Cetacean Identification: A Machine Learning Approach to Classifying Whales by Their Songs

Brody Dobson, Eliza Homer, Mark Thomas Watson, Zack Yancey
CS 201R, Fall 2023
Department of Computer Science
Brigham Young University

Abstract

This research aims to automatically identify whale species based solely on their vocalizations. Leveraging various machine learning models, we present an approach that utilizes scraped data from the Watkins Marine Mammal Sound Database. The data undergoes processing, feature extraction, and training on multiple machine learning models. Our results indicate significant success on test data using various classifier models. Through various featurization and model testing, we were able to achieve a 91.8% accuracy for correctly predicting if a cetacean was a whale or a dolphin on a test with the random forest model.

1 Introduction

Whales, dolphins, porpoises, and other cetaceans are renowned for their social nature, often forming tight-knit groups. They create their own form of whale song by producing a variety of clicks, pops, and whistles as a means of communication with fellow pod members. To gain deeper insights into the world of cetaceans and foster safer coexistence with these majestic mammals, it would be helpful to be able to automatically identify species using only their vocalization. Such an endeavor would aid scientific research in a myriad of ways. Using various machine learning models we hope to predict the different species of cetacean based on a small audio clip of their vocalizations.

1.1 Big Brained Mammals

Cetaceans are highly intelligent mammals. Besides possessing very large brains, they also exhibit many characteristics of intelligent life. They have self-awareness and exhibit problem-solving behaviors. Another sign of their intelligence is their playful nature. Some dolphins have been recorded to interact with humpback whales, so that the humpback will raise the dolphin out of the water then allow the dolphin to slide into the water with a splash [*How intelligent are whales and dolphins?,* 2020]. Whales also show signs of feeling love, loss, and friendship [Zeppetelli-Bédard, 2021]. Each pod of whales has a unique culture. The culture is passed down from mother to offspring as she teaches the young whale hunting techniques as well as the best places to migrate.

1.2 Whale Songs

Part of the reason whales and dolphins have developed such complex brains is to communicate with each other through sound. The sounds they make are often called songs due to the low melodic tones some whales produce; however, not all sounds fit this description. Depending on the species, as well as the type of communication, the cetacean noises could be pops, clicks, whistles, goans, squeaks, rumbles, buzzes or trills. Using various recording equipment, these sounds can be captured.

Since water molecules are much more dense than air molecules, sound can travel much faster in water than through the air. The sound frequency determines how far a particular sound can be heard, but some whale songs, such as the humpback whale, can be recorded over 10,000 miles from where it originated [Conlin, 2019]. Identification using purely the sound these mammals create would enable classification long before the cetacean would be close to the sound recorder.

1.3 Whale and Human Encounters

Whales are known to be gentle giants of the oceans. They are some of the largest creatures on the planet. The blue whale is the largest animal on the earth, growing up to 28 meters and weighing over one hundred tons [Goldbogen, 2020]. There are instances of whales helping and saving humans. Just a couple of years ago, video footage shows a large humpback whale protecting a driver from a 15-foot tiger shark [Specker, 2018]. The diver's life was spared as the whale helped her get back to the boat while avoiding the shark.

Whales have also reached out and seemingly asked for help from humans. A gray whale was recorded coming up to a whale watching boat and asking for the parasites on its head to be removed [Davis, 2023]. A sperm whale approached human divers to have a fishing hook removed from its mouth [Lepakko, 2023]. A young orca found a group of humans and called until they came and helped a large orca who was entangled [Sparke, 2023]. These are just a few of the examples of people helping whales.

Unfortunately, not all whale-human interactions are positive. Recently, orcas have started to attack boats. This behavior is unprecedented and scientists are working to figure out why these whales have begun to do this. Just this past year (2022) orca pods sank three boats off the coast of Spain [Pare, 2023]. These attacks have been coordinated and deliberate with the killer whales targeting the rudder of the boat then repeatedly ramming the side. The two main theories for their behavior are that the orcas are just playing with the boat, or an orca was hit by a boat and is now teaching others in their pod to attack boats [Morris, 2023]. Until this mystery is better understood, people are doing their best to avoid pods of orcas in this part of the world. Using machine learning to identify orca calls could offer the possibility of preemptive measures to avoid encounters with pods of orcas known to exhibit aggression.

1.4 Whale Conservation

More recently cetaceans have been facing new challenges such as climate change and human disruption. Climate change has affected the ocean environments as well as the food that whales and dolphins need to survive. Changing climates are causing them to display different behaviors, some of which are not sustainable. Humans create noise with boats and fishing that could interfere with the whales' natural calls. The fishing nets and gear are also a hazard that could entangle cetaceans. Classifying cetaceans through sound could help researchers monitor species populations and better understand the impact climate change and humans have on these remarkable creatures.

2 Data

2.1 Data Source

Due to the nature of the data needed to train the models and the land-locked state of the research team's location, it would have been impossible to record the selection of whale noises that would be needed at the scale required. We considered a few different sources but eventually stumbled on a repository of unlabeled audio clips curated by the New Bedford Whaling Museum. Going forward in this paper we will refer to the data source as the Watkins Marine Mammal Sound Database or more simply the Watkins Database [Watkins and Schevilll, 2023]. While we had a source that we could use, the issue remained that the data was not clearly labeled and varied significantly in file length and file type.

