Bat Echolocation Classification

Introduction

Bats represent the second most diverse order of mammals and compose approximately 20%, or 1400 species, of all mammal species. Despite this abundance, relatively little is known about bats compared to other mammals due to nocturnal and ultrasonic adaptations placing them out of view and hearing of humans. Within the past ten years, however, passive detection and recording of bat echolocation calls and software analysis of those calls has significantly expanded research opportunities of bat communities and activity. Due to the ultrasonic nature of bat echolocation, humans cannot identify bat calls by sound but a small number of machine learning classifiers have been developed that can identify call sonograms to the species level. Classifiers are extremely important due to the huge number of calls that can be recorded by passive detectors, which can be extremely difficult and time consuming to process manually. For example, at one location in Arizona, 24,000 calls were recorded in one weeks time. The two main classifier softwares, Kaleidoscope Pro by Wildlife Acoustics and SonoBat's classifier can be expensive, prone to error, and not specific to regions.

Project Goals:

1. Determine the accuracy of the Kaleidoscope Pro classification software for a bat echolocation library developed in Arizona, United States
2. Develop a machine learning classifier of bat echolocation calls for the Arizona call library

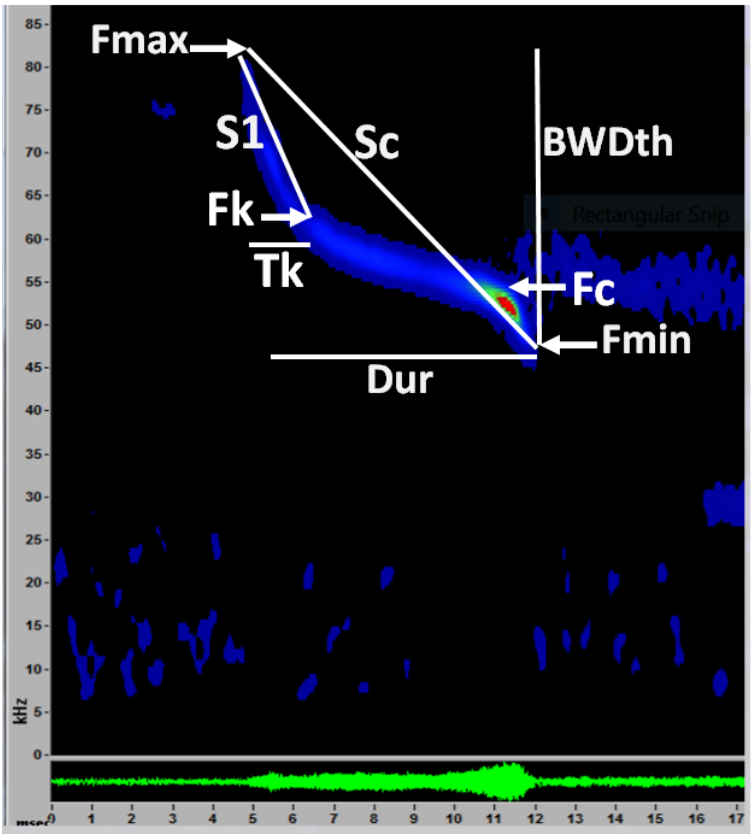
Echolocation Call Libraries Used In this Project

The primary library was developed in Arizona, USA and is composed of summary statistics for 5,849 individual calls. A second library developed in Mexico was also utilized in the beginning stages of the project but was determined to be incompatible with the Arizona Library and so was not used for modeling. The Arizona library was identified to species by the use of Kaleidoscope Pro software and manually identifying each call to species. Manually viewing sonograms to identify species are generally considered to be between 95% and 100% accurate.

Anatomy of a Bat Echolocation Call

North American bat echolocation calls generally range from about 15 kHz to 80 kHz, with 20 kHz being approximately the upper level of human hearing. Every species of bat makes a unique call which can be utilized to identify the bat to species. However, due to the ultrasonic and rapid nature of bat echolocation, calls must be recorded with ultrasonic microphones and identified by sonograms as plotted by computer software. Bats generally make three types of calls, 1. Search phase call which orient bats to their surroundings and help identify prey, 2. Approach phase call that bats use to zero in on prey, and 3. Feeding buzz calls that take place in rapid succession and are used to capture prey. By far, the vast majority of the time bats spend in flight they are making search phase calls with approach phase and buzz calls being made only a small portion of the time. Search phase calls are generally unique to individual species while approach phase and buzz calls are often very similar between species. For this reason, only search phase calls are used to positively identify bats to the species level. In real

world recording, however, removing approach and buzz calls from sonograms is extremely difficult.



