIO DATA

by

BENJAMIN FLANDERS

Major Professor: Adam Goodie

Committee: Delaram Yazdansepas
John Gibbs

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
MAY 2021

DEDICATION

I dedicate this thesis to my parents for always loving and supporting me.

ACKNOWLEDGMENTS

I would like to thank my research committee. Delaram Yazdansepas, for your guidance and mentorship through this project and your teaching, which led me to pursue this maters degree. John Gibbs, for your guidance and friendship. Adam Goodie, for your advisement and help in making this project possible.

I would also like to thank the National Parks Staff. Paula Capece and Brian Gregory, for allowing me to use this data and for their diligent effort in creating the research assistantship that allowed me to pursue this master's program. Michael Parrish, for your tireless work in improving this dataset and for helping me to understand this data. To the rest of my friends at the National Parks System, for your support and hard work. To the National Parks System, for protecting our Nation's most precious resources.

I would lastly like to thank my family and friends. Brandi Sarama, for your unyielding love and support. Michael Mazzone, for your friendship and grammatical suggestions on this paper. Joan and Rob Flanders, my mother and father, who raised me to never back down from a challenge.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

§ 1.1 Introduction

The South East Coast United States National Parks Inventory and Monitoring Division (SECN-NPS-I&M) collects and analyzes landbird audio vocalizations to determine species presence and populations throughout the southeast region. These statistics help to assess each park's ecosystem health and inform NPS decisions. Creating a machine learning pipeline for audio signal analysis can lead to the rapid assessment of landbird populations and improve NPS's ability to treat these ecosystems.

The unofficial internal landbird audio dataset produced by the SECN Wildlife Team is an example of a large dataset that is not immediately ready for machine learning use. There are various obstacles to overcome, such as the occurrences of landbird vocalizations are labeled, while non-vocalization data is not marked. To produce a binary classifier, we must label negative data through unsupervised learning or manual classification. Co learning combines the benefits of human-in-the-loop methodology with unsupervised classification, one such variant of co-learning being democratic co-learning. [1]

The Landbird dataset's complex label space reflects the high species diversity in the southeast region and reflects an imbalance in population sizes across bird species. One method for solving this data imbalance is through dataset balancing techniques and reusing data where possible.

In cutting-edge Artificial Intelligence, signal processing relies on neural network-based approaches performed directly on a signal, visual analysis of signal imagery, or feature extraction of signal data. Discrete wavelet transform is a modern solution for resolving signal data to its corresponding time-frequency domain; DWT is often preferred to its predecessor, Fourier Transform, due to its superior time resolution.

[1] Each vocalization occurrence is also labeled with a bird species.

The discrete wavelet transform uses the products of a signal and a predefined wavelet to produce DWT features which include an approximation coefficient and detail coefficients.

DWT output is structurally similar to a signal because each coefficient has a frequency value for each time sample. As a result of DWT output structure, individual decompositions can be input for feature extractions typically performed directly on a signal, such as generating Mel-frequency cepstral coefficients. Alternatively, summarizing features such as mean or median can mine valuable information from each decomposition.

After creating a machine learning-ready dataset of feature-label pairings, various classifiers can learn relations in this type of data, including Random Forest, Support Vectors, Neural Networks, or additional ensemble classifiers. Random Forest is one popular choice that uses an ensemble of decision tree classifiers to make a unified prediction. This lightweight classifier is ideal for rapid data labeling, given a sufficient feature space.

This thesis's first objective is to provide a feature space comparison for audio signal data that performs well when applied to landbird vocalization data. The first step will be to compare discrete wavelet transform (DWT) output with MFC output for the same signal input. Additionally, the combination of these two feature extraction methods, DWT-MFCC, is tested to see if it provides an improvement over its components. Next, these large feature spaces are reduces using feature reduction methods.

This paper's second overall objective is to propose a machine learning process aligned with the NPS SECN landbird collection procedure. This paper describes and justifies methods for the following: data processing, large audio file segmentation, short audio clip classification, and co-learning methods for labeling unknown data efficiently.

In addition to supporting the National Parks in their data collection efforts, this paper aims to expand on existing processes for rapidly labeling and mining useful information from existing data sources. The case of a partially labeled dataset is not a unique one, and modern AI benefits heavily from large labeled datasets. This paper gives one solution for efficiently producing labels for unknown data while relying on data already known. The level of expert knowledge required to utilize the practices described will vary significantly across problem spaces. The presence of bird sounds, for example, is easily detectable to the layman's ear. A more complicated problem that would require expert knowledge would be to classify the exact species of a bird. [2]

[2] This introduction loosely follows the methodology section to show the necessity of splitting the dataset into a segmentation portion and a separate classification dataset.

CHAPTER 2

RELATED WORK

§ 2.1 National Parks Wildlife Team Background

The National Parks South East Coast Network (NPS-SECN) Inventory Monitoring Group (I&M) includes a wildlife division that monitors wildlife species populations. The wildlife team primarily utilizes automated recording devices (ARDs) for their monitoring efforts. The team's process is formalized in MW et al., 2014. The wildlife team's objective is to monitor and study land-bird populations in the southeast United States; this objective has led to creating an internal audio dataset of audio recordings and landbird species labels. The SECN has designated landbirds as a 'vital sign' of the region's ecosystems, meaning population data can provide valuable insight into that ecosystem's health. According to MW et al., 2014 Landbirds are a vital sign due to the high correlation between land bird populations and the health of their associated ecosystem.

The SECN Wildlife Team passively records audio in pre-planned locations and then transcribes the collected audio sound with landbird species labels. Improvements in bird tracking technology can help scientists spend less time on manual processing of bird audio and re-focus on understanding the migratory and ecological implications of this audio. One key component of leveraging in-field audio recording devices is to separate bird sounds from background noises such as running water, traffic, other animal noises, and static of any sort.

Technicians deploy Audio Recording Devices (ARD) to monitor land birds. The ecologists responsible for data gathering will pre-determine site locations ARD site locations. These locations are chosen semi-randomly and cover as much of a park as possible. Further analysis of audio recordings helps determine the success of a site location, and proper labeling of the ARD output data is a predecessor for this analysis. These ARDs can also pick up numerous other sounds. Some examples are car noises, rain, streams, and other traffic.

The South East Coast National Parks I&M division does propose an automated analysis of audio data. Still, the current operation could benefit from an improved classification method and a classification system specifically built for the southeast bird species.

Much of the audio processing methodology is performed by human listeners, although some of the data is unclassifiable. This thesis project will handle unclassifiable data as non-vocalizations (or 'negative' for our purposes). This project intends to develop a process that pairs well with the ecologist's process and improved accuracy on a specifically tailored dataset to the SECN ongoing project.

This paper presents a workflow for leveraging the NPS SECN bird audio dataset's current state. The dataset is currently closed to the general public to prevent unintended consequences of species location knowledge. The dataset contains information about where in audio bird songs are occurring and the bird species vocalizing. Transcriptions of the audio data can be obtained at irma.nps.gov.

§ 2.2 Literature Review

§ 2.2.1 Dataset Preparation

Current signal processing methods include analyzing images derived from signals and extracting features directly from signal data. Convolutional Neural networks excel at the direct processing of image data, and this image data is often spectrogram imagery of an audio signal. Alternatively, feature extraction can be performed on signal data directly and be used to train various classifiers ranging from decision trees to neural networks. The correct choice of a classification method for signal processing continues to be an open question under investigation.

In 2013 the 9th annual MLSP competition tasked participants with classifying bird species in noisy environments Briggs et al., 2013 which demonstrates the openness of this problem space and the wide variety of processes available.

During this competition, the first step of representing the bird audio data primarily involved feature extraction from wave data, but teams also used spectrogram image representation of audio signals. Teams often used spectrogram image representations of audio signals and then applied FFT to the image frames to produce time-frequency domain features. Some teams found success using spectrum-summarizing features. For example, MFC could divide the spectrogram into bands, and the amplitudes of these bands are classifier features. Other summarizing features are used, too, such as band mean, median, min. The competition also demonstrated success that a focus on feature engineering with relatively simple classifiers, such as decision trees, was sufficient for performing well in the competition. Notably, some teams placed well using Convolutional Neural Networks that ignore the decision tree classifier methods' feature engineering requirements but lack the transparency of such a method. The neural network solution can be challenging to adapt and replicate, especially in this thesis which intends to construct a transparent classification pipeline.

The competition initially uses a segmentation form of the dataset where spectrogram data is highlighted as red or blue depending on if there is a presence of bird sound or not. This thesis will not rely on spectrogram-based methodology. The presence of a vocalization could also be judged by time, but the spectrogram gives more information on frequency ranges and bird sound power. Additionally, DWT provides better time resolution according to Yadav et al., 2015, so this tradeoff could result in higher accuracy. Some of the teams used a classifier that could automatically segment spectrogram images; however, others found success using features that did not require image segmentation or did not require segmentation at all. Spectrogram-based segmentation would allow for greater visibility and understanding for technicians involved but requires pre-labeling that is not a part of the SECN landbird dataset. Ultimately, to identify vocalization segments, classifiers will need both positive and negative classified segments of audio data.

The featured dataset had 19 bird classes, which is significantly less than the SECN's datasets of 99 classes. This paper also does not include the complexity of learning permanent data and ensuring its quality for future use in a situation where data must be validated significantly, as is the case with the SECN data verification methodology in MW et al., 2014. It is reassuring that these competition methods would likely perform well on a subset of the NPS 99 bird labels.

§ 2.2.2 Segementation of Audio Data

In the paper, L. Lu et al., 2001, authors first segment audio into speech and non-speech data and then, converse to the NPS procedure, apply a label to each non-speech data segment. The authors use a K-Nearest-Neighbor-based method of segmentation. This method simplifies labeling to four classes and classifies the remaining non-speech sounds into three classes. The authors find that this two-step segmentation-classification process is very successful when used with their feature extraction methods.

§ 2.2.3 Classification

Semi-supervised training techniques have been employed to handle datasets where some data is labeled and other data is not. A paper that demonstrates the successful use of co-training to training classifiers on a limited set of annotated audio data is Xu et al., 2005. This paper applies an ensemble of classifiers with various sets of feature data on a small dataset of music audio signal data. The author uses a summarized multi-view feature data from performing STFT, DWT, and MFCC to the signal data. The method involves labeling the entire unlabeled dataset and then only moving the two most likely classifications for each class to the labeled dataset. The resulting accuracy of the ensemble methodology presented is higher than any single classifier accuracy.