2.2 Extraction

We wanted to have consistent data for a variety of whale species, as that would provide the best results for our models. The Watkins Database had a section of Best Of Clips that they used to demonstrate the quality of their data and was labeled by species. However if we pulled the data by the bulk download service they provided the data was completely unlabeled. We created a simple python web scraper that would iterate through the data that was partially labeled, by scraping the hosted files and by placing them into subfolders which were labeled by species. This served as the foundation for the data which we would be able to feed into a processing pipeline to homogenize the data and extract features.

2.3 Processing Challenges

As we began analyzing the data that we scraped we began to notice that the data was very irregular and would need to be processed before we extracted features if we wanted our models to be able to treat the data fairly. First, we wanted to isolate just whales and dolphins in our dataset and the Watkins Database contained a selection of audio samples from other aquatic mammals such as seals and walruses. As part of our data processing pipeline we flagged these subfolders in our data and removed them. In order to gain more insight into the shape of the data we wrote a small program that would collect metrics of all of the audio files we sampled and display them for easy analysis (see figure 1).

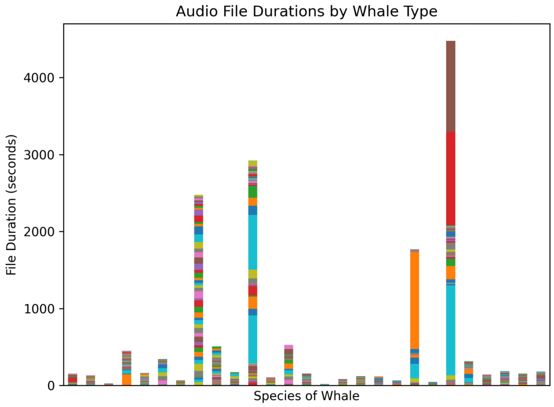

Figure 1: Length of audio file durations by whale type before processing.

This highlighted a large issue with our dataset. First, we noticed that some species were over-represented in the dataset, second, that file length varied from half a second to half an hour, and finally, there were several cases where a species combined data was no more than ten seconds. All of these served as significant challenges and could introduce bias into our data after we processed it. We knew that we needed to create a processing pipeline that would allow for our data to be ready for feature extraction. Eventually after

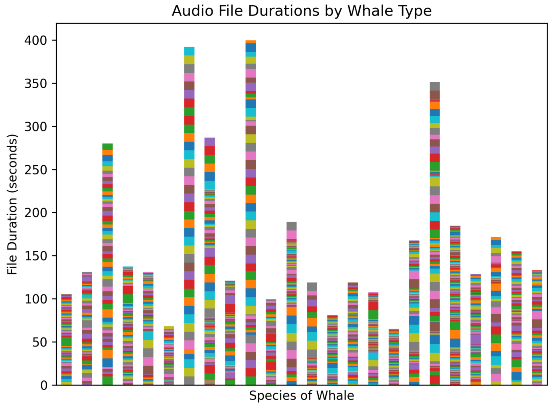

Figure 2: Length of audio file durations by whale type after processing

some testing with different metrics and flags we settled on the following rules for our data: 1) All data must be from either a whale or a dolphin 2) No file can be longer than 10 seconds in length 3) The combined length of all files for a species of whale or dolphin must be longer than 60 seconds 4) The combined length of all files for a species of whale or dolphin cannot exceed 400 seconds. It can be observed in Figures 1 and 2 that after this process the data was much more uniform.

After cleaning the data we began looking at the raw files and noticed was that the data was incredibly noisy, which could impact our ability to extract meaningful features. After evaluating some of our options we opted to leverage the python NoiseReduce library [Sainburg, 2020], which uses spectral gating to clean up audio signals and could be built into our data processing pipeline.

