A System for Collection and Machine Learning Analysis of Aquatic & Semi Aquatic Ecosystem Acoustic Data in North Carolina

Authors: Keenan Powers, Stewart Hammond, Victor Xia, Ishani Raha, Ryan DeVries, Jack Rhodes, Angela Torrejon

Abstract

In this project, we are delving into the realm of aquatic and semi-aquatic bioacoustics by using machine learning to classify different species based on their unique sounds within the North Carolina area. The goal lies in assembling a rich and varied dataset that encompasses a wide array of species vocalizations, such as koi fish and frogs, for accuracy in a vast machine-learning model. To transform these raw audio snippets into meaningful data for our machine-learning model, we are employing techniques such as spectrogram analysis and the identification of key acoustic parameters. Through a thoughtful selection of neural network architecture, we are training our model to discern the auditory fingerprints of each species. This project's impact is substantial; it could pave the way for a transformative tool in underwater monitoring and biodiversity research, offering an automated and streamlined method for identifying fish species through their distinctive acoustic expressions.

Online Google Drive folder containing project materials ("F23 Audio") is linked here.

Table of Contents

Title Page **0**
Abstract & Link to Online Project Folder **1**
Table of Contents **2**
I. Introduction **3**
A. Environmental Background 3
B. Machine Learning Background 4
III. Methods & Procedure **6**
A. HydroMoth 6
B. Piezo Sound Creation 7
C. Advanced Sound Creation 9
D. Mounting HydroMoth in Aquatic Environments 10
E. Machine Learning Audio Analysis - Traditional Algorithms 12
F. Advanced Machine Learning - Recurrent Neural Network 13
G. Dataset 14
IV. Results & Discussion **15**
A. Accuracy of Machine Learning Algorithms 15
B. Analysis of Data & Algorithm Results 17
C. Recurrent Neural Networks Performance 21
D. Limitations of the Hydromoth Recording System 23
VI) Conclusions & Future Considerations **24**
References **26**

I. Introduction

A. Environmental Background

In recent years, the integration of machine learning techniques, particularly neural networks, has revolutionized the field of wildlife monitoring and conservation. This project seeks to explore the application of machine learning neural networks in the detection of semi-aquatic and aquatic wildlife, assisting in monitoring species in aquatic ecosystems. As climate change and habitat degradation continue to threaten biodiversity, the development of efficient and accurate wildlife detection methods becomes crucial to ecosystem preservation. Our approach aims to contribute to the advancement of wildlife conservation efforts by using machine learning to provide an automated solution for identifying various frog (*Anura*) species in their natural habitats, such as the North Carolinian American east coast ecosystem found in the Sarah P. Duke Gardens. This project utilized customized, optimized machine learning code, such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectral contrast methods.

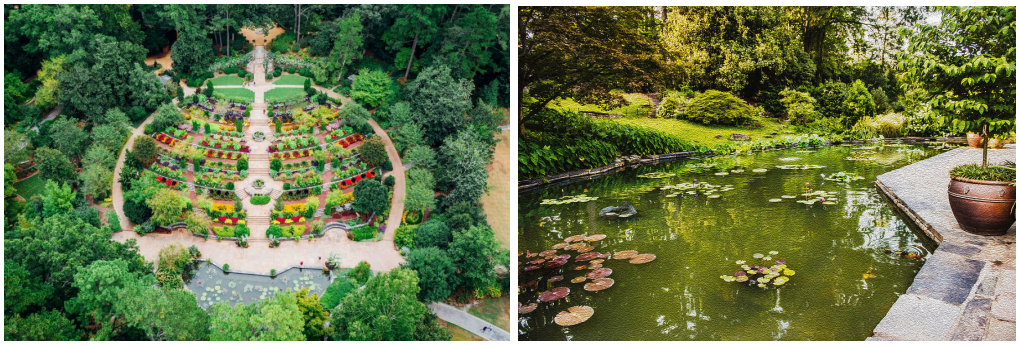


Figure 1: *a) The Sarah P Duke Gardens (left) and b) Koi Pond, located centrally within the garden proper (right).*

The goal of the project is to expand our understanding of the acoustic patterns and behaviors of North Carolina fauna. Due to geographic restraints and locality, we concentrated our efforts on the Sarah P. Duke Gardens (Figure 1a). The lack of comprehensive sound recording and classification techniques specifically tailored to the Sarah P. Duke Gardens negatively affects our ability to monitor and understand ways in which its ecosystem compares to ecosystems in and beyond North Carolina. Understanding these similarities and differences could offer insight into the optimal conditions for different frog species to thrive. The artificial changes in a relatively urban environment like the Sarah P. Duke Gardens may have implications for

natural selection within the frog species as the native species have not evolved for such an unusual environment. However, the absence of invasive frog species in the Gardens due to constant upkeep and cultivation may mitigate some of these effects.

Figure 2: *The Frog Sanctuary in the Sarah P. Gardens*

Audio machine learning algorithms are revolutionizing researchers' and conservationists' efforts to understand and safeguard ecosystem health. These algorithms can aid in the identification and monitoring of species through their unique sounds or calls, which aids in biodiversity assessment by continuously monitoring species presence, distribution, and behavior. By analyzing audio data, scientists can detect changes in species composition, providing insights into ecosystem dynamics (Sethi et. al, 2020). Such alterations may signal disturbances, climate shifts, or habitat degradation, crucial indicators of an ecosystem's health. Moreover, these algorithms facilitate early detection of invasive species, allowing for swift intervention to mitigate their potentially devastating impacts on native flora and fauna.

Such efforts can also help to identify endangered species and their habitats, supporting targeted conservation strategies. Through non-invasive acoustic monitoring, scientists have insight into wildlife behavior without causing stress or disturbance to the species being observed. This continuous monitoring offers a comprehensive understanding of the natural rhythms and patterns within ecosystems, aiding in the assessment of habitat health and the effects of human activities on biodiversity (Sethi et. al, 2020).

Overall, through the utilization of neural networks, this project seeks to delve into the potential of machine learning technology to enhance the precision, scalability, and real-time capabilities of wildlife monitoring systems, ultimately aiding in the sustainable management and protection of aquatic ecosystems worldwide in an ideally up-scalable process.

B. Machine Learning Background

Machine Learning is integral to this project, enabling users to obtain audio samples from animals and classify them according to their spectrogram–a visual representation of the frequencies within an audio signal over time. To make a productive contribution to bioacoustics, the classification system should produce results of 80% accuracy or greater in determining the species in given test audios. The first step in creating a model for the audio classification of animals and fish is the collection of labeled audio datasets. These datasets consist of recordings containing the vocalizations of various species, annotated with information about the species and context. The larger and more diverse the dataset, the better the model can learn to generalize across different species and environmental conditions. Once the datasets are curated, feature extraction is employed to convert the raw audio data into a format suitable for machine learning algorithms. Spectrogram analysis, which visualizes the frequency content of the audio signal over time, is a common technique. Other features such as pitch, duration, and intensity may also be extracted to comprehensively represent the vocalizations. Although in-person classification and analysis are possible when only minute differences are present in the data, it can be difficult to accurately classify based on species, and automated systems provide more efficient means for collecting useful data.

The process through which machine learning algorithms classify data is extensive. Initially, diverse and labeled datasets of animal sounds are collected, followed by preprocessing steps to ready the data. When many audio recordings are uploaded, it becomes possible for the machine to more accurately determine the identification of an animal. Next, feature extraction, involving identifying relevant acoustic features, is performed. This is done by reducing noise within the recording and amplifying specific frequency ranges within the sounds, as shown in Figure 3.