Co-learning is another semi-supervised learning method for utilizing classifiers to label unseen data. The paper Zhou and Goldman, 2004 presents one method of co-learning called democratic co-learning. Democratic co-learning is the process of training an ensemble of classifiers on unknown data and using their vote to label unknown data. Democratic co-learning relies on various classifiers rather than a singular classifier to minimize any single classifier's influence. Each classifier votes and the majority group's label is applied to the datapoint. One caveat is that for the label to be applied, the majority group's average confidence must exceed the average confidence of the minority group prediction.

Co-learning can potentially produce incorrect labelings and propagate these mislabeled data points into the training dataset, especially the training dataset is small. Human-in-the-loop methods can improve Co-learning classification by expanding the amount of known data.

In addition to labeling unknown data, the SECN landbird dataset has an issue with imbalanced classes. One option for resolving data imbalances in a multi-class dataset is to strategically sample data points for a more balanced dataset. The paper Yap et al., 2014 compares some of the most popular sampling methods in data mining in the context of a cardiac surgery dataset. These sampling methods include oversampling, undersampling, bagging, and boosting. The authors found that oversampling causes overfitting because of the duplication of data in minority classes. Oversampling also allows all majority samples to be maintained, preventing any loss of training data. The authors found oversampling and undersampling performed equally or better than boosting and bagging. The paper acknowledges that there are hybrid methodologies that are more complex, but they are often not yet implemented in software applications and not accessible to beginners. It seems that such complex methods are not conducive to building a flexible segmentation and classification process. This paper's insights are somewhat limited in their reach because of their basis in a binary dataset and do not necessarily extrapolate to a multi-class situation.

The problems of a multi-class large dataset with imbalanced class data are addressed further in Bhagat and Patil, 2015. The authors use random forest with "One-Vs-All" (OVA) classification and Synthetic Minority Oversampling Technique (SMOTE) to improve a variety of testing performance metrics. The paper states three primary methods of handling large and imbalanced datasets: Data Level Approach, Algorithm Level Approach, and Cost-sensitive Approach. Cost-sensitive approaches are essentially a combination of data level and algorithm level approaches. The authors' process involves converting their dataset labels into a series of One-vs-all labels (OVA) where labels are encoded as binary vectors rather than categorical variables (integers). The authors compared OVA classification via Random Forest with OVA + SMOTE with Random Forest and found that SMOTE + OVA had an excellent performance on various imbalanced UCI datasets.

The authors of Drummond and Holte, 2003 used cost curves to quantifiably compare over-sampling to undersampling results. The authors found that oversampling was ineffective and led to a poorer dataset. The experiment also found that undersampling provided accurate results, but undersampling injected randomness into an otherwise deterministic process.

Two examples of algorithm-level approaches to handling imbalanced class data are clustering methods (K-Means, hierarchy decomposition) and using classifiers specifically capable of handling this type of data (Ensemble classifiers, SVM).

One popular classifier for understanding the relations between MFC coefficients and a label is Naive Bayes. Bhakre and Bang, 2016 shows that Naive Bayes performs well with MFC-based feature sets, especially when the dataset is small. The conditionality and relations of an MFC feature vector are suitable for Naive Bayes classifiers. Unfortunately, Naive Bayes performs poorly at generating confidence scores, according to Pedregosa et al., 2011.

Support Vector Machines (SVM), an alternative to Naive Bayes, could perform well on a linearly separable dataset. Ahmad et al., 2016 shows that SVM outperforms Naive Bayes on gender identification when using MFC features on an audio dataset. SVM might perform better as long as there are not too few features, which Pedregosa et al., 2011 acknowledges could be a drawback to their library.

§ 2.2.4 Feature Extraction

There are four primary feature domain choices for signal data. These options are described in Sharma et al., 2020 as time-domain features, frequency domain features, time-frequency domain features, and deep-features. Time-domain features typically use a sliding window over audio data with feature extractors such as Zero-Crossing Rate described in Bachu et al., 2008. Alternatively, a signal can be converted to its frequency space using Fourier Transform (FT), which solves the problem of abrupt changes caused by windowing the data for time-domain data. One frequency-domain feature process is the extraction of LPC Coefficients in O'Shaughnessy, 1988 for compression of a signal into corresponding linear coefficients. LPC performs very well in speech but might not translate well to environmental sounds because it relies on encoding portions of sound that human vocal tracts are known for making rather than based on bird vocal tracts.

Another feature extraction method that relies on Fourier Transform is Mel-frequency Cepstrum (MFC) and is presented in Davis and Mermelstein, 1980. MFC generates a set of coefficients (MFCCs) used commonly as feature information in signal processing. MFC transforms signals into a Mel-Scale space that separates signals into a range roughly matching the same spacing between frequencies that humans perceive. This set of equally spaced frequency bands represents the power spectrum of a signal—the similarity to a human's perception has made it a popular feature extraction choice for syllable detection.

One method for converting signal data to the time-frequency domain is the Short-time Fourier transform (STFT) described in Bergland, 1969. This algorithm works by first converting a signal into time segments and then performing a Fourier Transform on each of these equal time segments. STFT suffers from a tradeoff between frequency and time resolution, which led to the advent of wavelet transforms.

Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) are wavelet-based methods that map a signal to the time-frequency domain. This transformation produces consistently high time resolution and lacks the tradeoff of frequency resolution reduction that SFTF has.

Discrete Wavelet Transform produces an approximation and detail coefficient at each decomposition level. Yadav et al., 2015 describes the specifics of this process in depth. The algorithm produces a detail coefficient that results from a mother wavelet convolution over a signal input and subsequent downsampling of the output; this is a high pass filter. DWT also produces an approximation coefficient by averaging and then downsampling the original signal. The detail and approximation coefficients both resemble the typical structure of a signal, as they have time and frequency components. Both components also have half the time and frequency space of the original signal input. Multiple wavelet decomposition levels are produced by repeatedly transforming the approximation coefficients of each wavelet transform.

Many feature extraction techniques in the time-frequency domain are adaptable to work with wavelet transform. The substitution of MFC's Discrete Cosine Transform for DWT is known as Mel-Frequency Discrete Wavelet Coefficients (MFDWC) and is described in Tufekci and Gowdy, 2000. In this algorithm, a signal input is framed and then translated by STFT. The output is then Mel-Scaled. Instead of MFC's discrete cosine transform, DWT is used and treats the Mel-Scaled output as a signal. This process produces wavelet transformed mel-scaled log filterbank energies of a speech frame. MFDWC demonstrates the application of DWT to MFC, but does not make up for STFT's low time resolution.

The paper Abdalla and Ali, 2010 shows an alternative to MFDWC where the output of DWT is input for the typical MFC algorithm. The algorithm uses STFT, but STFT acts on each approximation/detail coefficient rather than the original signal. Essentially, each coefficient is treated as separate signal input. The DWT-MFC coefficients are obtained from the Mel-scaled log filterbank energies of each approximation/detail-framed signal. This process does not eliminate the drawbacks of STFT but instead limits its influence by relying on DWT.

Neural networks are also widely used as a method for extracting features from signal data. A neural network's hierarchy extraction results are called 'deep-features.' Deep features have seen a rise in popularity and success due to recent advances in computing technology and access to large datasets. Examples include the autoencoder presented in X. Lu et al., 2013 or the Convolutional Neural Network in Piczak, 2015, both of which automatically reduce a signal to a feature vector through a learned mapping. These deep-feature mappings are learned via gradient descent and not designed to be human-readable. This lack of transparency makes it difficult to understand the meaning of any deep feature space.

Another example of deep feature representation of signal data is the paper Narasimhan et al., 2017, which segments and classifies simultaneously. The paper uses a CNN-based autoencoder capable of learning a mapping between a spectrogram image to a segmented version of that same signal. The neural network first encodes the spectrogram to a small number of digits representing the original signal, then decodes the image to its segmented version of itself. The encoded version of the signal is used as a feature vector because it represents the original signal. Still, it is unclear what any singular value in this feature vector means concerning the original signal.

Neural networks do tend to suffer from transparency and reproducibility issues, especially when trained on small datasets. Random forest, given a robust feature extraction method, performs well even on small datasets.

Mother Wavelets

The paper Wai Keng et al., 2013 discusses various approaches for selecting mother wavelets, both quantitative and qualitative in nature. Many of the methods presented were based on the similarity between the mother wavelet and signal data. The author acknowledges that more research is needed on the topic, especially around the accuracy of final results that a mother wavelet can produce.

The authors of Wai Keng et al., 2013 assess mother wavelet selection in the context of gear fault detection. The authors use passive recordings of electrical signals to detect faults. The authors note that Morelet is one of the most common mother wavelet choices across the field of machine condition monitoring and showed consistency across multiple signals. The authors also note that the basis for wavelet selection is often a function of minimizing the difference between the signal and the mother wavelet. Still, this method does not necessarily work in all situations. Lastly, the Daubechies wavelet shows near symmetrical properties at high orders (ex db44) but is otherwise asymmetrical.

Noise Reduction

Multiple thresholding techniques for signal data in the Time-frequency domain are proposed by Donoho in his paper Donoho, 1995. Two thresholding methods that Donoho proposes are hard thresholding and soft thresholding. Hard thresholding sets all data below some value to 0. Soft thresholding removes all data below a threshold and scales all other signal data towards 0 by that same threshold. The thresholding value for each algorithm is heavily influential. Donoho's universal threshold Donoho, 1995 and modifications of it in Aggarwal et al., 2011 have been found to reduce background noise.

CHAPTER 3

METHODOLOGY

This section presents a process for training a set of classifiers on the SECN Landbird audio dataset; each classifier aims to segment or classify unseen audio data. The methodology requires constructing two distinct datasets. The first dataset is a binary-labeled dataset of audio snippets classified as vocalizations and non-vocalizations. The second dataset is categorical and labeled with a vocalization's related bird species. There will be overlap in the two datasets, but the classifiers trained on each problem type will remain distinct.

§ 3.1 Data Preprocessing

The SECN Landbird dataset required preprocessing for use with both segmentation and classification. As described in section 2, the initial dataset was not ready for training use without some preprocessing. It also is not beneficial to know the existence of a binary indication of bird sounds in a multi-minute file. It is much more helpful to know the exact time that a bird vocalization happens and then count the number of occurrences in the larger audio file.