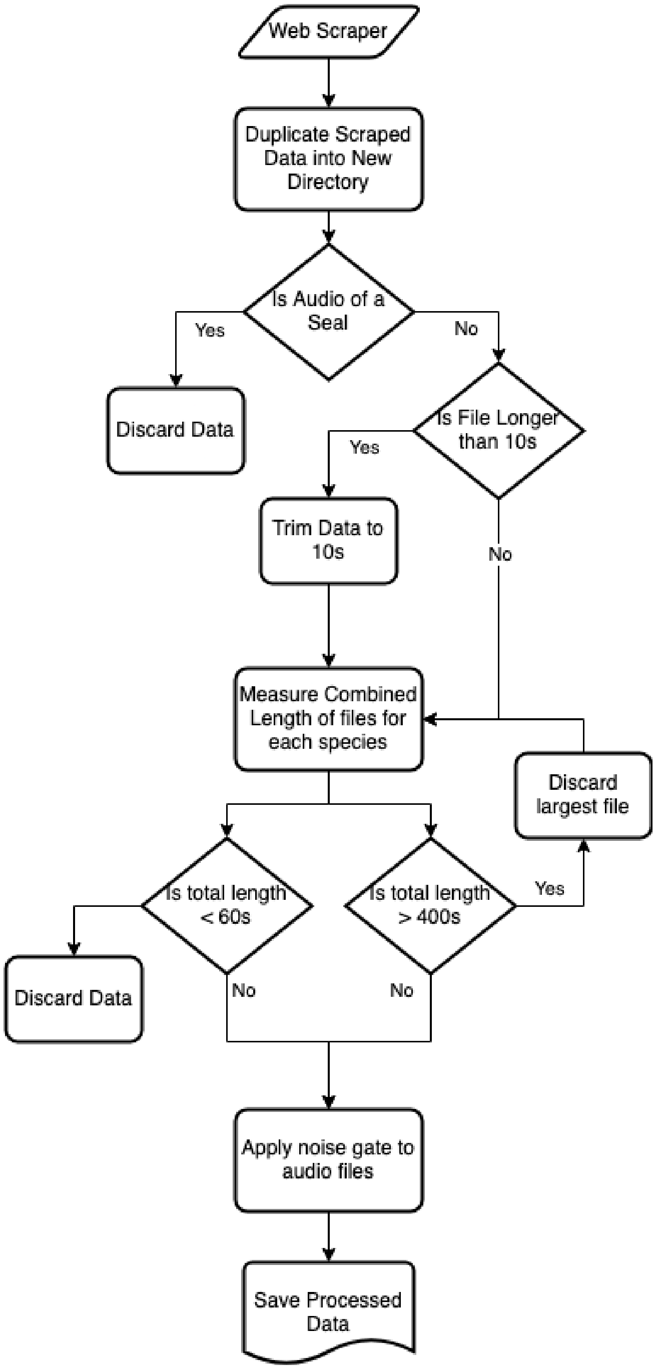

Figure 4: Audio Processing Flow Diagram

This worked quite well with our bioacoustic data as can be seen here

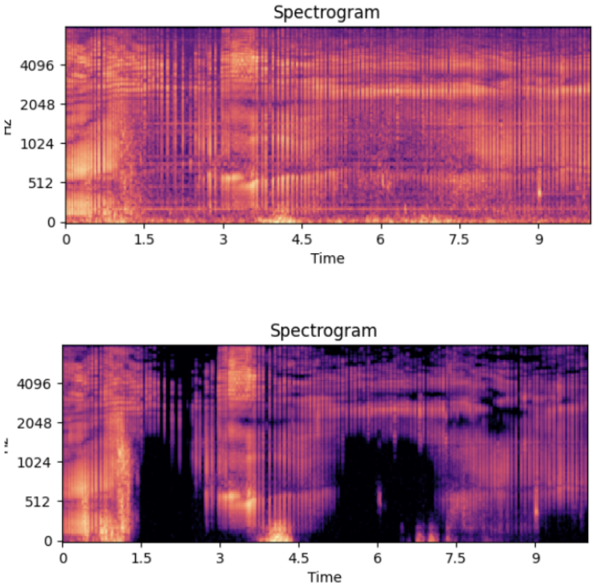

Figure 4: Audio Spectrogram before and after the noise reduction

and allowed us to have data which was ready for feature extraction.

Ultimately our data processing pipeline prior to feature extraction ended up following the flow described in Figure 4. As the initial audio data was not ours to distribute we used the scraper and data processing pipeline to ensure that all of the team members were working with the same data and anyone who wanted to reproduce our project would be able to do so as well without relying on us publishing our modified dataset. In addition it allowed for rapid iteration on our data to settle on the best features for our models.

2.4 Finding the Best Features

Classifying audio data is a balancing act. We knew that it would be far too easy to overfit our data if we provided too many features but we also didn't want to simplify our data in such a way that we would sacrifice meaningful information. After some research we decided to use LibRosa [McFee, 2015], a python library designed for audio processing as our method for feature extraction. After some reading we settled on Mel-frequency cepstral coefficients or MFCCs for our initial feature set

Mel-Frequency Cepstral Coefficients

MFCCs are a method of feature extraction which forms a compact representation of the spectral characteristics of the audio signal. While we will not explicitly go into depth about what these MFCCs represent, in brief, they take the audio signal, split it into frames of about 40 ms and then apply multiple transformations to the signal which results in an array of coefficients being returned for each frame. These two dimensional MFCC arrays work surprisingly well as a way of simplifying the data for a machine learning model while retaining data integrity.

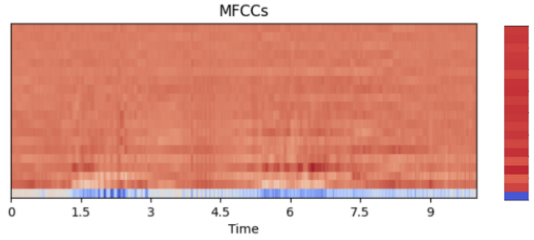

Figure 5: Visualization of Mel-Frequency Cepstral Coefficients over time. The bar on the right represents the MFCCs averaged over the entire length of the sample.

For demonstration purposes, this is what the two dimensional matrix of MFCCs looks like visually for the noise gated spectrogram shown earlier where cool to warm is represented by a coefficient between 0 and 5. This posed a problem for us though as we wanted to keep our features to a maximum of about 50, and a 10 second clip would result in about 10,000 MFCCs, assuming 40 coefficients per frame. We opted to take an average for each coefficient across all the frames to create a simplified token for the audio file of 40 numbers which could serve as features. This average is represented by the vertical bar to the right of the full MFCC chart.