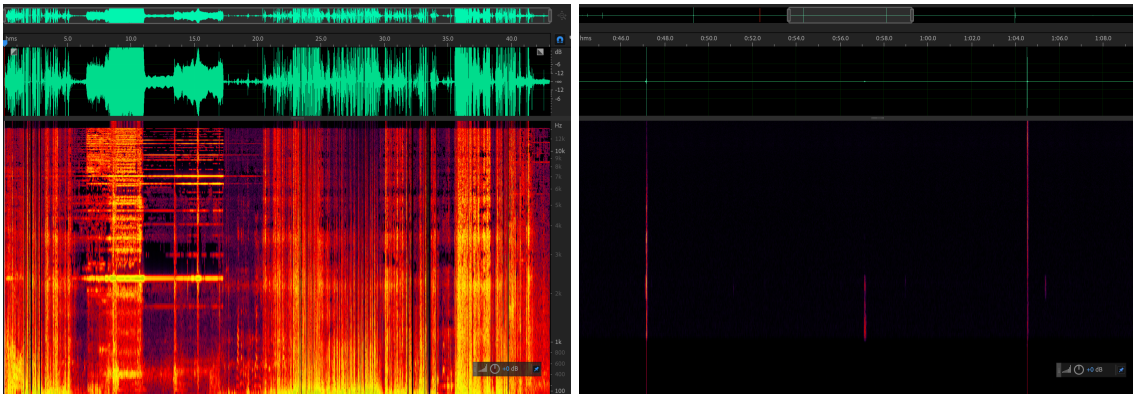


Figure 3. *a) Unadulterated Koi Spectrogram (left), b) Denoised and Amplified Koi Spectrogram (right)*

Then, the selected machine learning model, often based on neural networks, is trained on this data, adjusting internal parameters specific to each model process, such as certain frequency ranges, to minimize prediction errors. After evaluation on validation and test sets, the model is deployed for real-world use, classifying new audio data based on learned patterns. This process aids researchers in monitoring and conserving wildlife, providing valuable insights into biodiversity and behavior through automated bioacoustic analysis.

III. Methods & Procedure

A. HydroMoth

To capture the audio data needed to analyze the man-made environments in the Durham area, a recording device that can withstand the natural conditions of the spaces is needed. Created by the company Open Acoustic Devices, the HydroMoth is a versatile and open-source acoustic monitoring device designed for capturing wildlife sounds, particularly in the context of aquatic or semi-aquatic ecological or environmental research. The device is a low-cost, energy-efficient solution that allows scientists, conservationists, and other acoustic researchers to record sounds of natural environments with high quality (Open Acoustic Devices, n.d.). The Hydromoth provides a comprehensive solution for underwater audio recording with features tailored for marine and aquatic applications with the ability to record sound down to sixty meters underwater for short periods or longer periods, up to two months, at thirty meters (Open Acoustic Devices, n.d.). The HydroMoth uses a software application package that can be installed through the Open Acoustic Device's website. When the board is connected to the

computer using a USB to Micro USB cable, the device can be configured according to several properties including sampling frequency, gain, sleep, and cyclic recording using more broad scheduling of date(s) and 24-hour cycles.

The Hydromoth's underwater data collection abilities were initially tested in 3 locations. After learning to operate the Hydromoth, the device was taken outside to the Duke Pond, shown in Figure 4. The device was lowered into the water with a strap to submerge 1 meter deep. In this test, a turtle, fish, and several ducks moved toward the device. The sound recording was trimmed and the audio was leveled in an attempt to observe any distinct animal sounds, though none were apparent.

To obtain more sound recordings, the Hydromoth was taken to the Duke Gardens. The device was turned on and lowered into the Koi Pond, as shown in Figure 4. Several koi fish swam near the device during the 60 seconds of submersion, and the sound recording of the event contained two low-pitched croaks, a distinct sound of koi fish.

Figures 4. *Hydromoth underwater data collection at Duke Pond and Gardens Koi Pond*

The HydroMoth was then placed in a small, natural pond in the Duke Gardens, as shown in Figure 4. The water was obscured by moss, so it is unclear what animals would make sounds for the Hydromoth, though it was hoped that this would produce a larger diversity of collected sounds. This sound collection trial did not result in any discernible animal sounds to the human ear. To get more consistent underwater sound testing data from the Hydromoth, an underwater speaker was needed. This addition would allow for a broader sample collection as sounds from the internet could be played underwater, and an increased number of Hydromoth recorded samples, as more samples were needed for training a Machine Learning model. This was initially attempted with a Bluetooth waterproof speaker, though it was found that the Bluetooth signal

would be lost underwater, the speaker could not be submerged past 1 meter, and the speaker's audio output was distorted underwater as it was intended as an above-water speaker. Instead, a piezoelectric speaker system was created to test underwater sound production.

B. Piezo Sound Creation

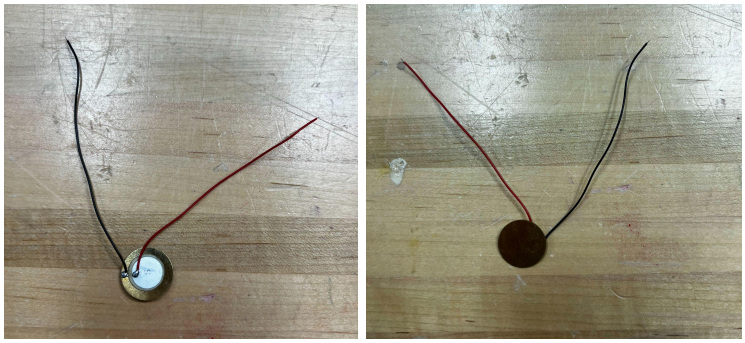


Figure 5. *The piezoelectric device, a) front and b) back.*

Compared to other speaker designs, piezoelectric speakers offer relative ease of operation. One key advantage of piezoelectric speakers lies in their resistance to overloads that could damage conventional high-frequency drivers. In underwater sound creation, piezoelectric variants serve as both output devices, generating sound underwater, and input devices, acting as components of underwater microphones. Their solid-state construction and resistance to seawater make them a practical choice in these demanding environments, outperforming traditional ribbon or cone-based devices (Hughes, 2016).

A single piezoelectric was attempted first above water. This was done by connecting the piezo speaker to an audio amplifier, which received input via an auxiliary cable connected to a laptop. This setup is shown in Figure 6, though with two piezo speakers connected.

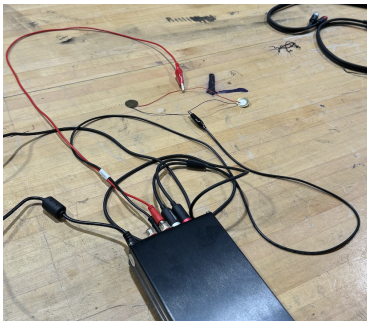


Figure 6. *Piezoelectric speakers connected to an amplifier and laptop*

This initial setup produced a very quiet sound traveling through air, so the next step was to attempt to wire two piezo speakers in parallel, to understand the impact of using more speakers on the overall output sound. The use of two speakers did produce a louder sound, though the metal and ceramic contacts of the two speakers were not protected from one another. Upon the two speakers physically touching while both connected in parallel, the circuit shorted and ignited a small flame. While using multiple speakers with more care is a viable way of amplifying the sound, this implementation would require more power. As shown in Figure 7, ten piezoelectric speakers would be needed to double the sound.