One-second samples were taken from the dataset and parsed into discrete vectors. The standardization of one-second audio clips allows us to extract features consistently regardless of the parent audio file size. Decoded audio signal data is in the form of a vector of numeric values, where each value in the vector is a single sample of audio data. The audio frame rate is 16000 frames per second, faster than the human ear perceives, yet a discrete sampling of a physical wave's speed.

The sample size of one second provides the convenience of having audio that is quick to listen to, minimal in data storage, and minimizes the resulting feature-space. One second empirically seems to be enough audio data to determine the binary presence of a bird vocalization. This one-second clip would not necessarily be enough for a human to determine bird species making noise. However, a simple solution is to add a buffer of time before and after a bird noise is found in a longer audio data piece and present that expanded time segment to a listener. A classifier could also predict a species for each one-second segment of an audio clip, and then the average prediction could be used as the final label.

For the segmentation audio dataset preparation, segments were labeled as 'Positive' if there was a known bird vocalization in the time segment; no label was applied to the dataset otherwise. Time segments of bird species classifications are not exact, but the time selected by researchers was assumed to be the start of a call. Many of the audio clips contained repetitive bird vocalizations that continued for many seconds after the time of record.

For each positive classification, processing produced three segments: one audio slice from the beginning of the vocalization for a total of one second and two more audio segments offset from the original clip by +- 500 milliseconds. This method of re-use creates more data while reducing manual classification.

In addition to the positively classified data, some negative (non-bird vocalization) data is necessary for machine learning classifiers to learn. The sheer amount of already classified vocalization data meant that the algorithm would not need any more (positive/species) vocalization data. The solution to zero initial negative data points was to classify data manually. As a starting point, I manually classified 100 seconds of non-vocalization data that primarily consisted of background noise with no bird vocalizations present. The subsection "Democratic Human In the Loop Co-Learning" further describes how algorithms continued this process of labeling negative vocalization data. This methodology is malleable to various co-learning situations regardless of dataset balancing. Notably, the segmentation (binary) data sampling from the overall dataset is non-deterministic throughout the multiple iterations of data gathered in the experiments. The positively labeled data heavily outweighs the non-vocalization data, so random undersampling corrects this and thereby adds randomness to the sampling process. In contrast, the exact dataset sampling used across the creation of the classification dataset is separate from the segmentation labeling. It is straightforward as the SECN dataset is already labeled to a high enough degree to be usable for bird label classification. Additionally, a bird species label requires expert knowledge, while a novice

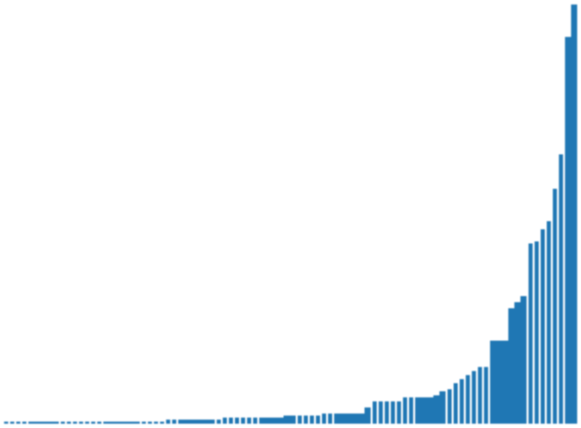


Figure 3.1: Bird count by bird species. Unlabeled to anaonymize dataset.

can determine and classify bird vocalizations even in a noisy environment. After processing, there appears to be a significant imbalance in class labels.

Figure 3.1 shows that the classification dataset is imbalanced in its labeling. The dataset reflects the monitored ecosystems, causing the dataset bird species classes to be imbalanced.

[1] Further information on this label imbalance issue is in the segmentation and classification portion of this methodology.

Signal Noise Reduction

As described in the background section, the impact of external noises could lead to reduced accuracies in classification. I apply soft thresholding with Donoho's Universal Coefficient. I intend to reduce noise in the audio data, improve accuracy, and improve scientists' manual classification ability without distorting the actual audio data.

[1] I removed all bird labels from 3.1 to preserve the anonymity of the dataset. Readers can find further information on published dataset findings at irma.nps.gov.

The Soft Thresholding function first performs a Discrete Wavelet Transform on audio data points and then soft-thresholds each detail coefficient and the approximation coefficient. The thresholding is applied to each level of wavelet decomposition with the Daubechies wavelet with eight decomposition levels.

Generally, soft thresholding works by replacing all values with an absolute value less than a set threshold with some substitute. The equation for soft thresholding is:

$$soft\_threshold(X) = \begin{Bmatrix} X/|X| * (|X| - thresh), & \text{if |X| - thresh} > 0 \\ 0, & \text{othwerwise} \end{Bmatrix}$$

The value X corresponds to some input signal data, such as wavelet decomposed coefficients which follow the same structure as their parent signal. 'THRESH' corresponds to a thresholding coefficient. Donoho presents an option for 'THRESH' called the 'Universal Threshold Coefficient.' In the context of this smoothing problem, Donoho's universal threshold coefficient presented in Donoho, 1995 roughy equates to the following:

$$threshold = \sigma n \sqrt{2 \log(N)} \tag{3.1}$$

N is the length of the soft thresholding input, and σ is the standard deviation of the noise at scale j. The noise present in this dataset is not necessarily Gaussian as Donoho, 1995 expects; the universal threshold could therefore have little impact.

§ 3.2 Feature Selection

I used three primary feature extraction methods for each audio data segment: DWT, MFCC, and DWT-MFCC. Lastly, K-Best attribute selection was performed based upon the ANOVA F-value of the selected samples to reduce the number of features. Feature extraction of audio data converts signal data from a waveform to a set of feature values for classifiers.

DWT is used for the first portion of the feature set and applied to the signal data. For each level of decomposition, there are multiple statistics generated from the feature data. The following is the set of operations performed on each signal's wavelet decomposition coefficients: mean, standard deviation, skew, variance, maximum value, median value, and minimum value. This process reduces each 1-second audio clip (16000 samples) to a total of 308 features. The primary wavelet used for the DWT operation is 'db8' with varying wavelets compared based on their accuracies. The maximum level of decomposition is always selected. The second feature extraction process is MFC. This method uses a filterbank computed from a signal, and then the log-absolute value of the produced feature set is computed. The final coefficients are a discrete cosine transform of the log-absolute values. The MFCC filter bank comprises 26 filters for every datapoint with the first 13 cepstral filters selected. Two more portions of the MFC feature set are delta and delta-delta. Delta is a measurement of change in the MFC feature values over the previous four frames, while the delta-delta is a measure of the rate of change in delta over four frames. DWT and MFCC both serve as control variables against the DWT-MFCC extraction method.

The third feature-set is DWT-MFCC which uses DWT coefficients as input to the MFCC algorithm. First, DWT decomposes an original signal input into multiple DWT decomposition levels. This process results in multiple detail coefficients and a singular approximation coefficient; each coefficient resembles a signal's structure. Each coefficient is a single input to the MFCC algorithm. Like MFC feature extraction, the delta and delta-delta features are appended for each of the MFC coefficient vectors. The complete process for DWT-MFCC is in figure 3.2.

§ 3.2.1 Feature Reduction

The DWT-MFCC extraction with a Daubechies scale eight mother wavelet produces a total of 1183 features per one-second audio clip. To reduce the number of features K-best selection is used with an f-score used for valuing each feature. This mapping can then be used later during testing or production without recomputing the feature selection list.

Feature Scaling

In addition to thresholding, a scaling function is used on sampled groups to prevent outliers further and improve training speeds. After all feature extraction and reduction are completed for a dataset, Scitkit-

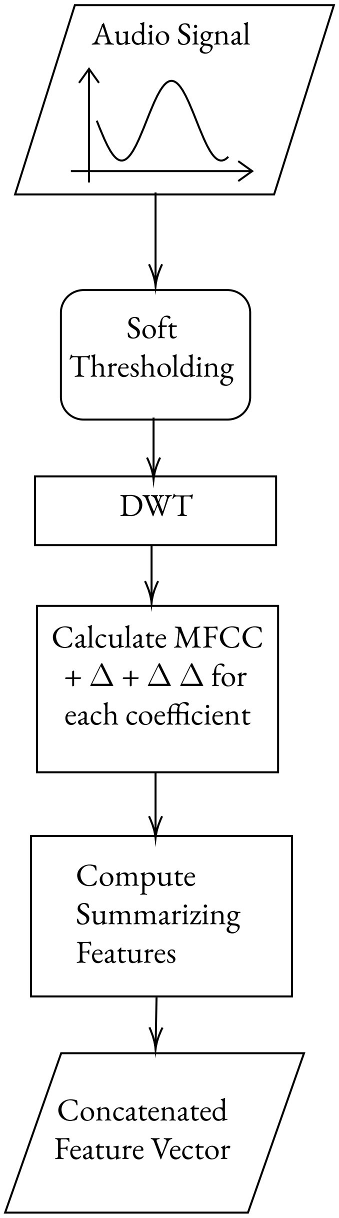

Audio Signal
Soft Thresholding
DWT
Calculate MFCC + Δ + Δ Δ for each coefficient
Compute Summarizing Features
Concatenated Feature Vector

Figure 3.2: DWT-MFCC Process

Learn's Standard Scaler, described in Pedregosa et al., 2011, is used on the entire dataset, which centers each feature to a zero mean and unit variance across the training set.

§ 3.2.2 Mother Wavelet

For convenience, the same mother wavelet is used for all DWT operations, including Soft Thresholding, DWT-MFCC computation, and DWT feature output. The primary choices of discrete mother wavelets are Haar, Daubechies, Biorthogonal, Coiflets, Symetrics, Morlet, Mexican Hat, and Meyer according to Lyons et al., 2020.

The land bird audio data abruptly cuts out in many audio clips due to the slicing from a parent audio clip and high background noise level. Tufekci and Gowdy, 2000 reports that antisymmetric wavelets tend to decrease discontinuities at the signal borders. The Debauchies wavelet will perform well on this type of problem because it is antisymmetric according to Vonesch et al., 2007.

The mother wavelet selected for this entire process is the Daubechies wavelet. This wavelet is also compatible with the Python Speech Features package presented in Lyons et al., 2020. The Debauchies wavelet with level 8 scaling, pictured in figure 3.3, is the mother wavelet choice in all methods that use a wavelet transform.