Ten different summary statistics of a bat call sonograms were contained in the Arizona Library. The measures of the call are as follows: Fc, Sc, Dur, Fmax, Fmin, Fk, Tk, S1, BwDth, PkDur. Refer to the labeled sonogram above to see exactly what each of these measurements are.

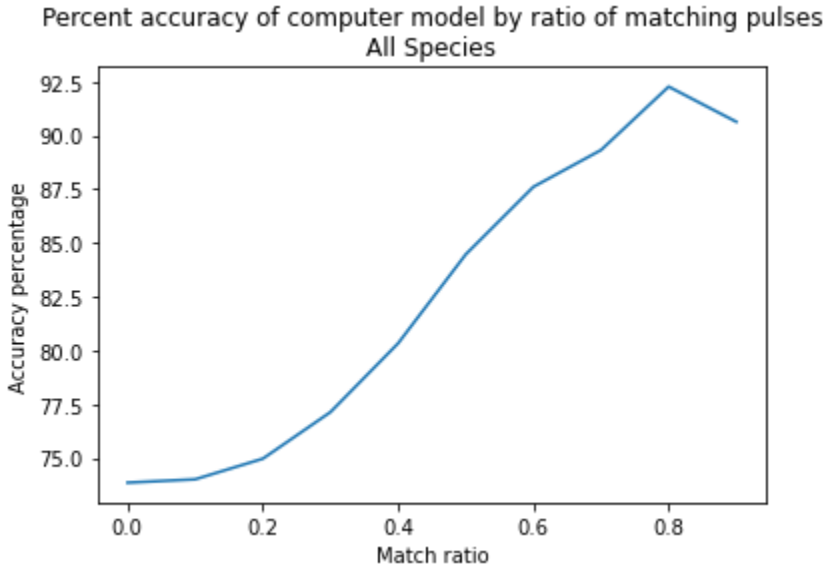
Definitions of Bat Echolocation Call Features

- Fc: Central frequency of the loudest portion of the call. Also known as the characteristic frequency of the call.
- Sc: Overall slope.
- Dur: duration in milliseconds of the call.
- Fmax: Maximum frequency of the call.
- Fmin: Minimum frequency of the call.
- Fk: Frequency of the knee, or the portion of the call with the largest change in slope.
- Tk: Duration of the call from start to the knee (Fk) of the call in milliseconds.
- S1: Slope of the first half of the call.
- BwDth: Range of the highest to the lowest frequency of the call.
- PkDur: Percent duration of the knee of the call to the entire call duration.

Machine Learning Methods for Identification of Sonograms to Species

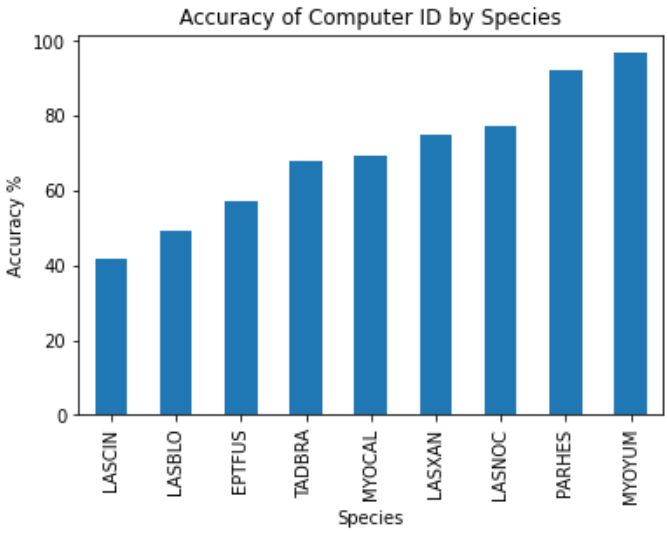
Sonograms of bat echolocation call recordings generally range from a single to potentially hundreds of individual calls that take place in rapid succession(also known as pulses). Software identifies individual calls in a recording and averages each of the above call features across the sonogram. In certain machine learning models each individual call is identified to a species. In a particular sonogram, it is typical that not all calls match well with a particular species. However, a sonogram is identified to the species with the highest number of overall matching calls. A second type of machine learning model identifies averages of each of the call features in a single sonogram. These averages are then utilized to identify the species to the sonogram.