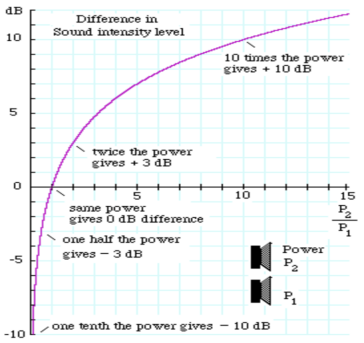


Figure 7. *Graph demonstrating the nonlinear reception of sound*

Moving forward with just one piezoelectric speaker, the next step was testing the speaker underwater. This was necessary to ensure that the speaker would work underwater and the vibrational sound waves from the speaker could be clearly picked up by the Hydromoth. It was found that the piezoelectric speaker operated properly underwater, without concern of electrical components shorting or loss of connection to the input sound (which was an issue with Bluetooth speakers underwater). The Hydromoth was placed into the bucket with the speaker, and the sound output by the speaker was successfully recorded by the Hydromoth. This proved as a viable method of recording underwater sound samples to build out a robust, accurate sample database for later comparison and machine learning training.

C. Advanced Sound Creation

To further improve upon the piezoelectric underwater speaker, a DAEX25W-8 Waterproof 25mm Exciter 10W 8 Ohm speaker was used. The speaker's IP67-rated waterproof enclosure allows for up to 1-meter deep submersion. The speaker also includes higher decibel ranges for sound output than the piezoelectric speaker, and outputs sound from a range of 70 Hz

to 20 kHz (Dayton Audio, 2023). This speaker has peak decibel levels of around 3kHz. This is ideal given fish and frogs typically reach up to the 3-4 kHz range (LSU, 2017).

The DAEX25W-8 speaker was tested with the same amplifier as the piezoelectric speaker, as shown in Figure 8. As observed by the team, this configuration produced a much louder sound out of water compared to the piezo speaker. This speaker could be heard without physical contact or closer proximity (<6 inches) of the speaker with the listener's ears.

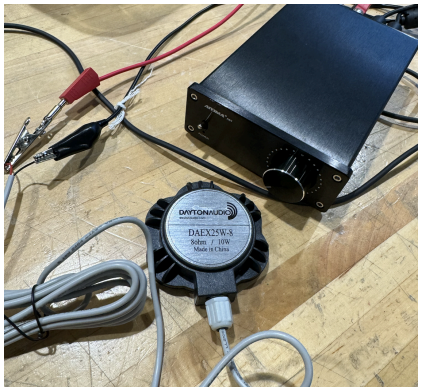


Figure 8. *DAEX25W-8 Speaker connected to amplifier*

The speaker was then tested in water, where it was found to be similarly louder compared to the piezo when both were submerged. At full amplification power, the DAEX25W-8 speaker was audible outside of the water when submerged 5 inches. This is shown in Figure 9, where the vibrations are visible on the water's surface. The Hydromoth was used to record this speaker's output underwater to be compared to the piezoelectric speaker.

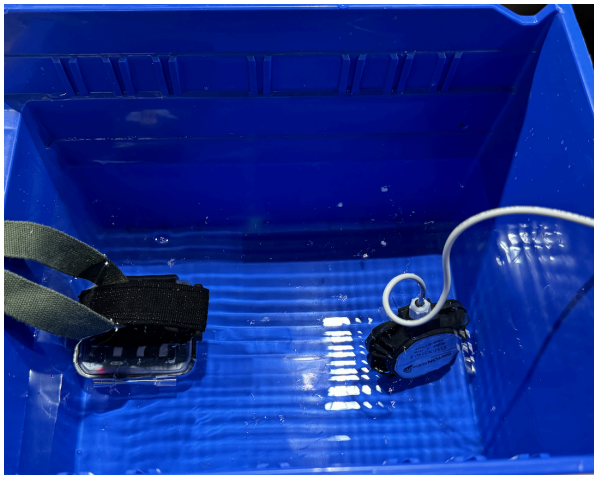


Figure 9. *DAEX25W-8 Underwater Sound Recording with Hydromoth*

This speaker enabled sound recordings with a higher signal-to-noise ratio than the piezoelectric speaker as recorded by the Hydromoth. To generate underwater recordings for training, the sample recordings must mimic the natural environment as closely as possible. The DAEX25W-8 speaker proves to be an ample device to create underwater sounds mimicking the real-world recordings that may be captured of fish species or anurans (frogs and toads) underwater in a pond.

D. Mounting HydroMoth in Aquatic Environments

In addition to considering the capabilities of the HydroMoth to record sounds in the local aquatic environments, the survivability and stability of the device were also of concern. The device comes in a clear plastic case with a woven fabric velcro strap which can be run through loops on the back of the case to secure the device to a pole. However, in the context of aquatic surveying, this method may not be enough to ensure device security for extended periods and can be difficult to replace by technicians. Therefore, a mounting device was constructed to contain the device securely in the environment by anchoring it in the ground while adding further protection for the HydroMoth. This design was created with tailored HydroMoth measurements within Fusion360 CAD software and rendered as shown in Figure 10. The CAD file as well as the featured images are located within the “HydroMoth Mounting System” folder of the project folder linked at the beginning of the report.

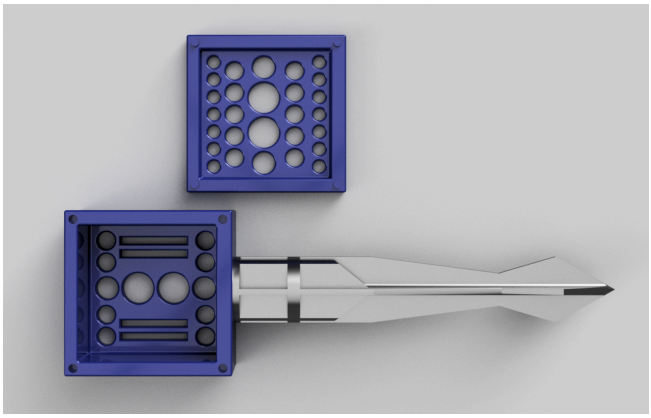


Figure 10. *Renderings of the designed HydroMoth mounting system*

The mount features enclosed housing for the HydroMoth in its waterproof case which is held in place by pin and hole joints and additionally by running the straps of the HydroMoth through the slits in the sides of the top and then through the back of the main enclosure. The back and front of the enclosure were made porous to enable sound waves to reach the recording device without too much added impedance. Additionally, attached to the main encasing is a robust, eight-inch spike designed to anchor the mount firmly in the ground of aquatic and semi-aquatic environments. The strap slits also enable the original mounting method of attaching to other wrapable features of the environment or an additional support. This system was not able to be tested properly within the period of the project due to issues with 3D Printing a prototype, but the mechanism would be useful for any future studies and recording using the HydroMoth for outdoor studies.

E. Machine Learning Audio Analysis - Traditional Algorithms

With little to no experience in the field of machine learning (ML), the first task was to attain a fundamental level of knowledge in using machine learning algorithms. We decided to start by generating our own simple machine-learning code. However, one of the most important parts of machine learning is determining which algorithm you are going to use. For machine learning on sound, it is especially important to determine which features are intended to be extracted from the audio files to use for identification characteristics in the audio model. So, several different features were tested as identification systems before moving on to more advanced code including Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff, Mel-Scale Spectrogram, Chroma Frequencies, and Mel-Frequency Cepstral Coefficients.