§ 3.3 Segmentation and Classification

Random Forest Classifiers were selected for both segmentation and classification. Additionally, a Support Vector Classifier and Ada-Boost Ensemble Classifier are used for the segmentation methodology to create a three classifier ensemble with a human-in-the-loop extension. Multi-class classification follows a more simplistic approach by solely using Random Forest Classification and relying on the dataset balancing section.

§ 3.3.1 Segmentation

The three classification networks for segmentation are Random Forest, Support Vector Machines, and Ada-Boost. Each classification algorithm provides its unique approach to classification and training while also being capable of running in a quick computational time. Each classifier predicts the class (0 or 1) for

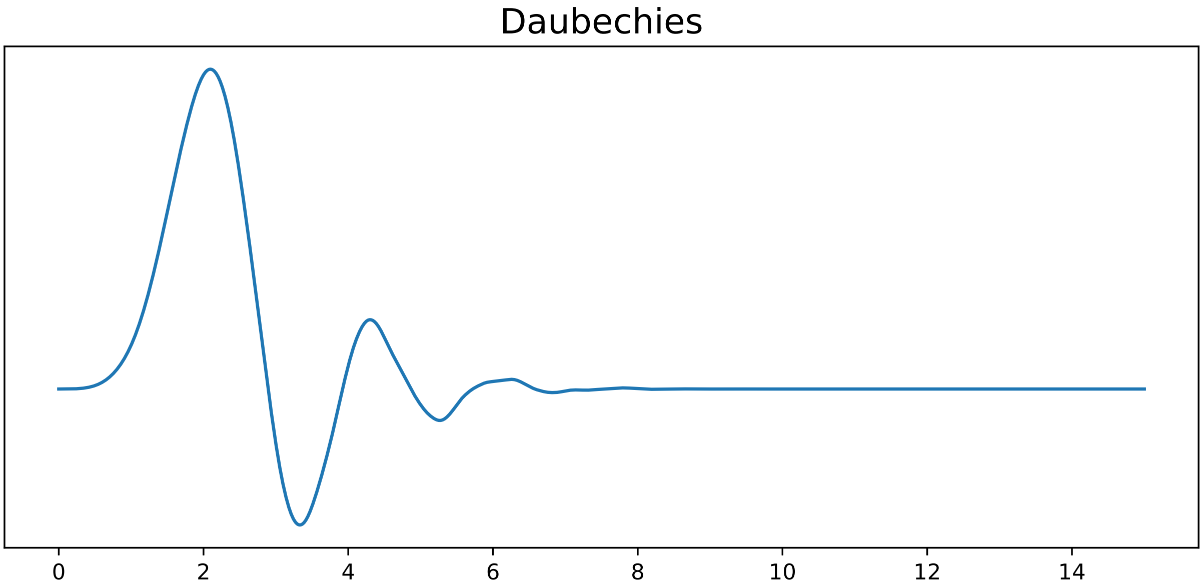


Figure 3.3: Daubechies mother wavelet with level 8 scaling.

each feature extracted training data. All three algorithms and their confidence scores are implemented via SciKit-Learn described in Pedregosa et al., 2011.

Random Sampling was used to select a subset of the positive classes, and the Edited Nearest Neighbor algorithm described in Wilson, 1972 undersampled the majority class further. SMOTE oversampled the minority class for the training data only. Both SMOTE and Edited Nearest Neighbor are accessible by Urbanowicz et al., 2017.

Democratic Human In the Loop Co-Learning

The Landbird audio dataset does not have audio labeled as a non-vocalization. It is not safe to assume that a lack of a vocalization label means there is no vocalization. There is nothing for a classifier to train without a 'negative' label based on its assumptions about a' positive' label. One method of solving this problem is to label audio snippets as 'negative' when there are no vocalizations present. To limit the amount of time it takes, we employed a human in the loop method towards data labeling.

A human in the loop methodology intends to make efficient use of a human's classification abilities. If machines can classify some of the data or determine a portion of a dataset that needs to be classified, less human work is required. Also, as we construct the dataset through co-learning, classifier scores should improve, and less human work will be required. The classifier employed avoid labeling data that it is uncertain about but can label data that it predicts with a high degree of certainty.

This paper's methodology uses a modified version of democratic co-learning presented in Zhou and Goldman, 2004. The modified process is presented in figure 3.3.1. The classifier ensemble will learn from data that the ensemble is confident of or has been pre-classified and mark all other data for human-in-the-loop classification. [2]

The classifier ensemble initially trains on the pre-labeled vocalization presence binary dataset. Next, each classifier votes by providing a label for each presented data point. If all labels were the same, then the label was assumed to be accurate and was applied to the data. This data goes into the training dataset for future use. If there was any disagreement in the label, then each classifier's confidence score was computed and used to settle the disagreement.

If the minority group's average confidence is higher than the average confidence of the majority vote group, then the datapoint is marked to be labeled by a human. I refer to this option as 'discarding.' This option forces the classifiers to discard data points that will have the highest impact on the classifiers while also minimizing the error introduced into the training dataset. [3]

Random Forest Classifier

The Random Forest Classifiers use 100 trees per forest, a Gini criterion of classification, and no maximum depth restriction.

Random Forest classifier is a decision-tree-based ensemble classifier. It is capable of handling the classification of both binary and categorical data. The random forest classifier is first proposed in Tin Kam Ho, 1995 and has amassed popularity due to its efficiency and accuracy, especially on small datasets.

[2] We initially labeled and marked 100 seconds worth of audio data that did not have any bird vocalization in it so that the co-learning ensemble would have a basis for training.

[3] Since the positive (presence of bird vocalization) portion of the dataset already exists, it does not make sense to increase the size of this portion of the dataset and is discarded.

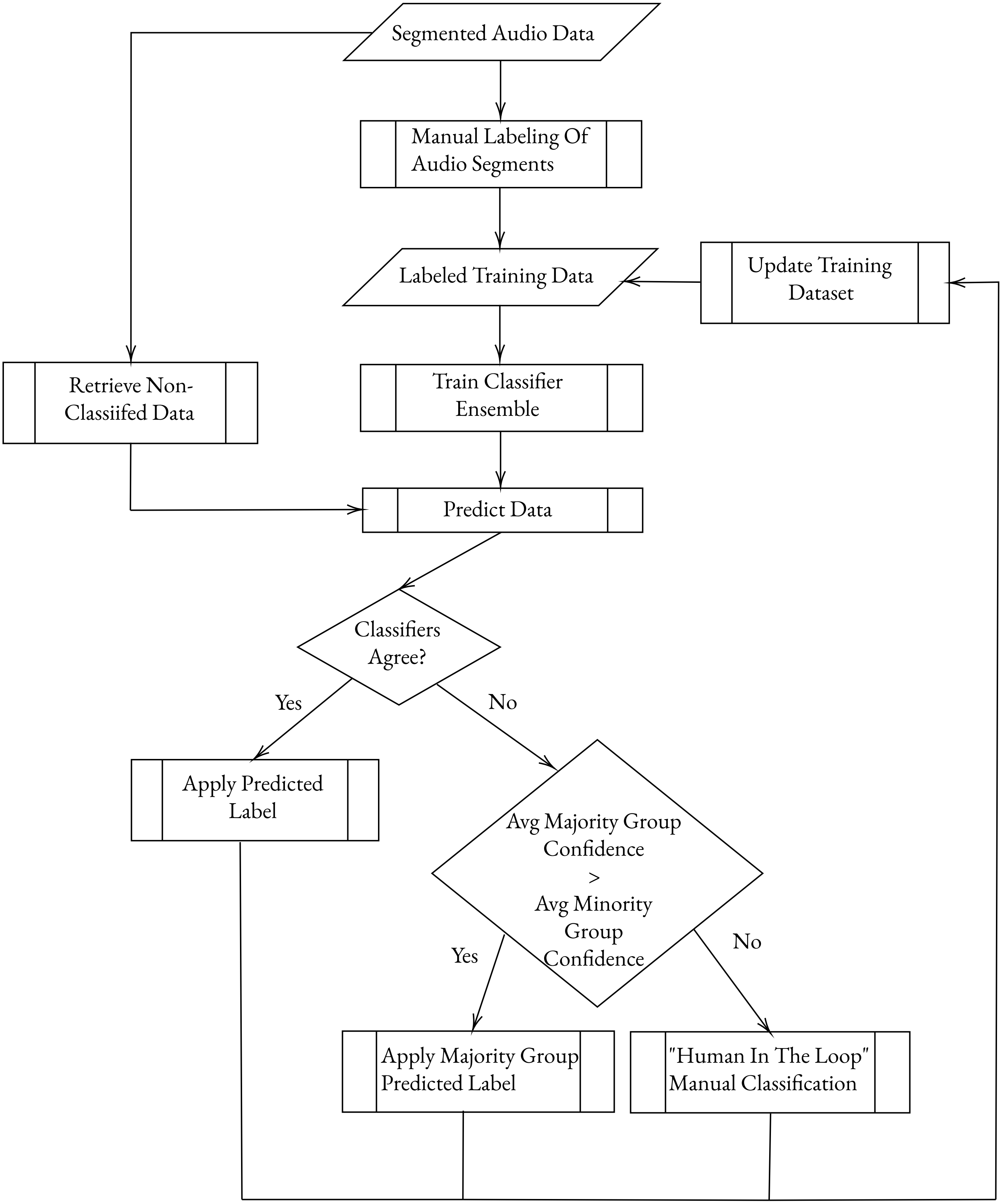

Segmented Audio Data
Manual Labeling Of Audio Segments
Labeled Training Data
Update Training Dataset
Retrieve Non-Classiifed Data
Train Classifier Ensemble
Predict Data
Classifiers Agree?
Yes
No
Apply Predicted Label
Avg Majority Group Confidence > Avg Minority Group Confidence
Yes
No
Apply Majority Group Predicted Label
"Human In The Loop" Manual Classification

Figure 3.4: Modified Democratic Co-Learning process.

The Random Forest ensemble trains individual decision trees that 'vote' on a final classification for some input. 'Gini Impurity' is used to determine when a split should occur.

The confidence prediction of a decision tree is the number of trees in the Random Forest to predict that label. This algorithm is implemented by Pedregosa et al., 2011.

Nu-Support Vector Classifier

The NuSVC is an algorithm in Pedregosa et al., 2011 accessed through Fan et al., 2008. It is an extension of Support Vector Classifiers where the number of support vectors is manipulable via a 'nu' parameter.