Additional Key Audio Features

Machine learning classifiers are quite good at classifying abstract data but we determined that too much data was being lost for each file by reducing it down to just the average MFCCs. As such we determined that we should find several more features which would serve to differentiate the audio files. Primarily we wanted several singular numbers which could represent both spectral characteristics of the data as well as rhythmic features of the data. Some features such as Tempo Estimation (BPM) or Most Prominent Pitch returned numbers but others returned an array which represented the value over the time series of the duration of the file. For these features we just took the mean value across the duration and returned that number. While, like with the MFCCs, we lost some information in this process, it provided very unique features that described the general shape of the audio data.

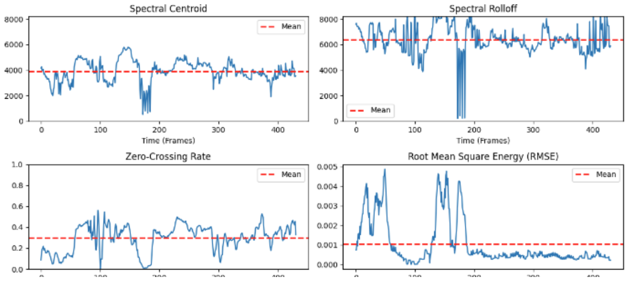

Figure 6: Visual Representation of additional features gathered

2.5 Optimizing Features for Training

After gathering all these features together we figured it would be beneficial to be able to have various forms of the dataset which we could perform our training on. Some of the additional features we added were on a very different scale to the MFCCs so we also created a version of the dataset which was normalized to ensure that we weren't introducing bias into the model due to the shape of the data.

Dataset	Description
mfcc_only.csv	For this dataset we only include the mean values generated by the MFCC algorithm.
mfcc_only_noisegate.csv	Same as mfcc_only but with the noise gate applied to the initial data.
full_data_not_normalized .csv	MFCC data with additional features [Chroma, Spectral Contrast, Bandwidth, Centroid, Rolloff, Zero Crossing Rate, RMS Energy, Tempo]
full_data_not_normalized_ noisegate.csv	Same as full_data but with the noise gate applied to the initial data.
full_data_normalized.csv	Same as full_data but normalized.
full_data_normalized_ noisegate.csv	Same as full_data but with the noise gate applied to the initial data and normalized.

Table 1 The six different versions of our dataset.

3 Models

To determine the optimal model for our dataset, we conducted an exhaustive evaluation of multiple models. Initially, we decided to test the models developed throughout the semester, including our perceptron, MLP, decision tree, and k-nearest neighbor models. However, we wanted to broaden our experimentation by incorporating additional models, such as support vector machines, random forests, and gradient boosting. For all our implementations of these models we use scikit-learn [Pedregosa *et al*., 2011]..

3.1 Simple Perceptron

As the number of features being used in this dataset is relatively high, it is unsurprising that the simple perceptron was not very successful at classifying the data. Default parameters using the scikit-learn package and any of the "whales only" datasets averaged over 500 instances of training resulted in training accuracy between 16% and 18%, and test accuracy slightly higher than 7%.

Hyperparameter tuning also proved ineffective, as variations to the learning rate, penalty, and normalization only seemed to decrease test accuracy. Some overfitting may still have been in effect, but the complexity of the dataset and the simplicity of the simple perceptron model meant that very little could be done to reduce it.

As the simple perceptron was only developed to use as a baseline, we felt it unnecessary to develop it any further. We determined that the naive classifier with only whale data should return an 8% accuracy, meaning that the simple perceptron performed worse on the test sets on average than a naive approach.

3.2 Multi-layer Perceptron

Following our use of the basic perceptron, we felt it was natural to implement a multi-layer perceptron with backpropagation. We used one hidden layer with 64 nodes. We also used an initial learning rate of 0.01 and no regularization. The MLP always had a higher test set accuracy than our baseline. The highest performance was seen on the “full data normalized” dataset with whales vs dolphins classification where it achieved an accuracy of 75% on the test set.

3.3 Decision Tree

The decision tree algorithm is a versatile and powerful classification tool. The decision tree algorithm is adept at capturing non-linear relationships and interactions among features, making it well-suited for discerning nuanced distinctions between different whale categories. During initial training and testing on our various datasets, we chose to run a decision tree with the default hyperparameters to get a baseline score. Our results indicated overfitting with a consistent 99% train accuracy, and 20-40% test accuracy.

From there, we fine tuned our hyperparameters to optimize our model accuracy on our datasets. Our finalized set of hyperparameters included the entropy criterion, a minimum impurity decrease of .025, and a balanced class weight. Running the decision tree model using these hyperparameters increased the test accuracy to 56% with no evidence of overfitting.