Accuracy of Kaleidoscope Pro ID for Arizona Bat Call Library



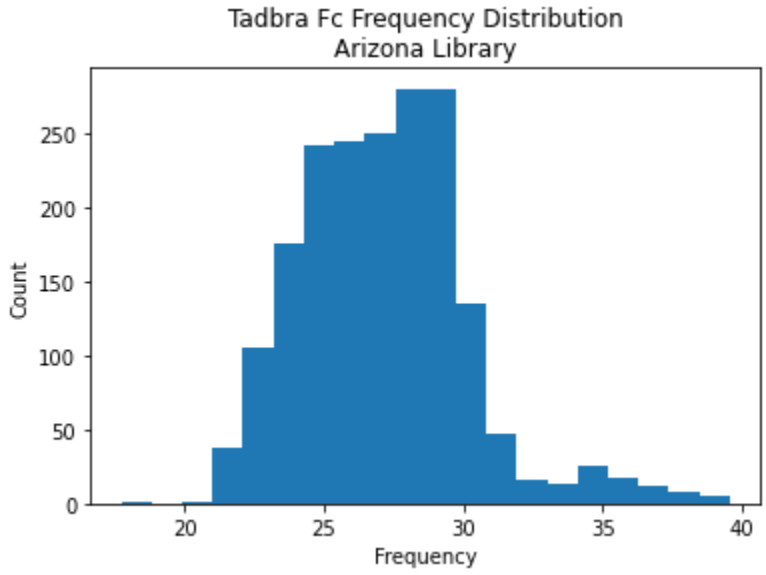
Machine learning models are extremely important in the identification of bat echolocation calls due to the large volume of calls collected in-field and the amount of time it takes to identify each individual call to a species level. The machine learning model for Kaleidoscope Pro classifies sonograms to species by the species with the highest number of matching calls within the sonogram. The species with the highest match ratio is then identified to the call. It would be expected that as a match ratio of a sonogram increases, the accuracy of the model would increase. For all species, it was determined that Kaleidoscope Pro was approximately 68% accurate for all match ratios. As the match ratio increased to 0.8, positive identification increased to 92%. A further increase to a 0.9 match ratio caused a slight drop in accuracy to 90%. This slight drop in accuracy is most likely due to sonograms with lower numbers of total calls and the tendency of noise to have a stronger influence on a smaller number of calls. These data indicate there is a lot of room for improvement in automating call identification. In this project, it will be determined whether machine learning can be applied to help improve this capability within the field.

Accuracy of Kaleidoscope Pro ID for the Most Common Species



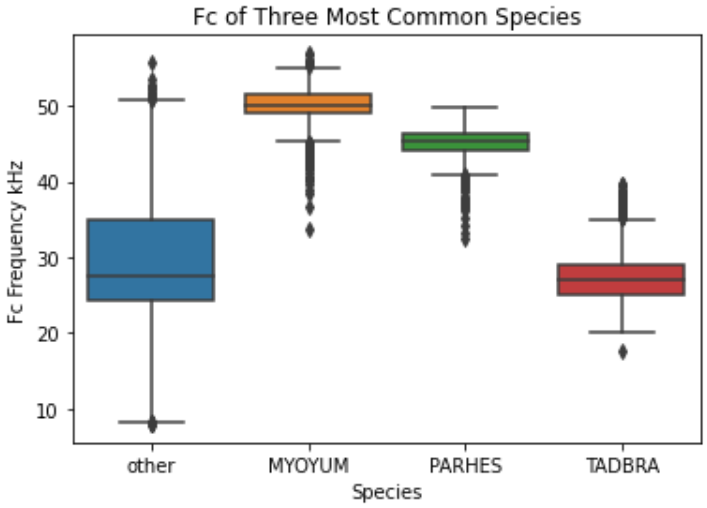
For the nine most common species within the Arizona library, accuracy of the Kaleidoscope Pro ID was dependent on species (see above figure). Differences in accuracy by species is most likely due to uniqueness of a species call. It would be expected that species with the most unique call features would have a higher accuracy than species with less unique features.