Spectral Centroid identifies the perceived brightness or center of mass (COM) within a spectrum, often utilized in music genre classification and sound analysis tasks due to its straightforward computation, offering a relatively fast processing speed (Giannakopoulos & Pikrakis, 2014). Spectral Bandwidth measures the frequency range width within a spectrum, aiding in distinguishing timbral characteristics in music analysis and sound classification, with a moderately fast processing speed suitable for real-time applications. Spectral Contrast captures spectral differences among adjacent frequency bands, enhancing audio feature discrimination in tasks like music genre classification, and processing at a moderate speed due to its detailed spectral analysis (Sable, 2021). Spectral Rolloff determines the concentration of spectral energy below a specific frequency, useful for distinguishing between harmonic and noisy components in audio signals, processing at a moderately fast speed suited for real-time applications (Tjoa, n.d.). Chroma features measure the presence of musical notes within an audio signal, crucial for tasks like chord recognition and melody extraction, with moderate processing speed suitable for music analysis applications.

Lastly, Mel-Frequency Cepstral Coefficients (MFCCs), known for capturing spectral characteristics akin to the human auditory system, are widely employed in speech and audio processing tasks such as speech recognition and speaker identification (Deruty, 2022). Although slightly more computationally complex, MFCCs offer efficient processing, making them suitable for real-time applications and preferred due to their superior performance in various audio analysis tasks. These algorithms were incorporated in machine learning code to determine which were most effective at identifying the sounds of semi-aquatic sounds, specifically anurans, when trained with an online database of frog and toad sounds.

F. Advanced Machine Learning - Recurrent Neural Network

After the traditional ML methods were trialed and analyzed, the team decided to briefly touch on the deep learning method. Deep learning is a machine learning technique that usually involves one or more layered and sophisticated functions, commonly referred to as deep neural networks, to be optimized to fit the dataset. This technique is often seen in modern cutting-edge scientific breakthroughs and is potentially very powerful.

Due to the team's entry-level understanding of the technique, a suitable tutorial is found (Nandi, 2021) to provide the first draft of the code. The team then improved upon the tutorial to develop a model more suitable for this project.

A recurrent neural network (RNN) is being used for this project. RNNs take in a sample data segment and output their prediction of the next segment. It is well known to be very effective in processing time series data such as sound recordings. Adding several fully connected layers on top of the RNN, a model suitable for audio classification is created. Specifically, the RNN is a long-short-term memory (LSTM) structure imported from tensorflow. The entire NN thus has the following structure:

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_21 (LSTM)              (None, 1024)              8294400

 dense_84 (Dense)            (None, 256)               262400

 dropout_63 (Dropout)        (None, 256)               0

 dense_85 (Dense)            (None, 64)                16448

 dropout_64 (Dropout)        (None, 64)                0

 dense_86 (Dense)            (None, 24)                1560

=================================================================
Total params: 8574808 (32.71 MB)
Trainable params: 8574808 (32.71 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 11. *RNN structure and amount of parameters*

Layer Type	**Function**
LSTM layers	Sequence data processing.
Dropout layers	Prevent overfitting by randomly omitting a portion of the data during training.

Dense layers	Act as fully connected neural network layers for classification.
Final Layer	Uses a softmax activation function (hidden in the graph) to output a probability distribution over the possible species classes.

Table 1. *Layer types of the neural network machine learning model and their functions*

To optimize the model, a loss function and an optimizer are also required. The Adam optimizer is chosen for its efficiency in handling the complex landscapes of audio data optimization. It adapts learning rates for each parameter, ensuring faster convergence and effective learning, crucial for the nuanced task of audio classification.

In the meantime, Sparse Categorical Crossentropy is selected for the loss function. This choice is due to the nature of the classification task, where each audio sample is associated with a single label from multiple classes. This function efficiently handles multi-class classification problems, providing a measure of the model's performance in predicting the correct species from the audio data, which is essential for accurate species identification. One major concern of using such a model and technique is the computational resource required for fitting the many parameters to the desired optimum. The team set up a local Windows Subsystem for the Linux environment with an Nvidia RTX 4070 graphics card for the task. Should the need emerge, Duke University also provides its cloud-based Duke Computer Cluster access for students.

G. Dataset

The team successfully collected a series of sound recordings from various sample places, such as Duke Pond, but the total amount of data was eventually not enough for model training. Thus, the dataset for Rainforest Connection Species Audio Detection (Kaggle Competition, 2023) is used. The data contains abundant sound recordings of birds and frogs, which is similar to the project’s goal. In the future, after more data is collected, the same model can be easily adjusted for the project’s data.

IV. Results & Discussion

A. Accuracy of Machine Learning Algorithms

Before performing any testing, it was assumed that the Mel-Frequency Cepstral Coefficients method of identification would likely be the most accurate. Since the algorithm is designed to process speech patterns, it would perform the most meticulous identification of the varying audio patterns found in wildlife vocalizations. Spectral Contrast should also perform fairly well in classification, as it may differentiate the tempo, pitch, and melody of the different anuran species. Due to their simplicity, Spectral Bandwidth, Centroid, and Rolloff were not expected to perform well, though Spectral Rolloff was expected to perform better than Spectral Centroid or Spectral Bandwidth since it is less susceptible to noise.

Algorithm	**Parameter Values**	**Percent Accuracy (n=5)**
MFCC	n/a	97.14%
Spectral Centroid	n/a	14.01%
Spectral Bandwidth	n/a	12.93%
Spectral Contrast	n/a	88.58%
Chroma Frequencies	n/a	72.20%
Spectral Rolloff	95% cutoff	13.36%
	90% cutoff	14.01%
	85% cutoff	15.52%
	75% cutoff	14.44%
	50% cutoff	12.93%
	35% cutoff	11.42%

Table 2 *Results of Each ML Sound Identification Algorithm for Anuran Sounds*

To test this hypothesis, machine learning code was created to test the algorithm of each method, located in the attached “Frog ML Code” Google Drive folder as “frogConfusionML.py”. The audio data used to train and test the system was a Kaggle data set uploaded by Turker Tuncer (in F23 Audio Folder in Frog Simple ML Code), consisting of 50

audio files for each species (Tuncer, 2023). The accuracy results for each method, and parameter values for each algorithm, are listed in Table 2.