Hyperparameters

In the implementation of our Decision Tree model for whale classification, specific hyperparameters were carefully chosen to achieve optimal performance. The criterion parameter was set to "entropy," indicating the use of information gain as the criterion for making splits in the tree. The parameter max_depth=None allowed the tree to grow without a maximum depth limit so we could visualize the entire tree. To control the complexity of the tree and avoid overfitting, min_impurity_decrease was set to 0.025, specifying the minimum impurity decrease required for a node to split. Additionally, the class_weight parameter was set to "balanced," adjusting weights based on class distribution to handle potential imbalances in the dataset. These hyperparameter choices contribute to the Decision Tree's effectiveness in discerning patterns within the whale classification data while addressing considerations of tree structure and class distribution.

3.4 Support Vector Machine

Support Vector Machines provide a more robust capacity to discern intricate patterns within diverse datasets. Before hyperparameter fine tuning, our accuracy score maxed out around 23% using our initial dataset. After hyperparameter fine tuning, that accuracy increased to 31%. This result was promising and produced one of the highest accuracies among the other models. However, when we ran the same model on our updated datasets, we only saw a 1% increase in accuracy.

Hyperparameters

In the implementation of our Support Vector Machine (SVM) for whale classification, we chose a radial basis function (kernel='rbf') to capture non-linear relationships in the data. The choice of gamma='auto' adjusts the kernel coefficient automatically, enhancing adaptability. The shrinking=True setting enables the SVM to leverage a shrinking heuristic during training, optimizing computational efficiency by reducing the number of support vectors. Together, these hyperparameters contribute to the SVM's effectiveness in discerning intricate patterns in the whale classification task.

3.5 K-Nearest Neighbor

The K-Nearest Neighbor algorithm was another fairly simple algorithm compared to the complexity of the dataset, yet its test results were surprising when hyperparameters and preprocessing were optimized. For hyperparameter tuning, the “all animals” datasets were used, and results were averaged over 100 trials.

Hyperparameters

Running on just the MFCC features with default parameters resulted in a test accuracy of around 15-17%, only slightly better than the naive approach. However, the default parameters did not include distance scaling, the inclusion of which increased test accuracy by 5-10 percent. Changing the distance calculation to manhattan distance instead of Euclidean distance increased test accuracy by a similar amount. K, the number of neighbors to consider, varied widely depending on the test-train split and the exact dataset used, but the best values were usually between 8 and 20. 14 was used for k in the model used for head-to-head comparison with the other models. Other hyperparameters such as maximum leaf size did not have a great impact on test accuracy.

Dataset Selection

Of the “all animals” datasets, the MFCC only datasets resulted in the worst test accuracy. Normalization and noisegating provided almost negligible increases, but at optimum hyperparameters, the test accuracy hovered around 23% - 25%.

However, the full datasets increased accuracy. Of these three sets, the normalized and noisegated one surprisingly performed the worst, around 35% - 38%. Removing noise gating increased test accuracy to around 45%, but removing both noisedating and normalization gave the k-nearest neighbors algorithm its highest and only majority-correct score of 53% test set accuracy. This seems to indicate that there is more value in the additional data for this algorithm than there is in the MFCC data, as the datasets with more weight on these features performed better than those without those features, and better than the datasets that attempted to equalize the importance of these features through normalization.

Of the other dataset types used in head-to-head comparison, the K-nearest neighbors algorithm performed best on the “whale only” dataset, at a maximum accuracy of 75% as seen in Table 5, which is very respectable compared to the other, more complicated algorithms.

3.6 Random Forest

Ensembles are very powerful in machine learning. They combine the predictions from multiple runs of algorithms to classify. They oftentime outperform non-ensemble algorithms. With this and our base understanding, we wanted to see how an ensemble approach would work on our whale classification problem. We expected to see the same trend and have the ensembles outperform our basic machine learning models. We selected to use the random forest classifier. This is an ensemble of the decision tree algorithm. It creates multiple decision trees and makes a prediction by choosing the class that most decision trees would “vote” for.

Overall Performance

As we predicted Random Forest outperformed all the other models we tested. With every variation of dataset and feature selection random forest had the highest test set accuracy. It also achieved the highest test set accuracy overall. We were able to achieve a 91.8% accuracy for correctly predicting if a cetacean was a whale or a dolphin.

3.7 Gradient Boosting

With such success with the random forest ensemble, we employed another ensemble model with hopes that it would achieve even higher accuracy on our test data. We choose to use another ensemble of decision trees: gradient boosting. This ensemble learns each tree sequentially and works to minimize a lost function.

Overall Performance

Unforantanly gradient boosting did not do better than the random forest ensemble; however, it scored higher than most of our other basic models. This shows that ensembles are more powerful than a single model implementation.

4 Results

We played around with different data extraction methods and resulted in six different datasets. We also worked to separate our whale and dolphin data in a way that made the most sense to do classification.

4.1 Initial Results

First we just used our MFCC only dataset composed of all 24 whales and dolphins (12 whales and 12 dolphins) and ran it through all of our models. Our baseline was 1/24 which is 4.17%. While most of the models, excluding the perception, did better than our baseline, our overall test set accuracy was not stellar. Our high test set accuracy score came from our random forest model at 36%.