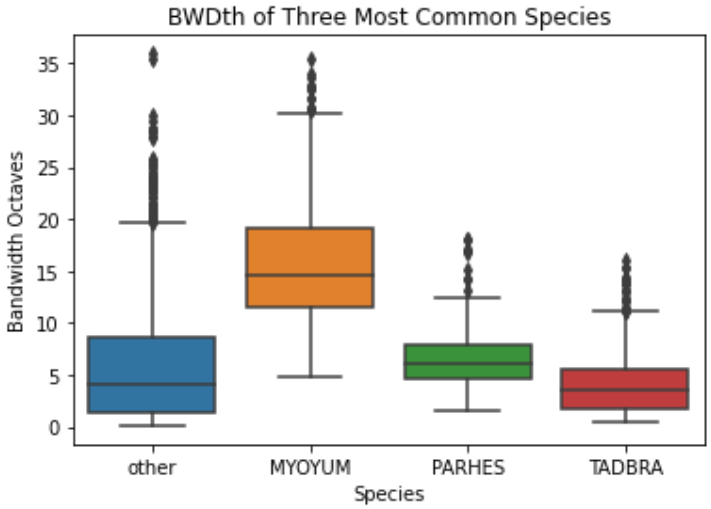
Distribution of Data for Arizona Bat Call Library

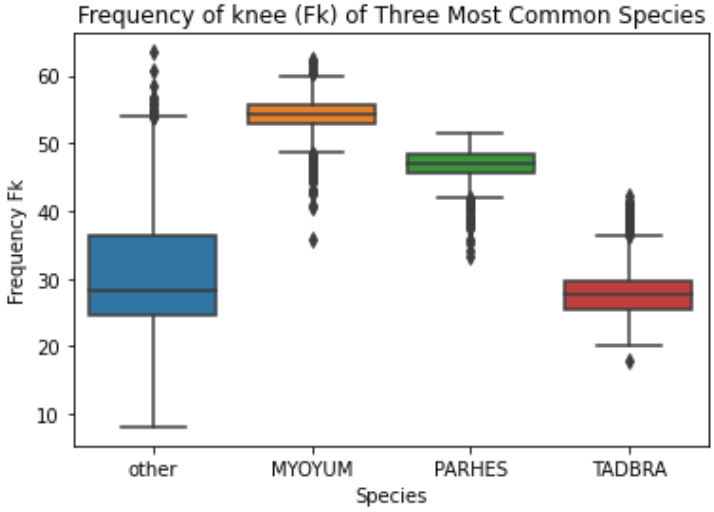


The above histogram demonstrates non-normal distributions for the Fc for Tadbra. Non-normal distributions typically demonstrated right skewness, such as seen above. In order to test normality, a Shapiro-Wilk Test was carried out for each species for each call feature. This test demonstrated that all call features for all species had non-normal distributions. All p-values were < 0.05.