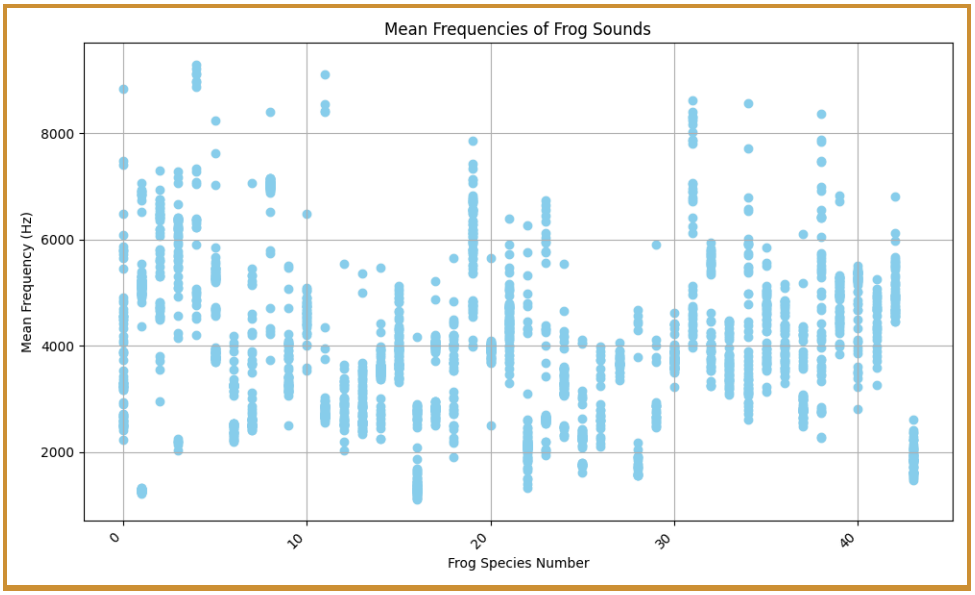
As predicted, MFCCs were the most accurate method by far. As we mentioned, this makes sense since MFCC is designed to distinguish particular speech patterns. Although originally designed for human speech, frog “speech” seems to be similar enough for the method to still work. Spectral Contrast also did well as expected, providing an accurate prediction around 9 of 10 times. Classifying frog ribbits as though they were different genres provides a fairly accurate model. Similarly, Chroma Frequencies, which measure the presence of musical notes, were fairly accurate as well, providing an accurate prediction in almost 3 out of 4 of the cases. The methods that performed poorly were Spectral Bandwidth, Spectral Centroid, and Spectral Runoff. All of these methods have very simple feature extractions, trading complexity for speed. Spectral Bandwidth and Centroid were not expected to have done well since they essentially provide a range and COM of the frog sounds, it would have been surprising if this was enough to differentiate 43 frog species. These two methods are also very susceptible to noisy systems. Repeated runs showed Spectral Centroid consistently performing slightly better than Spectral Bandwidth; nonetheless, both still lacked significant accuracy. What was surprising, however, was how Spectral Centroid performed about as well as Spectral Runoff, even better in some cases, depending on the cutoff point. Upon further thought, we believe this may be due to the fact Spectral Runoff only cuts off noise *above* a certain frequency (i.e. the *cutoff point*). This means, for instance, when the Spectral Runoff has a *cutoff percent* of 85%, it will return the highest frequency for which 85% of the noise is below it. Since frog ribbits are notoriously low frequency, this may be negatively affecting this prediction. Therefore, the effect of the ambient noise, which is likely of a higher frequency than the frog sound, is even more pronounced in this method.

As a new hypothesis and the results of the previous trials, it was predicted that using cutoff points that cutoff sounds *below* a certain frequency would more accurately capture the notoriously low frequency of frog sounds. This new method is called Reverse Spectral Rolloff and is located in the attached “Frog ML Code" Google Drive folder as “reversingSpectralRolloff.py”. This new algorithm was run on the same data with cut-off points at 90%, 85%, 50%, and 25% which produced mean accuracy results of 5.17%, 6.47%, 5.60%, and 6.03% respectively over three trials for each parameter value. While a bit better than just

guessing (2.33%), Reverse Spectral Rolloff performed the worst of any algorithm. This suggests, contrary to what we originally believed, that frog sounds are more concentrated/differentiated in the higher relative frequencies. Upon further research, these results make sense: against common wisdom, frog croaks are often high frequency. It is only a few notable exceptions, in particular the bullfrog, that have given them this reputation. Smaller frogs, which make up the majority of frog species, have a higher frequency (Narins, 1995). Moreover, tree frogs, due to strong ambient noise in the rainforests, often have even higher frequencies to differentiate their calls (Gridi-Papp, 2014). In fact, frogs have the highest auditory processing capabilities of any amphibian (University of California, Los Angeles, 2009).

B. Analysis of Data & Algorithm Results

After performing this simple machine learning analysis, we decided to take a deep look into the data we are using. The first analysis follows the discussion on the high frequency of frog sounds after the surprising failure of the Reverse Spectral Rolloff feature extraction method (Figure 12).



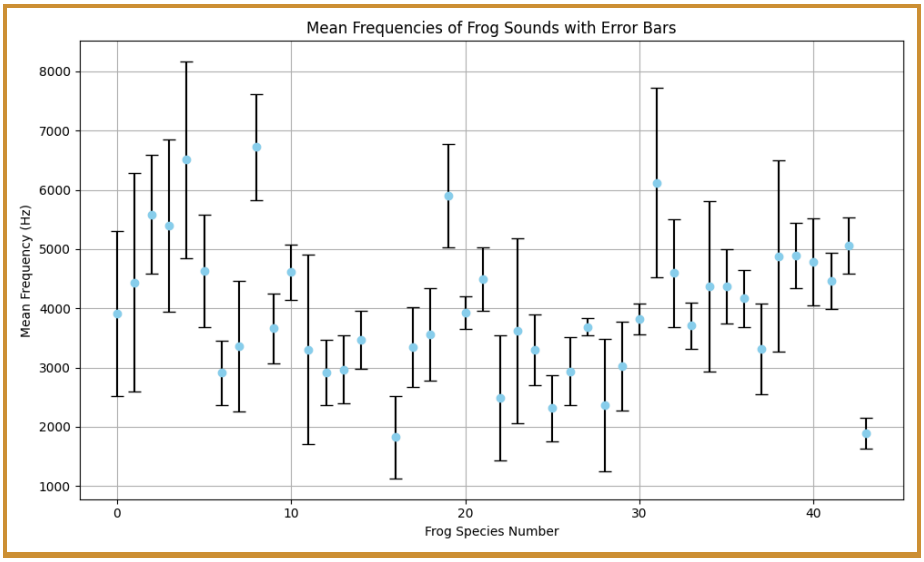


Figure 12. *Graph of mean frequencies for each recording in the used dataset, grouped by frog species number (top), and the average frequency across all sounds particular to a frog species, plotted with error bars at ±1 standard deviation (bottom).*

Interestingly, there is a large variance among the sound frequencies of each frog species. Frogs, 20, 27, 30, and 43 are the only notable exceptions, with 27 having an especially low variance. Additionally, all frequencies averaged over 1500 Hz for each sound recording with some as high as 9 kHz. For reference, the frequency of human speech ranges from 75 to 300 Hz (Lofft 2020). The record for the highest frequency sung by a male is D#8, around 5 kHz (Songstuff 2023). So, as we predicted in our conclusions from the Reverse Spectral Rolloff method, these frog sounds are high frequency. While we do not have the exact names of the frog species, we can conclude that these frog species are of a smaller variety, likely different types of tree frogs.

Confusion matrices are like a scorecard that helps us see how well our algorithm method understands the sounds. Along the diagonal, it shows how many frog sounds the computer correctly guessed. Off the diagonal, it displays the wrong guesses along with which frog species it thought the sound came from. In essence, it helps us understand where the algorithms made mistakes and how accurate they were for each sound it guessed. Therefore, confusion matrices were created for the outcome of each machine learning characteristic algorithm explored in this project, as displayed in Figure 13.