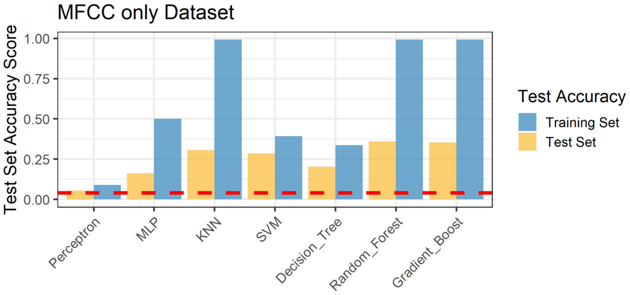

Figure 6: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data but only using the MFCC features.

4.2 Feature Improvements

Seeing as our first results, we concluded that much work was needed to increase our accuracy since every model was returning less than 50% test accuracy. Our next step was to try using noisegating on the MFCC data as a feature improvement. We found that the noisegating the MFCC features did not increase our accuracy, if anything it caused our test set accuracy scores to decrease.

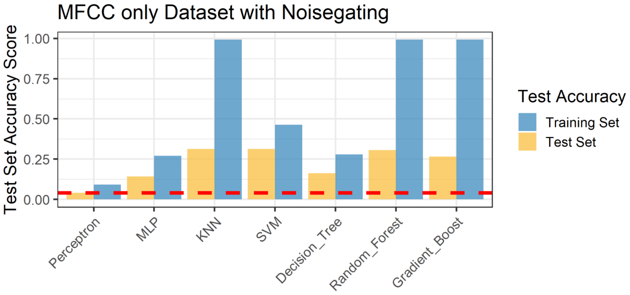

Figure 7: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data but using the MFCC features and noisegating.

We then resolved to collect and add other features to the dataset aside from the MFCCs. Initially we did not normalize any of the features. Even though noisegating was not effective with the MFCC only, we wanted to give this method a second chance. With our new dataset with MFCCs and other features (called fulldata in Table 1), we tried training models on this data with and without noise gating. We saw a huge jump in our test set accuracy with the addition of the other features besides MFCCs.

The highest accuracy score for the full data not normalized WITHOUT noise gating dataset was 78% from the random forest model. The highest accuracy score for the full data not normalized WITH noise gating was 61% from the random forest model. Once again we see that the noise gating decreased the accuracy score instead of improving it.

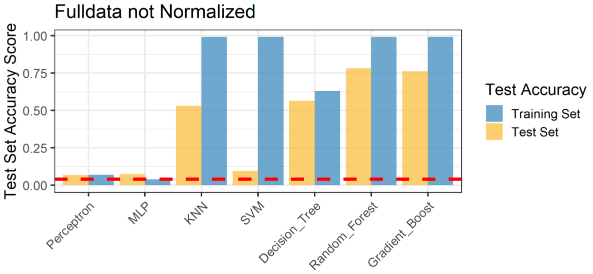

Figure 8: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data but using MFCC features as well as our additional features. We did not normalize them.

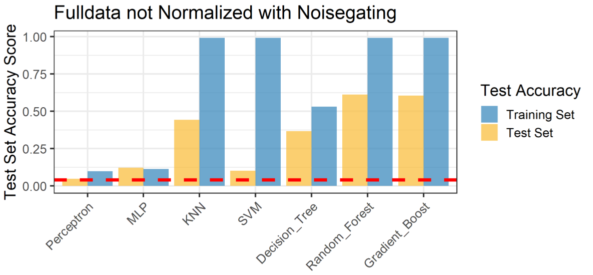

Figure 9: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data but using MFCC features as well as our additional features. We did not normalize them both but we did use noise gating.

Looking at the additional features we added, we found that some of them possessed very large numbers. These large values may prevent higher accuracy scores since the larger numbers would have more weight. To mitigate this problem, we decided to try normalizing all of our features. We applied normalization to both the noisegated and not noisegated versions of our full data set.

Here we found that noisegating made a slight, but not significant improvement. Overall random forest had a 63% test set accuracy for the full data set normalized without noise gating and a 64% test set accuracy for the full data set normalized with noisegating. Gradient boosting actually outperformed random forest on the full data set normalized with noisegating with an accuracy score of 68%.

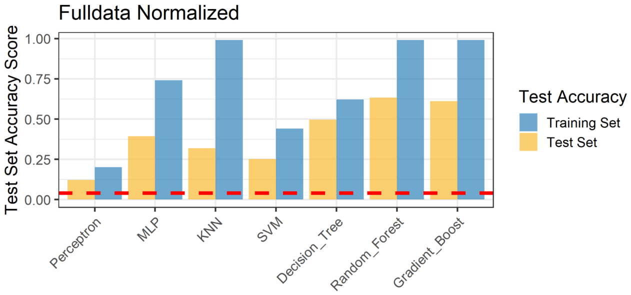

Figure 10: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data. This is for the full data, without noise gating but with normalization.

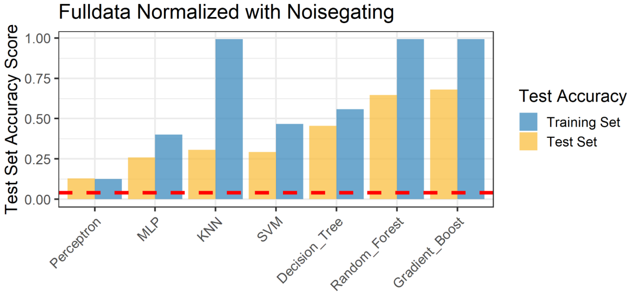

Figure 11: The dashed red line shows the baseline accuracy. This shows data from predicting on a dataset with both whale and dolphin data. This is for the full data, with noisegating and normalization.