Call Features of Three Most Common Species



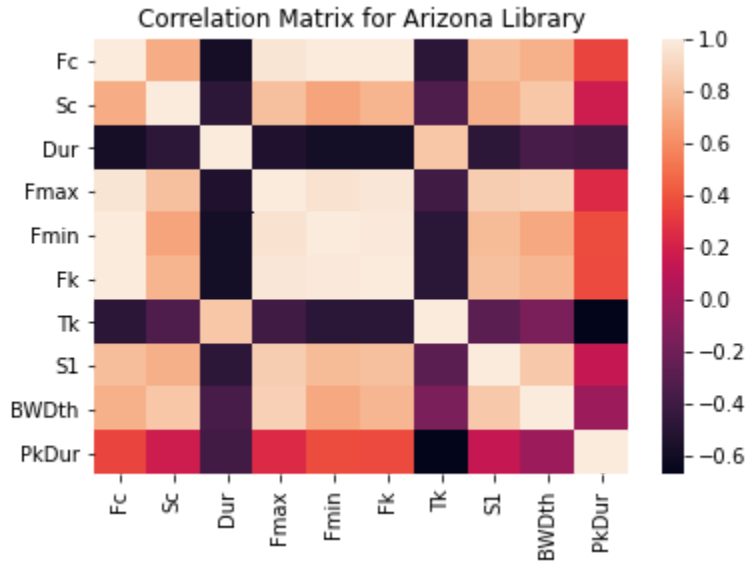




The above graphs demonstrate the Fc, BWDth, and Fk for the three most common species. Due to non-normal distributions as demonstrated by the Shapiro-Wilk Test, a Kruskal-Wallis test for each feature of the three most common species was carried out. The three most common species were Tadbra, Myoyum, and Parhes. An ANOVA test comparing each individual call feature between these species was also carried out. Both Kruskal-Wallis and ANOVA demonstrated a statistically significant difference ($p < 0.05$) between the three most common species for each of the ten call features. The below table displays ANOVA and K-W test scores and p-values for each of the call features.

Feature	ANOVA (Test score, p-value)	K-W Test (Test score, p-value)
Fc	(20077.744059978777, 0.0)	(2394.8620088675666, 0.0)
Sc	(5105.901157284977, 0.0)	(1747.7612246876329, 0.0)
Dur	(2265.649039462254, 0.0)	(1963.1640421742625, 0.0)
Fmax	(11418.705458036427, 0.0)	(2409.0727201866084, 0.0)
Fmin	(20838.43486293496, 0.0)	(2355.623793437008, 0.0)
Fk	(20494.94029736887, 0.0)	(2411.1383511262643, 0.0)
Tk	(1376.0166825576352, 0.0)	(1622.2487693440598, 0.0)
S1	(3360.4762474392587, 0.0)	(2049.163163532451, 0.0)
BWDth	(2982.0052347446185, 0.0)	(1812.0126850017564, 0.0)
PkDur	(746.6649638402916, 1.849247775160946e-266)	(974.6976323327767, 2.223813574255156e-212)

Feature Correlation Matrix



The correlation matrix for the Arizona Library demonstrates a significant amount of high correlation between call features. This is likely because the majority of call features utilize frequency as a dependent variable. For example, as Fc increases across species, Fmin, Fmax,

and Fk will also increase. In other words, as Fc increases, the overall frequency of all other frequency measures increase also. The two slope features, Sc and S1, also have a somewhat high correlation with dependent variables measured by frequency. This is likely because species with higher frequency calls have calls with greater slopes. Time dependent variables such as Dur and Tk were found to have strong negative correlations for all features.

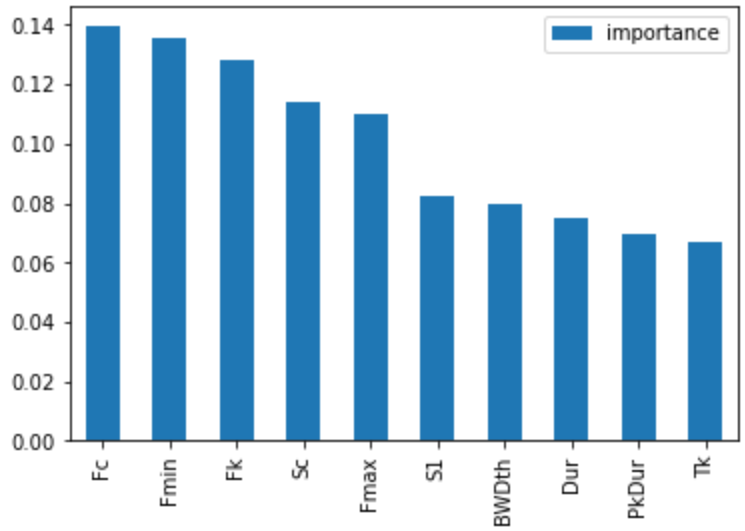
Feature Importance

VIF scores determined that frequency features of bat call sonograms had high multicollinearity. For this reason, Fmin, Fmax, and Fk were removed and only Fc be retained due to high multicollinearity. Fc was retained due to it being the loudest portion of the call, making it the most likely to be consistent between call sonograms. Other features were retained for model building.