Unlike the ensemble classifiers, NuSVC in Pedregosa et al., 2011 uses cross-validation to compute an input's confidence score.

The Support Vector Classifier uses a radial basis function kernel (RBF) to separate the support vector decision space into a non-linear plane.

The nu value used is 0.3, and the tolerance to stopping training is 0.0001.

Ada Boost Ensemble Classifier

The AdaBoost-SAMME algorithm proposed in Hastie et al., 2009 and implemented by Pedregosa et al., 2011 fits a classifier and copies of the classifier to a training set with a weighted focus on challenging data points.

Pedregosa et al., 2011 implements the AdaBoost classifier confidence score by computing the weighted mean predicted class probabilities of the ensemble classifiers.

The AdaBoost base estmiator is the Decision Tree and a total of 100 estimators were used in the ensemble.

§ 3.3.2 Classification

In the heavily imbalanced classification dataset, some rebalancing is done for comparison purposes. As a control point, I chose to oversample classes with less than a set number of data points and undersample the remaining categories to get to that same data count. This sampling method could lead to overfitting to specific bird sounds rather than learning the range of a species's vocalizations.

The final process of this thesis used SMOTE to oversample minority classes and did not perform undersampling. This approach reweights minority labels without destroying potentially helpful training data.

§ 3.4 Complete Pipeline

The simplified yet complete process is demonstrated in figure 3.4.

Although the democratic classifier ensemble was re-trained on the co-learned dataset, all AI-labeled data was kept separate and not mixed with the pre-classified dataset. This is to ensure that the dataset can be verified later on and improve visibility into the process. Additionally, each labeling round's AI classifications were kept separate from earlier rounds to determine the cause of any introduced training error.

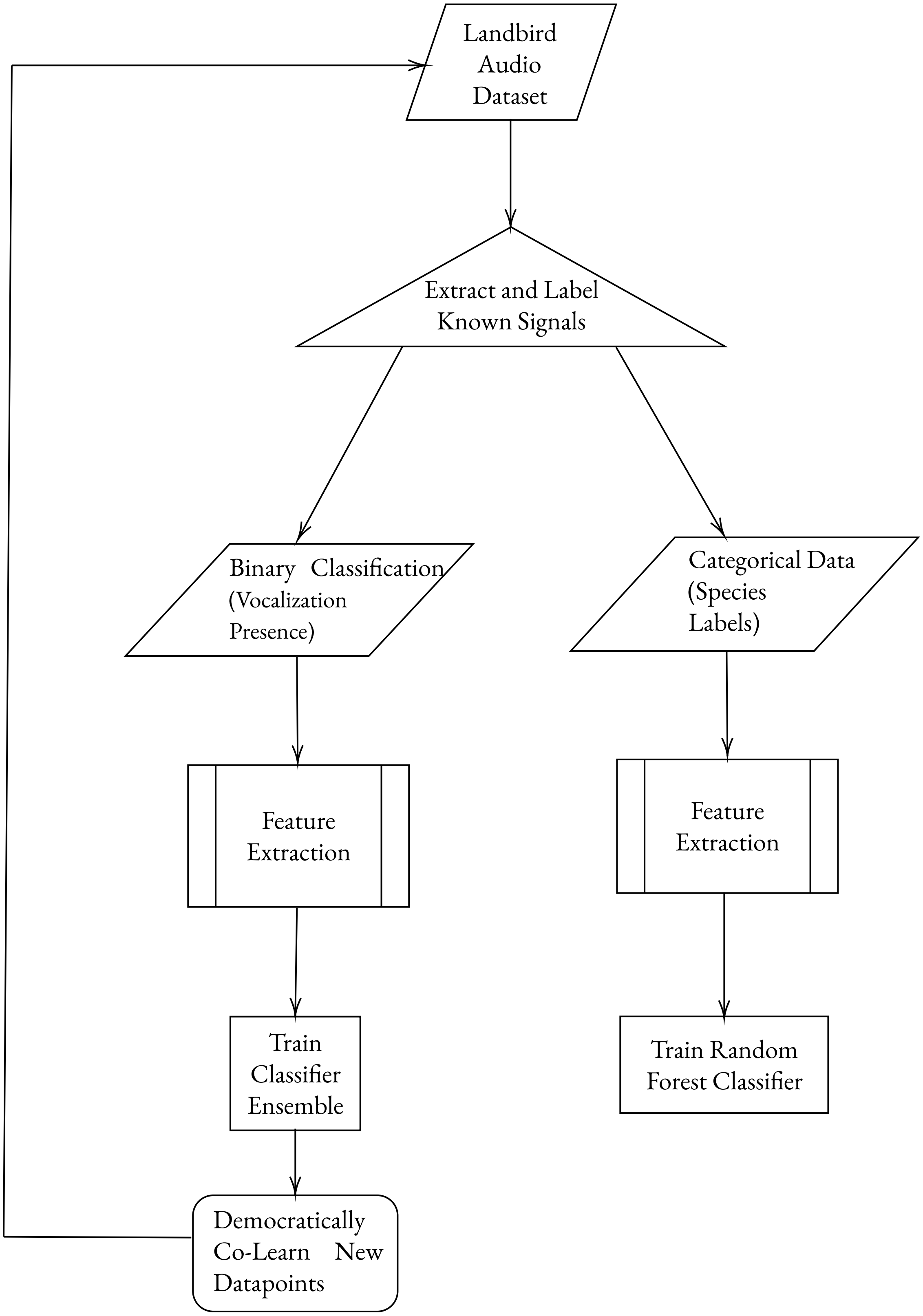

Landbird Audio Dataset
Extract and Label Known Signals
Binary Classification (Vocalization Presence)
Categorical Data (Species Labels)
Feature Extraction
Feature Extraction
Train Classifier Ensemble
Train Random Forest Classifier
Democratically Co-Learn New Datapoints

Figure 3.5: Combined process for training the segmentation and categorical classifier.

CHAPTER 4

RESULTS

§ 4.1 Dataset Processing

There are a total of 82,848 seconds of audio data classified as bird vocalizations present (positive). Unclassified data cannot be assumed to be background noise (or other non-vocalization data) and was therefore manually classified. The manually classified (non-bird-vocalization) data (with a negative label) accounts for 896 seconds.

Manual classification was accomplished by using the co-learning methods described in this thesis's methodology section and manually verified to ensure the dataset's quality.

A total of forty-nine attributes were computed from the discrete wavelet transform summarizing features and 1183 features from DWT-MFCC.

The classification dataset contains a total of 82,848 seconds of audio data spread across 99 labels. The median number of audio data seconds per class is 156.0 seconds, the mean is 836.85 seconds, and the standard deviation is 1724.39 seconds.

Figure 4.2 is an example of a spectrogram representation of a bird vocalization. This signals corresponding Discrete Wavelet Decomposition is in figure 4.3.

§ 4.2 Mother Wavelet Selection

The accuracy comparison for the antisymmetric mother wavelet configurations using random forest is as follows:

From the above information, there is not a clear mother wavelet that outperforms the others. We select Daubechies with a scale of 8 for future processing.

Table 4.1: Multi-level Daubechies wavelet and Haar wavelet accuracy comparison.

	db2	db4	db8	db12	Haar
Accuracy	89.33%	89.33%	90.03%	89.55%	89.32%

§ 4.3 Smoothing

We used audio visualizations and a novice's opinion of the sound transformation to ensure that Soft thresholding and Donoho's universal coefficient were applied correctly. Thresholding does not seem to alter the vocalization portion of a signal. Thresholding does visually and auditorily reduce noise.

The audio visualization in figure 4.1 includes a one-second bird vocalization that is detectable in the original unprocessed audio.

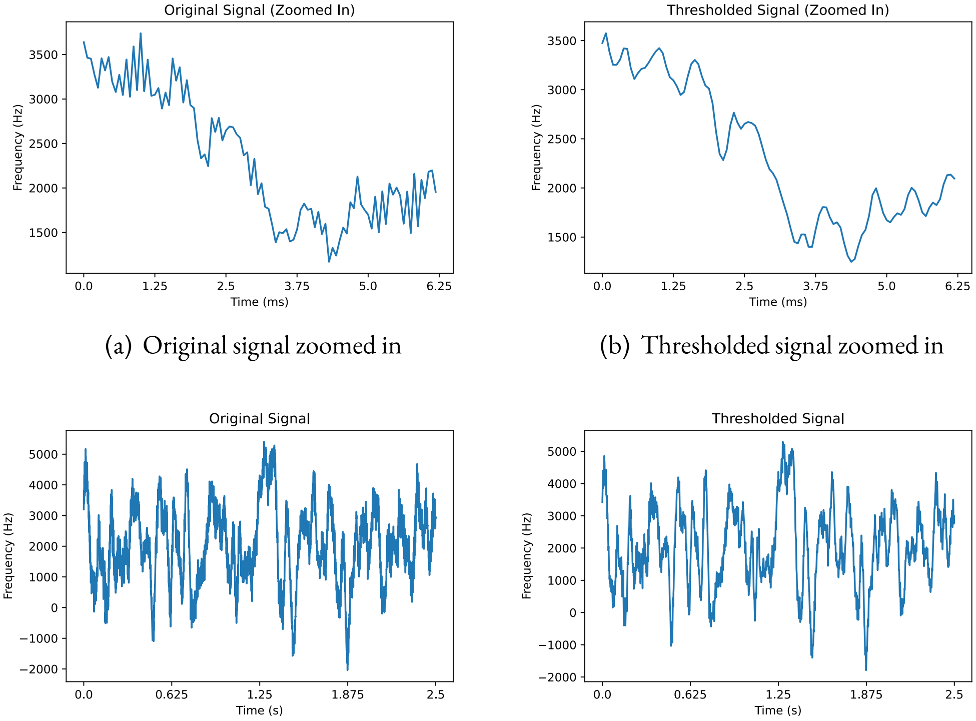
There does appear to be a visual difference in smoothing that is much more prominent in the comparison of the two zoomed-in images in 4.1. This comparison involves soft thresholding of an original signal with Donoho's Universal Coefficient as the thresholding value.

There is also a clear auditory difference between the original sound and the thresholded audio. The bird vocalization can be heard both before and after the thresholding is applied. The audio appears to be quieter in the thresholded audio. There is no conclusive subjective view on which audio clip sounds better and is easier to understand based on the smoothing visuals and audio alone. Soft thresholding reduces a signal closer to the mean according to Donoho, 1995 and will affect the bird vocalization pitch rather than just reducing noise.