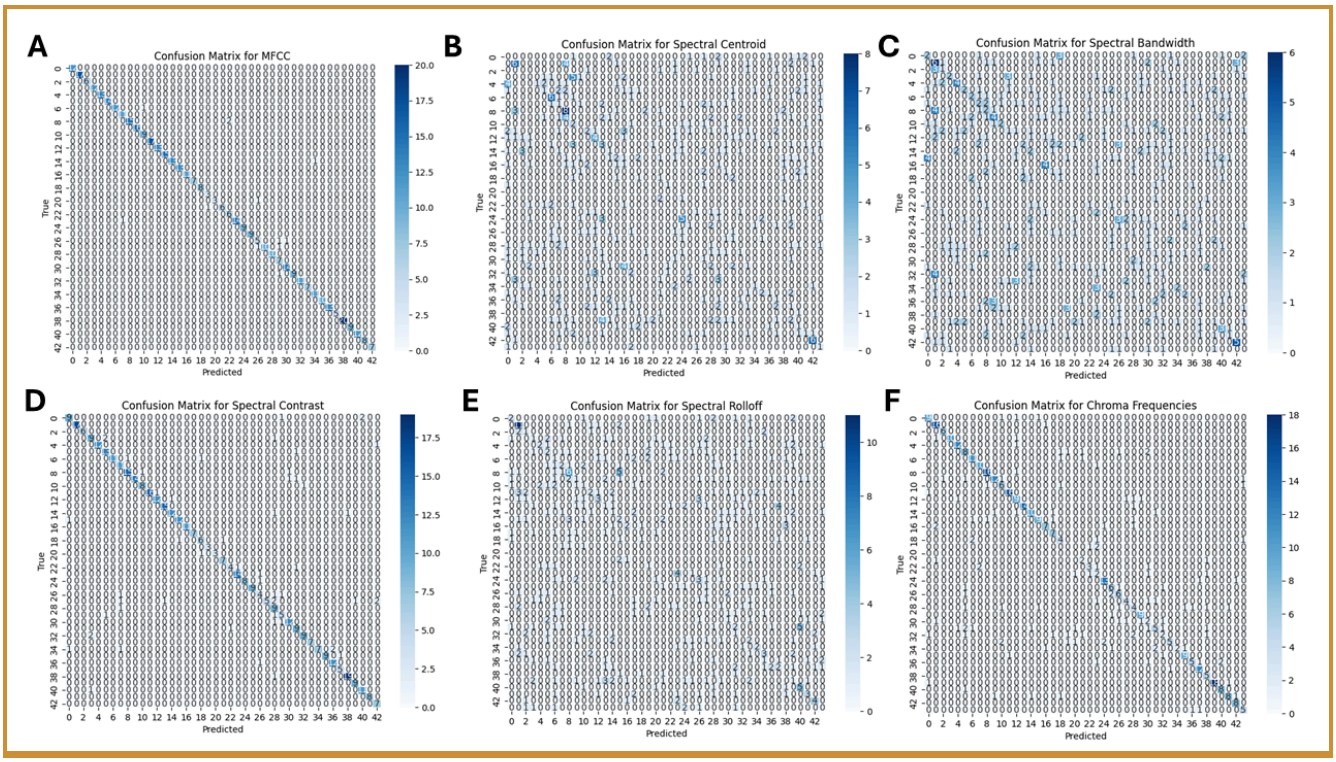


Figure 13. *Confusion matrices representing the results of A) MFCC, B) Spectral Centroid, C) Spectral Bandwidth, D) Spectral Contrast, E) Spectral Rolloff, F) Chroma Frequency algorithms.*

The accuracy for MFCC is 97.41%, so apart from a few scattered mistakes, the model was very accurate. Moreover, all mistakes were singular apart from 2 Frog 8 sounds being predicted to be Frog 22, as shown in Figure 13a. Other notes are that Frog 29 was incorrectly predicted twice, and Frog 26 was mistaken twice. Ultimately, it is a sign of high accuracy that we can see a clear diagonal line in the data. Lastly, while MFCC had the highest accuracy of any method, it was also the least efficient in processing speed. In contrast to MFCC, the Spectral Centroid Confusion Matrix does not have a clear diagonal band, as shown in Figure 13b. As a reminder, the accuracy for Spectral Centroid is 14.01%. So, the Spectral Centroid is inaccurate for this dataset. The spectral centroid measures the Center of Mass of the sound. Despite the inadequacy of this model, a few sounds, Frogs 1, 6, 8 12, 24, and 42, have several correct glances. However, when you account for the different sizes in each frog species sample set, you see that only Frog 42 was predicted accurately more than 50% of the time. This suggests Frog 42's center of mass is fairly distinct. Accuracy for Spectral Bandwidth is even worse than Spectral Centroid, averaging at 12.93%. The analysis of this data set is similar to Spectral Centroid, as shown in Figure 13c. There were a few sounds with a fair number of correct

predictions, but Frog 42 seemed to stand out as the only one with over 50% accuracy. As a reminder, the accuracy for Spectral Contrast is 88.58%. The Spectral Contrast method is fairly reliable, second in accuracy only to MFCC, as shown in Figure 13d. It is also the best of the methods with average processing efficiency (Spectral Contrast and Chroma Frequencies). Similar to MFCC and in contrast with Spectral Centroid and Bandwidth, there is a clearly defined diagonal line. However, unlike MFCC, there are a few instances of the same incorrect prediction for a particular Frog sound. To note on Frog 42, which seems to be an outlier in previous methods, it was predicted accurately 100% of the time. As a reminder, the accuracy for the Spectral Rolloffs (with a cutoff of 85%) is 15.52% which is very poor, but is the best of our time-efficient methods (Spectral Centroid, Bandwidth, and Rolloffs), as shown in Figure 13e. Interestingly, the accuracy of Frog 42 fell below 50% for the first time, suggesting while its center of mass and bandwidth may be distinct, its rolloff is not. In this case, only Frog 1 has an accuracy above 50%, predicting correctly 11 of 17 times. However, it should be noted that Frog 1 is likely overfitted since it was overpredicted and wrong 13 of 24 times. As a reminder, the accuracy for Chroma Frequencies is 72.20%. This was the third best, though the worst of the algorithms with medium-to-high processing cost (Spectral Contrast, MFCC, and Chroma Frequencies). Interestingly, there seems to be a divide in accuracy across species. For instance, some sounds were highly accurate, with Frogs 2, 17, 38, and 42 even having 100% accuracy (again, Frog 42 seems to have a unique sound), as shown in Figure 13f. Yet, others, like Frogs 19, 30, 32, and 34 had very low accuracy. Frog 19 was never accurately predicted—Frog 20 was also never correctly predicted, but in this particular random sample a sound from it was never selected for testing. Since Chroma Frequencies are used to extract musical notes from sounds, perhaps some Frog species have a more musical quality, with an emphasis on notes, than others.

Overall, we found that the more complex feature extraction methods resulted in higher accuracies. The MFCC method, which is used to model human speech, performed best. Second to MFCC, the methods model musical rhythm and notes did moderately well. The method of measuring notes was the most divisive, performing accurately for some Frog sounds, but not others. The simple methods of center of mass, range, and, and measuring lower/upper points of the range, did very poorly. Lastly, we wanted to investigate Frog 42, which seemed to have a unique sound. Upon listening to the audio files, we discovered that it was in fact a recording of

bird sounds, not frogs. This is promising for applications of the MFCC software on species other than anurans since MFCC was able to classify Frog 42 just as well as the other sounds.

C. Recurrent Neural Networks Performance

Before the RNN training is started, a key challenge in machine learning, particularly in a complex task like audio classification, is overfitting. This occurs when the model performs well on training data but poorly on unseen data. To mitigate this, dropout layers are used, and the model's performance is closely monitored on both training and validation sets. Adjustments are made to the model architecture and training process based on these observations.