Considering all the different test set accuracies, we found that the highest score for the classifying our 24 different species of whales and dolphins was our full data dataset without normalization or noisegating. The random forest model has an accuracy of 78% on this dataset.

4.3 Final Results

Our last step to improve our accuracy was to break our classification problem into different pieces. Up until this point we have just been classifying all 24 of our whale and dolphin species together. Here we show our best test set accuracy we find from the six different feature versions of our dataset when we filter those datasets to include (1) both whales and dolphins, (2) whales only, (3), dolphins only, or (4) two classes: whales or dolphins.

For this table we see that the random forest classifier performed the best (78.2%) followed closely by the gradient boost classifier (76.2%). This was encouraging since the models were trying to classify 24 different species.

Model	**Max test accuracy**
Baseline	0.04167 (1/24)
Perceptron	0.129
MLP	0.395
KNN	0.531
SVM	0.313
Decision Tree	0.565
Random Forest	0.782
Gradient Boost	0.762

Table 2: Dolphin and Whale combined max test accuracy per model

For Table 2 we turn this into a binary classification problem with just whales versus dolphins. We also reach one hundred percent accuracy on the test set. The highest score was the random forest classifier with 91.8% accuracy.

Model	Max test accuracy
Baseline	0.50 (12/24)
Perceptron	0.612
MLP	0.755
KNN	0.844
SVM	0.60
Decision Tree	0.748
Random Forest	0.918
Gradient Boost	0.912

Table 3: Whale vs dolphin max test accuracy scores

For Table 3 we limited our dataset to just include whales. This allowed there to only be twelve different classes instead of 24.

Model	Max test accuracy
Baseline	0.0833 (1/12)
Perceptron	0.382
MLP	0.632
KNN	0.750
SVM	0.574
Decision Tree	0.765
Random Forest	0.897
Gradient Boost	0.853

Table 4: Whale only max test accuracy scores

Model	Max test accuracy
Baseline	0.0833 (1/12)
Perceptron	0.228
MLP	0.418
KNN	0.532
SVM	0.418
Decision Tree	0.620
Random Forest	0.785
Gradient Boost	0.810

Table 5: Dolphin only max test accuracy scores

From these four tables we can conclude that the models performed best when we turned the classification problem into a binary classification problem of whales versus dolphins.

5 Discussion and Conclusion

Through different feature variations, models, hyperparameters, and breaking up our classification problem into smaller or different pieces, we can predict cetacean noises with decent accuracy. From these findings we suggest that the best approach to classify a cetacean is to first determine if it is a whale or a dolphin. From there, it can be determined what species of whale or dolphin.

For processing the audio files, we found it best to use the MFCC algorithm in combination with other sound features such as Tempo Estimation (BPM) or Most Prominent Pitch. Our experiences using these features in conjunction with normalization seemed to indicate that normalization was detrimental to the classification accuracy of any individual model. A quick comparison of the “added” features compared to the MFCCs showed that the “added” features were usually much larger in magnitude. From this information, we determined that it was likely that these extra features were more important to classification success than the MFCCs were. Considering the data loss that occurred in computing the MFCCs, we considered this to be a reasonable assumption.

The ensemble machine learning models, random forest and gradient boost, constantly had the highest test set accuracy. The random forest model outperformed the gradient boost model the majority of the time.

6 Future Work

There were several tasks that we felt would lead to interesting conclusions but were unable to test while working on this project. These tasks are left for future research.

Binary Classification

One of the methods used as an intermediate step to ensure some degree of successful classification was binary classification. Similar to the whale-dolphin split data, this approach consisted of taking each species and comparing it exclusively with another species. This approach was only used as a reassurance when any specific model seemed to be underperforming, and as such was not fully developed in this report. However, it was noted that most models struggled and succeeded with similar species comparisons. No analysis was conducted to attempt to explain these correlations, but future work could consist of performing an exhaustive comparison for each model and researching the difficult-to-categorize species to try and find similarities or explanations.

Whale Geolocation

Another area of research includes the location that the different species of whales frequent (breeding grounds, feeding areas, general habitats). This could be used for at least two different purposes. First, it could be used as a feature that could help any model to differentiate two species that may sound similar but that may be physically separated.

Considering the vastness of the ocean and the tendency of biologists to classify different species that are very physically similar based on migratory patterns, this may help to improve test accuracy for any or all of the models. If this research is to be used to try to classify whale sounds real-time, the matter of geolocation becomes even more useful. Researchers could tag each sample with the location of its recording, which could be used to provide more accurate predictions when testing on novel data in the research field.

Dory Validation Set

Another useful and much more entertaining application of this research is affectionately termed the Dory Validation Set by the research team. This set is thus named because it could include clips of the character Dory (Ellen DeGeneres) from the Disney/Pixar animated feature films *Finding Nemo* and *Finding Dory* to determine the "language" that Dory is "speaking" in the movies.