§ 4.4 Results of Democratic Co-Learning

The results of training all three classifiers and computing their composite democratic score at various training levels are Figure 4.4.

The democratic co-learning ensemble consistently outperforms Random Forest classifiers, SVM, and Ada-Boost. Support vector machines consistently perform slightly worse than Ada-Boost.



(a) Original signal zoomed in

(b) Thresholded signal zoomed in

(c) Original signal

(d) Thresholded signal

Figure 4.1: Thresholding Comparison

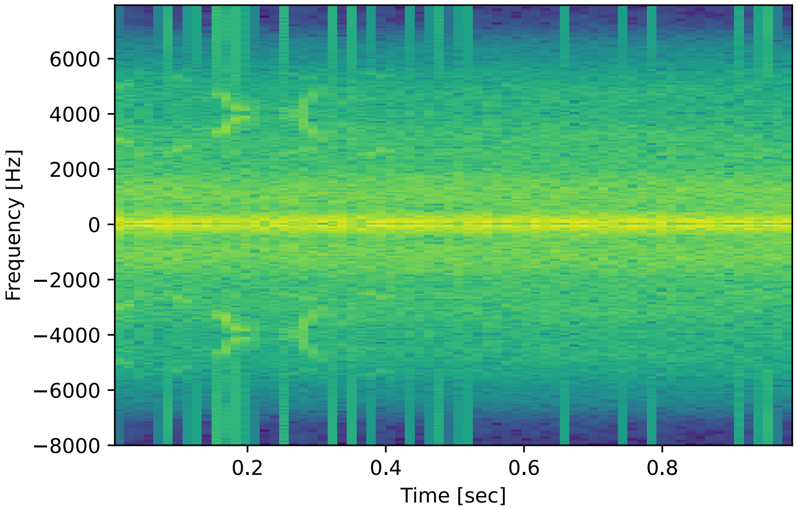


Figure 4.2: Thresholded spectrogram signal data.

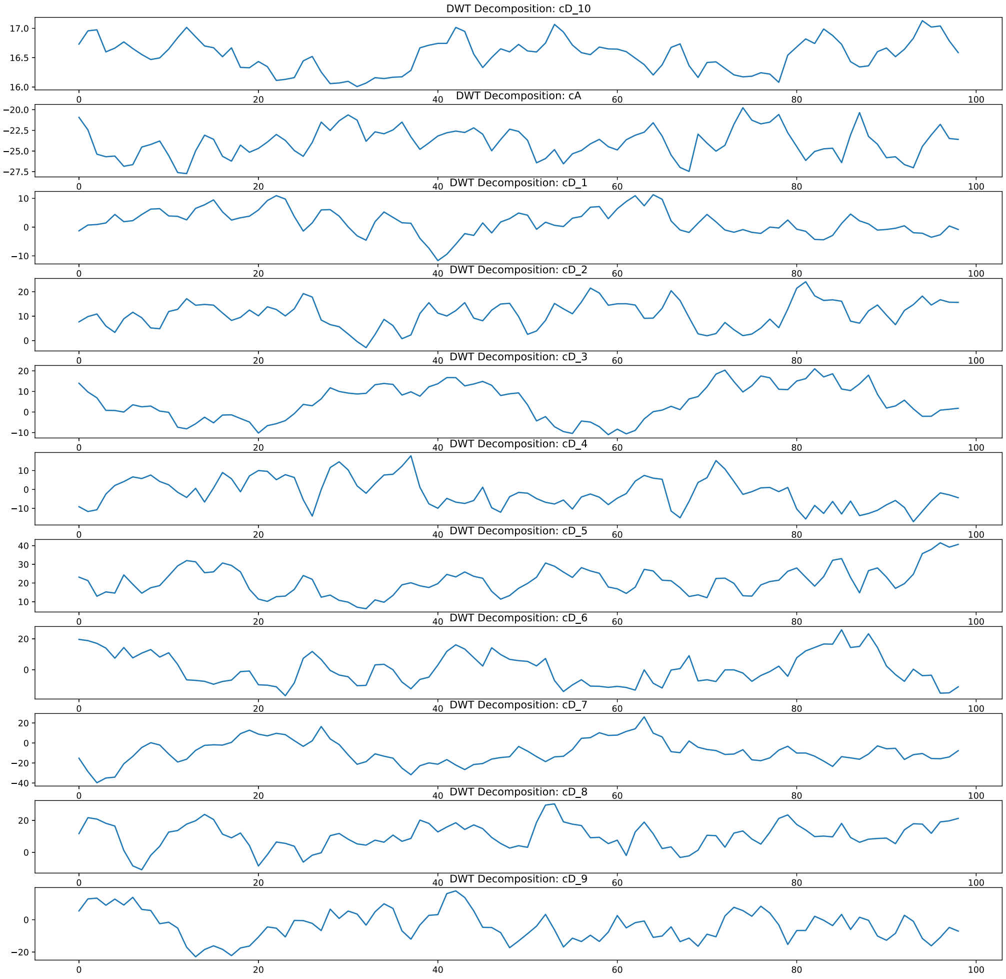


Figure 4.3: Multi-Level DWT

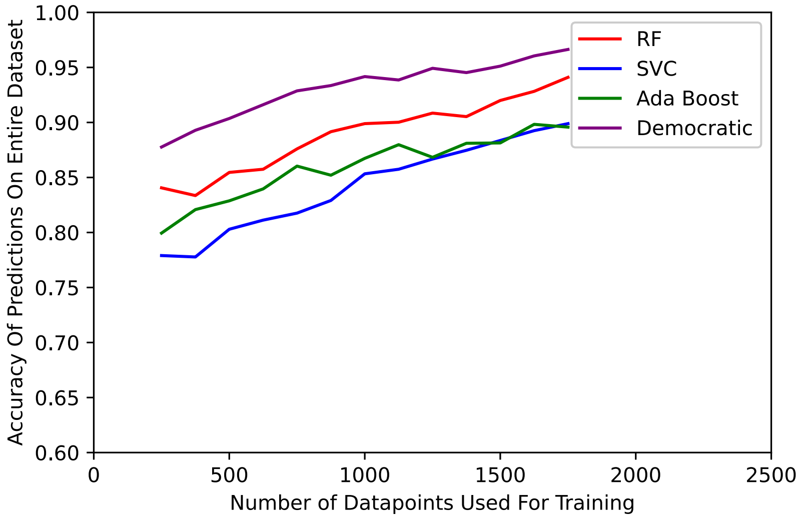


Figure 4.4: DWT-MFCC Feature Extraction accuracy with no feature reduction based on varied training data sets with 80-20 training splits.

While random forest performs poorly on a small amount of training data, the democratic co-learning boosts the overall accuracy significantly. Figure 4.4 does not show the amount of classified data and only offers the predicted data's accuracy.

We sampled a separate 80% training-20% testing set from the segmentation and output the confusion matrix results in Table 4.2. The results show that 9.57% of the testing set was selected to be 'discarded,' which would require manual human classification in this experiment's methodology. False negatives account for 44.44% of the algorithm's error, while the testing set's proportion of negative data is 24.25% of the testing dataset and 33% of the sample overall.

§ 4.5 Feature Engineering

Two feature reduction methods are compared in figure 4.5. The first selection method is relief attribute selection introduced in Kononenko et al., 1997 and implemented in the python library Skrebate at Urbanowicz et al., 2017.

Table 4.2: Democratic co-learning with classifiers trained on an 80/20 split of the dataset on the DWT-MFCC Problem set with no feature reduction nor smoothing. And a 9.57% discard rate.

		prediction outcome			
		p	**n**	**discard**	**total**
actual value	**p′**	547	16	46	609
	n′	20	144	31	195
	total	567	160	77	

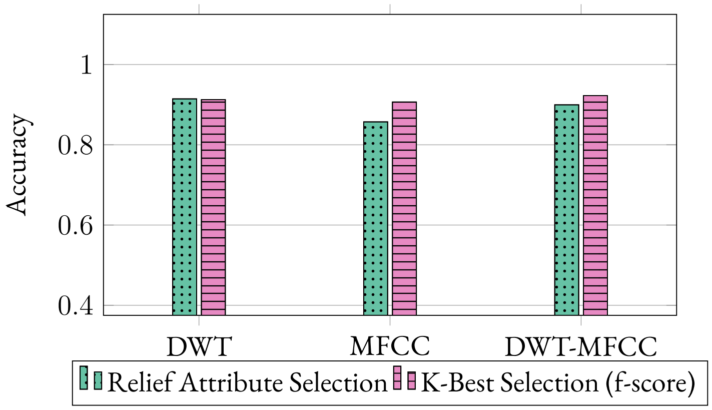


Figure 4.5: Performance comparison of feature reduction methods.

The second feature selection method is the univariate K-best selection with f-score as the metric for computing the attribute score computation and implemented in the Python Library Sci-Kit Learn at Pedregosa et al., 2011. The f-classification algorithm is formulated by Lowry in Lowry, 2014. K-best feature selection ranks each attribute according to its f-score. F-score is a measurement of the certainty that the two classes' means will differ significantly for a given attribute.

There is little difference in accuracy scores. However, k-best runs significantly faster, so the k-best selection method is the feature selection method for all following method portions.

Table 4.3: Segmentation accuracy comparison across each feature set, with feature reduction performed by K-Best Selection (F-Score metric).

Feature	DWT		MFCC		DWT-MFCC	
Count	normal	smoothed	normal	smoothed	normal	smoothed
25 features	0.9125	0.92088	0.9066	0.87126	0.9225	0.911764
100 features	X	X	0.92370	0.89372	0.92763	0.93835
250 features	X	X	0.91601	0.92857	0.94344	0.944
No Reduction (49 / 1287 / 1183)	0.92154	0.909495	0.94670	0.92788	**0.950481**	0.93972

Table 4.3 shows a comparison of the three feature sets at varying feature reduction levels. Each feature type is from a sampling of a single 80-20 dataset split where positive data accounts for 2240.0 data points (71.4% of the sample), and negative data total 896 data points (28.57% of the samples).