Similar to previously mentioned in the previous section IV.B, the audio data needs to be processed to extract features from it. As we determined in our traditional machine learning, MFCCs, which effectively represent the power spectrum of sound in a way that mimics human speech, were the most accurate. Therefore, this is the feature we extracted for our RNN. These coefficients are extracted using the Librosa (Librosa Development Team., n.d.) library, to capture the relevant audio characteristics.

The audio data, once processed into MFCCs, is log-scaled to ensure consistent amplitude ranges across different samples. This step is crucial for effective learning. The dataset is then divided into training, validation, and test sets, maintaining a balance to avoid biased learning.

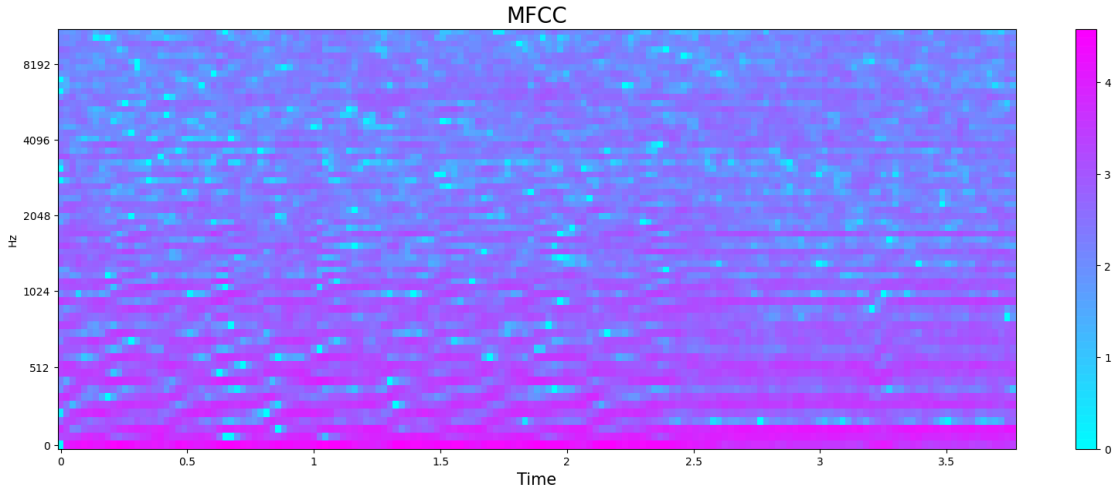


Figure 14. *MFCC feature extracted and scaled from a segment of sound recording*

The RNN is then trained with 150 epochs with a batch size of 512. These hyper-parameters are selected through experimentation of different sets and with one that yields the best result. The training stops at the 150th epoch, as Figure 15 below shows, the critical performance metrics, accuracy, have stopped improving.

Note that the loss during training peaked several times. This indicates that the model was trying to move out of local minima to a better-fitting position.

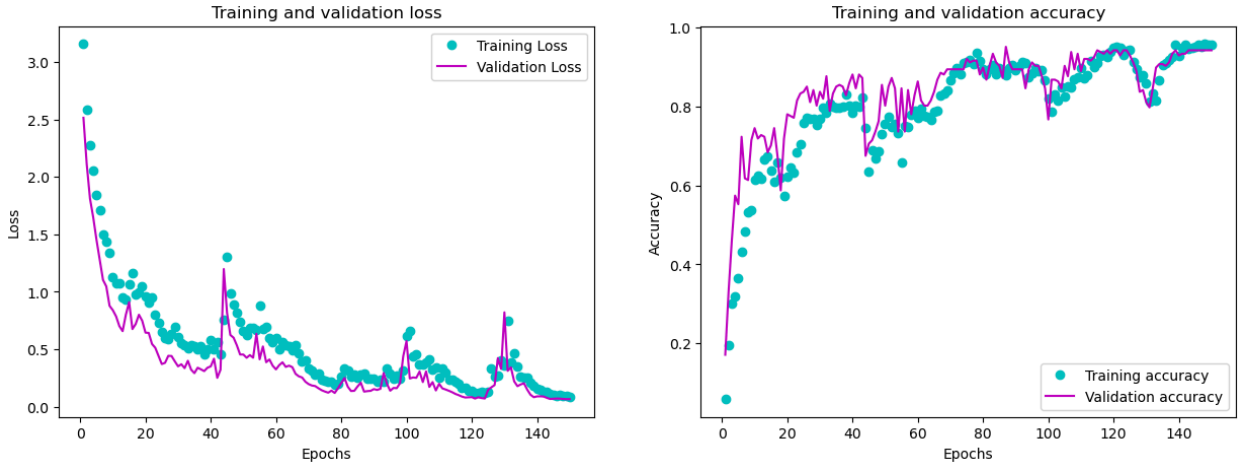


Figure 15. *Training history*

Eventually, the trained model is tested on a reserved test data set to gauge its performance. This test data set was set aside from the very beginning and was never seen by the model. It ensures the fairness and fidelity of the model's performance when applied to the analysis of newly collected data. The test eventually yields a 95% accuracy. The confusion matrix, which gives details of the model's performance and biases, is shown below in Figure 16.

```
Confusion_matrix:  tf.Tensor(
[[12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  9  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 11  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0  1  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 13  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 11  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 25]]
```

Figure 16. *Confusion matrix of the RNN*

As mentioned in the previous section, a good model would have its prediction results lie on the diagonal of the matrix, which is pretty much the case. One exception is on the 13th species, which is too similar to the 8th species in terms of MFCC, and the RNN is having a hard time distinguishing between them. This is where the absolute most of the errors come from.

Note that this error is associated with this specific dataset, and is expected to be automatically mitigated when switched to aquatic sound recording data in the future. If one was to temporarily allow this error, in acknowledgment of the model's limitations, the resulting overall performance would be a 99% accuracy, which is sufficiently reliable for this project.

D. Limitations of the Hydromoth Recording System

The Hydromoth exhibits certain limitations that warrant consideration in the context of its application for underwater acoustic recording. Notably, the HydroMoth demonstrates a lower signal-to-noise ratio compared to SoundTraps, suggesting potential challenges in detecting faint signals, particularly in noisy environments (Lamont et al., 2022, 362-378). Its frequency-dependent detection ability, matching closely with South Trap below 15 kHz but exhibiting differences at higher frequencies may result in missed detections of marine mammal sounds above 40 kHz. This limitation could be critical in situations where the target signals are relatively quiet. Moreover, there's directional bias present within the Hydromoth. The asymmetrical design of the Hydromoth casing introduces a directional bias, impacting its sensitivity to sounds coming from different directions. This bias is evident in the study's findings, where back-to-back HydroMoths showed discrepancies in detecting dolphins (Lamont et al., 2022, 362-378). The need for multiple HydroMoths oriented in various directions emphasizes a potential limitation in applications where precise sound source localization is crucial. Moreover, uncalibrated sensitivity and frequency response could cause discrepancies within the data collected. The lack of calibration for HydroMoth introduces uncertainty regarding its sensitivity and frequency response.