More productively, this validation set could also include sound clips from other sources than the Watkins Database. In this capacity, tests could be done (with only the "serious" sound clips) and compared with test set accuracy from the Watkins database. This could be useful in determining whether the ML models that have been tested in this report are truly picking up on traits of the whale calls, or if they are merely finding similarities in the static inherent in each recording.

Unfortunately, the time required to get the additional sound clips, clean them sufficiently, and pass them through the validation pipeline was more than the research team could afford to expend, but it is highly recommended to any potential future peer reviewers as a rewarding (and entertaining) use of time and resources.

Deep Learning

Although outside of the scope of this project, deep learning is likely to be more than adequate at classifying this problem. Deep learning is exceptionally well-suited for problems in which features that are close to each other contain valuable information. This approach is already used in image recognition in various applications, to great success.

The spectrograms used to make the MFCCs used in this project are nothing more than visual representations of the sound files being processed. The mel-frequency cepstral coefficients are essentially a numeric representation of the sound, taken from samples of each few milliseconds of sound. The research team found it necessary to average these coefficients across the entire time of each sample in an attempt to extract general pitch and resonance data from the audio files in an easily digestible way, but each one of these processes resulted in substantial amounts of data loss.

Deep learning would allow future researchers to train ML models directly on the entire spectrogram and would remove the need for MFCCs altogether. This could result in the downstream models being able to detect nuances such as the orders and relative durations/intensities of sequential pitch changes, in contrast to current models that could only tell the existence of each pitch in the combined average data. It is postulated that with an effective deep learning stage, even the simple perceptron could return a very satisfactory test set classification accuracy.

Acknowledgments

This work would be entirely impossible without the amazing curation done by those at the Woods Hole Oceanographic Institution and the repository of whale noises they have collected in the Watkins Marine Mammal Sound Database. In addition, without tools and libraries such as LibRosa for feature extraction, or NoiseReduce for audio isolation, we would not have seen the success we have with our models.

References

[Conlin, 2019] Thomas L. Conlin. *The Humpback Song*. Journey North. 2019. . https://journeynorth.org/tm/hwhale/SingingHumpback.html#:~:text=Researchers%20believe%20that%20some%20of,part%20of%20the%20whales'%20songs.

[Davis, 2023] Margaret Davis. *Gray Whale seeks assistance from whale-watching captain in Mexico, unveiling a fascinating interaction between animals and humans.* Science Times. July 08, 2023. https://www.sciencetimes.com/articles/44755/20230708/gray-whale-seeks-assistance-watching-captain-mexico-unveiling-fascinating-interaction.htm

[Goldbogen, 2020] Jermey Goldbogen. *Blue whales*. Curr Biol. 2020 Dec 7;30(23):R1399-R1400. doi: 10.1016/j.cub.2020.10.068. PMID: 33290699.

[*How intelligent are whales and dolphins?,* 2020] *How intelligent are whales and dolphins?.* Whale & Dolphin Conservation USA. April 16, 202. https://us.whales.org/whales-dolphins/how-intelligent-are-whales-and-dolphins/#:~:text=Whale%20and%20dolphin%20brains%20contain,seems%20they%20are%20deep%20thinkers!

[Lepakko, 2023] Zeek Lepakko, *Gigantic Sperm Whale Asks Divers for Help Being Freed from a Fishing Hook.* A-Z Animals. November 10, 2023. https://a-z-animals.com/blog/gigantic-sperm-whale-asks-divers-for-help-being-freed-from-a-fishing-hook/

[McFee, 2015] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. *Librosa: Audio and music signal analysis in python*. In Proceedings of the 14th python in science conference, pp. 18-25. 2015

[Morris, 2023] Amanda Morris. *Are the orcas out to get us? What to know about recent attacks.* Washington Post. July 14, 2023. https://www.washingtonpost.com/wellness/2023/07/14/orcas-boats-attacks-spain/ Amanda Morris

[Pare, 2023] Sascha Pare. *Orcas have sunk 3 boats in Europe and appear to be teaching others to do the same. But why?*. LiveScience.com. May 18. 2023. https://www.livescience.com/animals/orcas/orcas-have-sunk-3-boats-in-europe-and-appear-to-be-teaching-others-to-do-the-same-but-why

[Pedregosa *et al*., 2011] Pedregosa *et al*., *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830, 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[Sainburg, 2020] Sainburg, Tim and Thielk, Marvin and Gentner, Timothy Q. *Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires.* In PLoS Computational Biology Volume 16. 2020

[Sparke, 2023] Trinity Sparke. *Young Killer Whale Seeks Human Help to Rescue Entangled Mother.* OneGreenPlanet.org. August 2023. https://www.onegreenplanet.org/animals/young-killer-whale-seeks-help-rescue-entangled-mother/

[Watkins and Schevilll, 2023] William Watkins, William Schevill. *Watkins Marine Mammal Sound Database*, Woods Hole Oceanographic Institution and the New Bedford Whaling Museum.

[Zeppetelli-Bédard, 2021] Laura Zeppetelli-Bédard. *What do we know about intelligence in whales and dolphins?*. Whale Scientists. June 23, 2021. . https://whalescientists.com/intelligence-whales-dolphins/