The feature reduction method employed in Table 4.3 is K-Best Selection with F-score as the metric for determining each feature's value. Each feature extraction method produces a unique number of features. The "No Reduction" row corresponds to each extraction method's complete feature set, where the feature amounts are enclosed in parentheses and follow the ordering of the feature extraction header list. This k-best selection with an f-score metric (ANOVA) is of each feature reduction is formulated in Lowry, 2014.

The DWT-MFCC set of features with no feature reduction outperforms all other feature sets, with the MFC feature set performing similarly.

The poorest performing feature set is MFCC, with soft-thresholding and no feature reduction.

Table 4.3 shows that few thresholding cases improve accuracy from their parallel non-smoothed version.

§ 4.6 Class Distribution

A Random Forest Classifier trained on 80-20 splits of the complete classification dataset at various class selection levels and various under/oversampling methods. The results are shown in figure 4.4. The first three listed methods in 4.4 are subsets of the overall segmentation dataset. For example, the first listed

Table 4.4: A Random Forest comparison of class selection and prediction accuracy. The dataset used is DWT-MFCC with no feature reduction.

Class Selection	Accuracy
over 5000 datapoints (4 classes)	**0.89138**
over 1000 datapoints (19 classes)	0.846125
over 500 datapoints (33 classes)	0.8407989
all classes (99 classes)	0.832106
all class (99 classes with random oversampling/undersampling to 500 datapoints)	0.618260
all classes (with SMOTE)	**0.84369**

subset only includes classes with over 5,000 seconds of data associated with a label (totaling four labels/bird species).

The highest performing class subset is "all classes with over 5,000 data points," which amounts to four classes, and 30,104 seconds of audio data. The worst performing class selection is all 99 classes with oversampling and undersampling to reach 500 seconds of audio per class type. (No other class selection method involved random oversampling.)

Of the three class-selection methods that only consider overall class sizes, the best performing group uses only four classes. This group (over 5000 data points) performs about 6% better than the full 99 class selection option.

The training group created with Synthetic Minority Over-sampling Technique (SMOTE) does lead to slight improvements over the entire 99 class dataset with no balancing techniques.

§ 4.7 Multi-Class Species Classification

Next, I tested the various feature selection options on the species classification dataset, where an f-score metric determines feature reduction subsets. This experiment's results are shown in table 4.5.

Table 4.5 follow the same structure as table 4.3.

The top-performing classification configuration appears to be DWT-MFCC with soft-thresholding and no feature reduction. The worst performing selection is MFCC, with thresholding and a feature space reduced to 25 features. Soft thresholding consistently improves the accuracy of the DWT-MFCC feature space but has little effect otherwise.

Table 4.5: Categorical classification accuracies with feature reduction performed by K-Best Selection (F-Score metric).

Feature	DWT		MFCC		DWT-MFCC	
Count	normal	smoothed	normal	smoothed	normal	smoothed
25 features	0.97694	0.97754	0.80265	0.795654	0.841762	0.9809897
50 features	X	X	0.81309	0.795051	0.846288	0.982679
100 features	X	X	0.81050	0.79969	0.847001267	0.982619
No Reduction (49 / 1287 / 1183)	0.98002	0.979541	0.82341	0.81593	0.84369	**0.983343**

CHAPTER 5

DISCUSSION AND CONCLUSION

§ 5.1 Discussion

The explored feature sets allow for separating bird vocalizations from non-vocalization data, especially when used as input for a Random Forest classifier. The proposed algorithm demonstrates high accuracy on the SECN Landbird Dataset but is not a replacement for the wildlife team's standard operating procedure described in the background section.

There is evidence that the accuracy of binary classification improves with more binary data. The addition of more negatively classified segmentation data points would improve accuracy and lead to more minor manual classification being necessary. These segmentation results indicate that the Democratic ensemble can label unseen data points with minimal human input.

The democratic classifier causes a small amount of mislabeling error, shown in figure 4.2. This data's ideal labeling would be unknown in a production environment, so it would likely end in the training dataset. Outlier detection methods could scan the training dataset later for outliers, but this does come with the fault of having some misclassified data in the dataset.

Democratic co-learning does seem to improve audio data segmentation both in accuracy and its ability to 'set aside' data for manual user classification. The classifier ensemble's accuracy is seemingly enhanced by the variety of classifiers, as the testing accuracy is consistently above the top classifier (usually random forest). The second benefit of democratic co-learning is highlighting problematic data, which could alternatively be done using the probability estimate of any classifiers independently. However, the combination of probability estimates helps to identify further data that can be successfully classified while minimizing both error and the amount of 'discarded' data.

The machine learning pipeline described in this paper did not need the benefits of democratic ensemble classification for the second portion of the text (classification) because Random Forest, an ensemble technique, performed accurately without additional classification votes. Also, co-learning was not needed since the dataset is extensive and the cost of labeling more audio data was too high for this current project.

Across the Democratic ensemble, random forest consistently performs the best on this segmentation training set, while SVC performs the worst. The SVC's 'RBF' kernel should have handled the dataset's non-linearity. However, the inconsistencies caused by vocalizations occurring at varying time frames could have led to a need for relations between features. One option for moving forward would be to generate linearly separable features or use a different classifier.

The K-best selection feature reduction method outperforms Relief Attribute selection but reduces the segmentation classifiers and categorical classifier's accuracy. The K-best selection method with f-score as the attribute valuation method likely performs well on all three feature sets because each feature set comprises independent variables, a requirement of F-Score use. Each feature set relies on summarizing features rather than coefficient values themselves, allowing for variables' independence. Accuracy scores reduce with feature selection, which is explainable by a reduction of valuable features. K-best selection seemingly removes features that improve understanding of the problem space, and further investigation could result in an ideal feature selection count, but this would also cause the time complexity of the overall process to increase.

Results indicate that smoothing leads to better performance in multi-class vocalization classification while not improving the segmentation problem. Donoho's coefficient likely requires more fine-tuning and could benefit from expert knowledge on the range of bird songs and the expected range of background noise. Further investigation into the settings that the ARDs record in could also indicate what modifications could naturally reduce noise. It is also possible that too much information is removed from the signal data within the range of bird vocalizations. The bird vocalizations could be overpowered by other noises, such as car horns, and the bird vocalization could be overly reduced.

§ 5.1.1 Research Application

The process proposed helps create a tool that can section out bird noise data from other noise sources and could be used in an application by training a single classifier with high accuracy and iteratively 'scanning' data. Because the application only parses one second of noise data at a time, there is a limitation in algorithm applicability. We could overcome this by converting all signal data to DWT-MFCC feature space at once and immediately using the learned feature selection space. All preset and pre-learned, so the computation time for reuse is drastically lower than the original time costs. The process would then parse the feature set into set intervals corresponding to a single segment and classified from this final state.

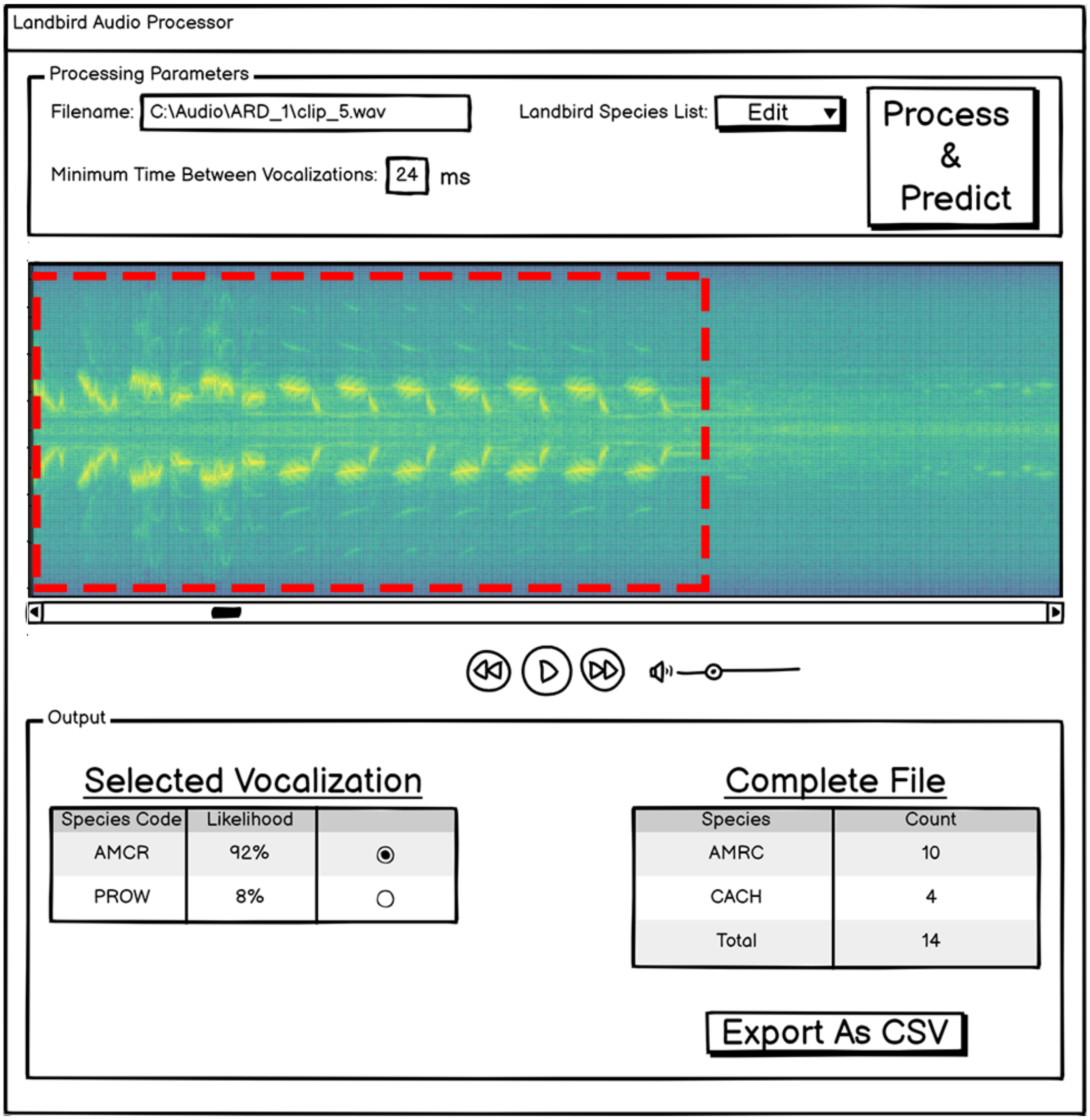
An example user interface that would pair well with the created classifiers is shown in figure 5.1. This application would begin by framing a singular audio file into second-long segments. Next, the binary classifier assigns a score to each of these second long segments and predicts whether a vocalization is present. Lastly, the application assigns a species prediction to each of the discrete vocalizations.