Without proper calibration, it becomes challenging to use HydroMoth for applications requiring accurate quantification of sound pressure levels. This limitation restricts its utility in studies that compare absolute sound pressure levels across different ecosystems or assess the impact of noise pollution on aquatic environments. The instrument error observed in simultaneous recordings from different HydroMoths suggests that small differences in the placement of recording devices can contribute to variations in recorded data. Users need to be

cautious when comparing outputs from different HydroMoths or hydrophones, considering the potential impact of instrument error on the interpretation of results. While HydroMoths are cost-effective recording instruments, the study highlights that operational costs associated with deployment, retrieval, and data management may still be significant. However, hopefully, our design for the HydroMoth mounting system would help ameliorate this issue. Nonetheless, this is an essential consideration for users planning large-scale deployments or programs with high operational overhead. The reliance on an open-source support forum for technical assistance may pose challenges for users who require more structured technical support with warranties. In contexts where users lack the expertise for independent troubleshooting, HydroMoth might not be the most suitable and cost-effective solution.

VI) Conclusions & Future Considerations

To improve species classification accuracy, future work should refine feature extraction and explore advanced neural network architectures. For algorithmic accuracy, fine-tuning existing models, especially MFCC, by adjusting parameters and architectures is recommended. Ensemble learning methods and data augmentation, incorporating variations in pitch and background noise, can enhance classification accuracy by combining strengths and expanding datasets.

Expanding the database is crucial as well. Future efforts should include a diverse representation of species, especially those with more unique vocalizations. Although the data was primarily tested on rainforest amphibians, further testing out of this niche would prove useful to observe accuracy across a wider range of sounds. Incorporating recordings from various geographical locations accounts for environmental variations; additionally, long-term monitoring captures seasonal changes and behavioral shifts, providing a comprehensive model adaptable to dynamic ecological conditions in aquatic and semi-aquatic ecosystems.

Additionally, considering the complex nature of acoustic data in aquatic environments, future research could explore the implementation of more sophisticated neural network architectures. Neural networks, particularly deep learning models, have shown success in handling intricate patterns and temporal dependencies. Utilizing advanced architectures, such as Convolutional Neural Networks (CNNs) or attention mechanisms, may offer improved feature

learning and enhance the model's ability to discern subtle variations in species-specific sounds, further elevating the system’s capacity for accurate species identification across a diverse variety of ecosystems.

In conclusion, our study delves into the application of advanced machine-learning techniques and specialized recording devices for monitoring aquatic and semi-aquatic ecosystems. Noteworthy findings include the superior accuracy of the MFCC method in frog sound classification over other methods such as Spectral Centroid and Spectral Bandwidth, despite certain processing inefficiencies. The HydroMoth recording device utilizes features for underwater audio data collection, contributing to its widespread use in ecological studies. Our work aligns with current trends in aquatic ecosystem research by employing Recurrent Neural Networks with Long Short-Term Memory units and emphasizing the significance of MFCCs for accurate feature extraction.

By integrating these features into the existing algorithm, future students can build upon our work to create a more accurate, adaptable, and comprehensive system for aquatic and semi-aquatic soundscape analysis. This iterative process is fundamental to the evolution of both machine learning techniques and ecological monitoring strategies across numerous ecosystems.

References

DAEX25W8 Waterproof 25mm Exciter 10W 8 Ohm. (n.d.). Dayton Audio. Retrieved December 13, 2023, from https://www.daytonaudio.com/images/resources/295-232--dayton-audio-daex25w-8-specifications.pdf

Dayton Audio. (2023). *DAEX25W-8 Waterproof 25mm Exciter 10W 8 Ohm*. Dayton Audio. Retrieved December 14, 2023, from https://www.daytonaudio.com/product/1181/daex25w-8-waterproof-25mm-exciter-10w-8-ohm

Deruty, E. (2022, September 16). *Intuitive understanding of MFCCs. The mel frequency cepstral coefficients*. Medium. Retrieved December 14, 2023, from https://medium.com/@derutycsl/intuitive-understanding-of-mfccs-836d36a1f779

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB® Approach*. Elsevier Science. https://doi.org/10.1016/B978-0-08-099388-1.00004-2

Gridi-Papp, M. (2014, September 23). Is the Frequency Content of the Calls in North American Treefrogs Limited by Their Larynges? *International Journal of Evolutionary Biology*, *2014*. Hindawi. https://doi.org/10.1155/2014/198069

Hughes, M. (2016, June 29). *How Piezoelectric Speakers Work - Technical Articles*. All About Circuits. Retrieved December 14, 2023, from https://www.allaboutcircuits.com/technical-articles/how-piezoelectric-speakers-work/

Kaggle Competition. (2023, March 2). *Rainforest Connection Species Audio Detection*. Kaggle Competition. Retrieved December 15, 2023, from https://www.kaggle.com/c/rfcx-species-audio-detection/data

Lamont, T. A.C., Chapuis, L., Williams, B., & Dines, S. (2022, January 4). HydroMoth: Testing a prototype low-cost acoustic recorder for aquatic environments. *Remote Sensing in Ecology and Conservation*, *8*(3), 362-378. Zoological Society of London. https://doi.org/10.1002/rse2.249

Librosa Development Team. (n.d.). *librosa — librosa 0.10.1 documentation*. Librosa. Retrieved December 15, 2023, from https://librosa.org/doc/latest/index.html

Lofft, A. (n.d.). *Audio Oddities: Frequency Ranges of Male, Female and Children's Voices.* Axiom Audio. Retrieved December 14, 2023, from

https://www.axiomaudio.com/blog/audio-oddities-frequency-ranges-of-male-female-and-childrens-voices

LSU. (2017, April 10). *Frequency Hearing Ranges in Dogs and Other Species*. LSU. Retrieved December 14, 2023, from https://www.lsu.edu/deafness/HearingRange.html

Music Note Fundamental Frequencies. (n.d.). Songstuff. Retrieved December 14, 2023, from https://www.songstuff.com/recording/article/music_note_fundamental_frequencies/

Nandi, P. (2021, March 1). *Recurrent Neural Nets for Audio Classification | by Papia Nandi*. Towards Data Science. Retrieved December 15, 2023, from https://towardsdatascience.com/recurrent-neural-nets-for-audio-classification-81cb62327990

Narins, P. M. (1995). *Acoustical Society of America - Frog Communication*. Acoustics.org. Retrieved December 14, 2023, from https://acoustics.org/pressroom/httpdocs/swa9501.html

Open Acoustic Devices. (n.d.). *AudioMoth*. Open Acoustic Devices. Retrieved December 13, 2023, from https://www.openacousticdevices.info/audiomoth

Sable, A. (2021). *Introduction to Audio Analysis and Processing*. Paperspace Blog. Retrieved December 14, 2023, from https://blog.paperspace.com/introduction-to-audio-analysis-and-synthesis/

Sethi, S. S., Jones, N. S., Fulcher, B. D., Picinali, L., Clink, D. J., Klinck, H., Orme, C. D. L., Wrege, P. H., & Ewers, R. M. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. Proceedings of the National Academy of Sciences of the United States of America, 117(29), 17049–17055. https://doi.org/10.1073/pnas.2004702117

Tjoa, S. (n.d.). *spectral_features*. Music Information Retrieval. Retrieved December 14, 2023, from https://musicinformationretrieval.com/spectral_features.html

Tuncer, T. (2023, March 2). *Rainforest Connection Species Audio Detection*. Kaggle. Retrieved December 15, 2023, from https://www.kaggle.com/c/rfcx-species-audio-detection/data

University of California, Los Angeles. (2009, May 9). Ultrasonic Communication Among Frogs. *ScienceDaily*. https://www.sciencedaily.com/releases/2009/05/090508192231.htm