A general overview of how this research could fit into the NPS SECN Wildlife process is shown in figure 5.2. This process would allow for automatic data collection and gradual improvements in algorithm accuracies. The classifiers other than random forest seem to perform worse consistently, so once the random forest begins to consistently produce high accuracy, we could remove the other democratic classifiers.

§ 5.1.2 Future Research

Further research would be to work on a smoothing method that is adaptive in the same way that the feature extraction methods are. A more adaptive smoothing method could improve over time and could use some meta value that determines how much smoothing to apply.

One significant improvement for future research would be to better use the entirety of the unlabeled portions of the dataset. Much of this unlabeled data is simply background noise, but even this data could lead to an improved segmentation algorithm. Currently, some manual classification is required, but other unsupervised methods of learning from this unlabeled data could be used to make better use of it.



Filename: C:\Audio\ARD_1\clip_5.wav

Landbird Species List: Edit ▾

Minimum Time Between Vocalizations: 24 ms

Selected Vocalization

Species Code	Likelihood	
AMCR	92%	●
PROW	8%	○

Complete File

Species	Count
AMRC	10
CACH	4
Total	14

Figure 5.1: The intended final product wireframe that is not yet implemented, but could be based on this project's produced classifiers.

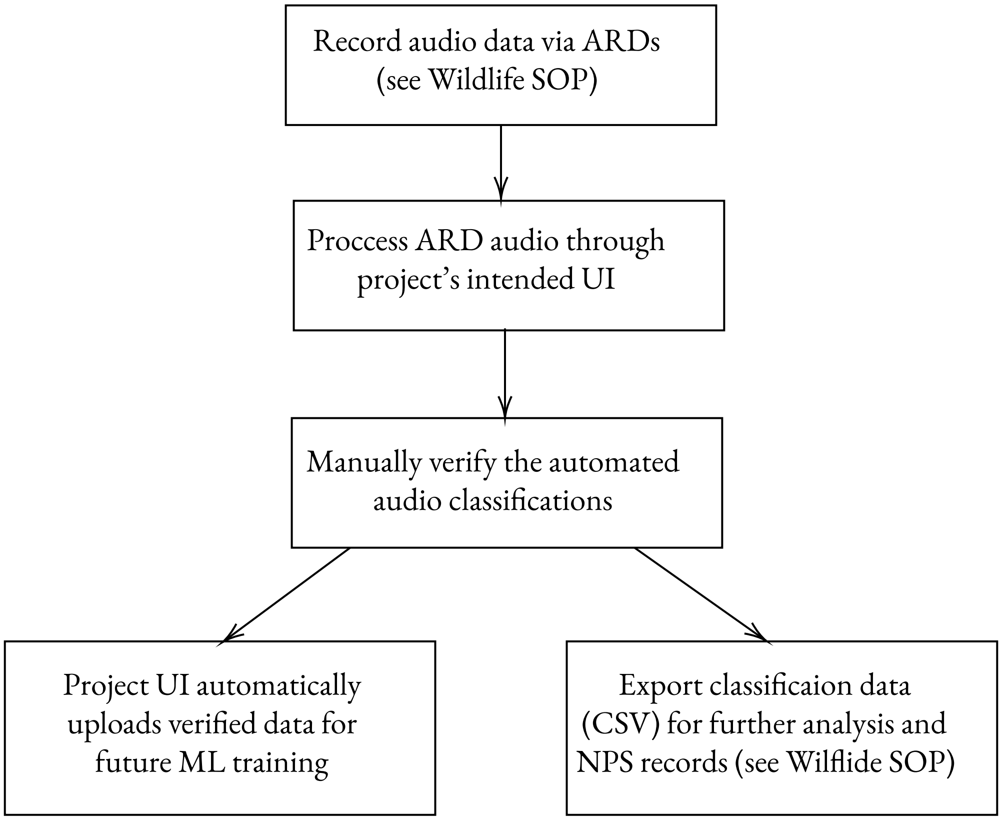

Record audio data via ARDs (see Wildlife SOP)
Proccess ARD audio through project's intended UI
Manually verify the automated audio classifications
Project UI automatically uploads verified data for future ML training
Export classificaion data (CSV) for further analysis and NPS records (see Wilflide SOP)

Figure 5.2: An integrated workflow adoptable by the NPS SECN Wilflide team.

Lastly, Democratic co-learning could increase the sample size of imbalanced bird classes and identify specific bird species from the larger audio dataset. This process requires expert knowledge to ensure that a selected audio clip fits a particular label but could lead to the rapid acquisition of potential data points for a given label.

BIBLIOGRAPHY

Abdalla, M. I., & Ali, H. S. (2010). Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models. *arXiv preprint arXiv:1003.5627*.

Aggarwal, R., Singh, J. K., Gupta, V. K., Rathore, S., Tiwari, M., & Khare, A. (2011). Noise reduction of speech signal using wavelet transform with modified universal threshold. *International Journal of Computer Applications*, *20*(5), 14–19.

Ahmad, J., Fiaz, M., Kwon, S.-i., Sodanil, M., Vo, B., & Baik, S. W. (2016). Gender identification using mfcc for telephone applications - a comparative study.

Bachu, R., Kopparthi, S., Adapa, B., & Barkana, B. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal, In *American society for engineering education (asee) zone conference proceedings*.

Bergland, G. D. (1969). A guided tour of the fast fourier transform. *IEEE Spectrum*, *6*(7), 41–52. https://doi.org/10.1109/MSPEC.1969.5213896

Bhagat, R. C., & Patil, S. S. (2015). Enhanced smote algorithm for classification of imbalanced big-data using random forest, In *2015 ieee international advance computing conference (iacc)*. https://doi.org/10.1109/IADCC.2015.7154739

Bhakre, S. K., & Bang, A. (2016). Emotion recognition on the basis of audio signal using naive bayes classifier, In *2016 international conference on advances in computing, communications and informatics (icacci)*. https://doi.org/10.1109/ICACCI.2016.7732408

Briggs, F., Huang, Y., Raich, R., Eftaxias, K., Lei, Z., Cukierski, W., Hadley, S. F., Hadley, A., Betts, M., Fern, X. Z., Irvine, J., Neal, L., Thomas, A., Fodor, G., Tsoumakas, G., Ng, H. W., Nguyen, T. N. T., Huttunen, H., Ruusuvuori, P., … Milakov, M. (2013). The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment, In *2013 ieee international workshop on machine learning for signal processing (mlsp)*. https://doi.org/10.1109/MLSP.2013.6661934

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, *41*(3), 613–627. https://doi.org/10.1109/18.382009

Drummond, C., & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *the Journal of machine Learning research*, *9*, 1871–1874.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), 349–360.

Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, *7*(1), 39–55.

Lowry, R. (2014). Concepts and applications of inferential statistics.

Lu, L., Jiang, H., & Zhang, H. (2001). A robust audio classification and segmentation method, In *Proceedings of the ninth acm international conference on multimedia*, Ottawa, Canada, Association for Computing Machinery. https://doi.org/10.1145/500141.500173

Lu, X., Tsao, Y., Matsuda, S., & Hori, C. (2013). Speech enhancement based on deep denoising autoencoder., In *Interspeech*.

Lyons, J., Wang, D. Y.-B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., Birch, S., Zhihe, L., Hölzl, J., Lesinskis, J., Almér, H., Lord, C., & Stark, A. (2020). *Jameslyons/python_speech_features: Release v0.6.1* (Version 0.6.1). Zenodo. https://doi.org/10.5281/zenodo.3607820

MW, B., JC, D., CJ, W., E, T., & CD., J. (2014). *Protocol for monitoring landbird communities in southeast coast network parks. natural resource report. nps/secn/nrr—2014/853.* (tech. rep.). National Park Service. Fort Collins, Colorado.

Narasimhan, R., Fern, X. Z., & Raich, R. (2017). Simultaneous segmentation and classification of bird song using cnn, In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. https://doi.org/10.1109/ICASSP.2017.7952135

O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE Potentials*, *7*(1), 29–32. https://doi.org/10.1109/45.1890

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks, In *2015 ieee 25th international workshop on machine learning for signal processing (mlsp)*. https://doi.org/10.1109/MLSP.2015.7324337

Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020. https://doi.org/https://doi.org/10.1016/j.apacoust.2019.107020

Tin Kam Ho. (1995). Random decision forests, In *Proceedings of 3rd international conference on document analysis and recognition*. https://doi.org/10.1109/ICDAR.1995.598994

Tufekci, Z., & Gowdy, J. N. (2000). Feature extraction using discrete wavelet transform for speech recognition, In *Proceedings of the ieee southeastcon 2000. 'preparing for the new millennium' (cat. no.00ch37105)*. https://doi.org/10.1109/SECON.2000.845444

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2017). Benchmarking relief-based feature selection methods.

Vonesch, C., Blu, T., & Unser, M. (2007). Generalized Daubechies wavelet families. *IEEE Transactions on Signal Processing*, *55*(9), 4415–4429.

Wai Keng, N., Leong, M., Hee, L., & Abdelrhman, A. (2013). Wavelet analysis: Mother wavelet selection methods. *Applied Mechanics and Materials*, *393 (2013)*, 953–958. https://doi.org/10.4028/www.scientific.net/AMM.393.953

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137

Xu, Y., Zang, C., & Yang, J. (2005). Semi-supervised classification of musical genre using multi-view features, In *Icmc*. Citeseer.

Yadav, V. K., Jain, A., & Bhargav, L. (2015). Analysis and comparison of audio compression using discrete wavelet transform. *International Journal of Advanced Research in Computer and Communication Engineering*, 310–313.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets (T. Herawan, M. M. Deris, & J. Abawajy, Eds.). In T. Herawan, M. M. Deris, & J. Abawajy (Eds.), *Proceedings of the first international conference on advanced data and information engineering (daeng-2013)*, Singapore, Springer Singapore.

Zhou, Y., & Goldman, S. (2004). Democratic co-learning, In *16th ieee international conference on tools with artificial intelligence*. https://doi.org/10.1109/ICTAI.2004.48