Logistic regression feature importance was carried out for the three most common species with Fmin, Fmax, and Fk removed. With unscaled data, for Tadbra, Dur had the highest significance; BDWth for Myoyum, and Tk for Parhes. Fc was found to be second in importance for both Myoyum and Parhes. With scaled data, Fc had the highest significance for Myoyum and Parhes, and Dur for Tadbra. The chart below demonstrates the log-odds for each of the features for each species. The high log-odds for Myoyum and Parhes for Fc is related to the high frequency of their calls. Myoyum had the highest log-odds for Sc due to having the greatest slope of the three species. Parhes had high log-ods for Tk and PkDur due to having the longest duration of the knee of their call. Myoyum had the greatest BWDth for their call, corresponding to the highest log-odds for that feature.

	MYOYUM	PARHES	TADBRA
Intercept	-1.000	-1.000	2.409
Fc	0.444	0.521	-0.027
Sc	0.018	-0.095	0.010
Dur	0.404	0.050	0.288
Tk	-0.807	1.080	-0.161
S1	-0.002	0.015	-0.007
BWDth	0.178	-0.387	-0.017
PkDur	-0.047	0.078	-0.043

Random forest feature selection told a slightly different story. When high multicollinear features were not removed Fc, Fmin, Fmax, Fk, and Sc were the top five features for importance for both the three species data and all species data. With high multicollinear features removed, Fc, S1, Sc and BWDth were the features with the highest importance for both three species and all species data. The below graph displays the random forest feature importance.



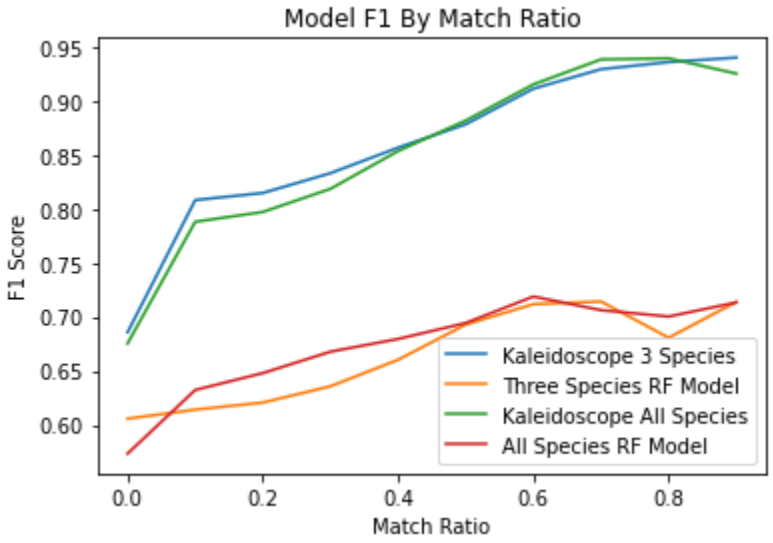
EDA Conclusion

The overall accuracy of Kaleidoscope Pro software for identifying bat species was approximately 68%, leaving significant room for improvement in model development. Due to the disproportionate number of calls between species in the Arizona Library, it is suggested that a model should only be developed for the three most common species, TADBRA, MYOYUM, and PARHES. Models for all species could also be developed for comparison. Strong multicollinearity of frequency features was found and it was determined that Fc should be retained while other frequency features removed for modeling.

Modeling

Machine learning modeling was then carried out in order to develop a model to classify echolocation call data to the species level. Random Forest, KNN, XGBoost, and Multinomial Logistic Regression were used to develop models. Data used in these models was either scaled or unscaled, with or without high multicollinear features, and for all the species present in the data set and with only the three most common species. Grid Search was carried out on each of these models to determine the best hyper parameters. Results of the best performing model for each classification are in the table below. Of the tested models, Random Forest performed best for both all species and three species datasets.

Model	All Species Score	Three Species Score	Data Scaled	Multicollinear Data	Best Hyperparameters
Random Forest	0.6629	0.7613	No	Not removed	Criterion: gini Max depth:11 Max features: auto N estimators: 220
Knn	0.6426	0.7421	Yes	Removed	Distance: Manhattan Weight: distance N neighbors: 19 (all species) 16 (3 species)
XGBoost	0.6507	0.7579	No	Not removed	Learning rate: 0.05 Max features: 3 (all species) 2 (3 species) N estimators: 20
Multinomial Logistic Regression	0.5962	0.6695	Yes	Removed	C: 0.001 Penalty: Ridge reg. (all species) None (3 species)



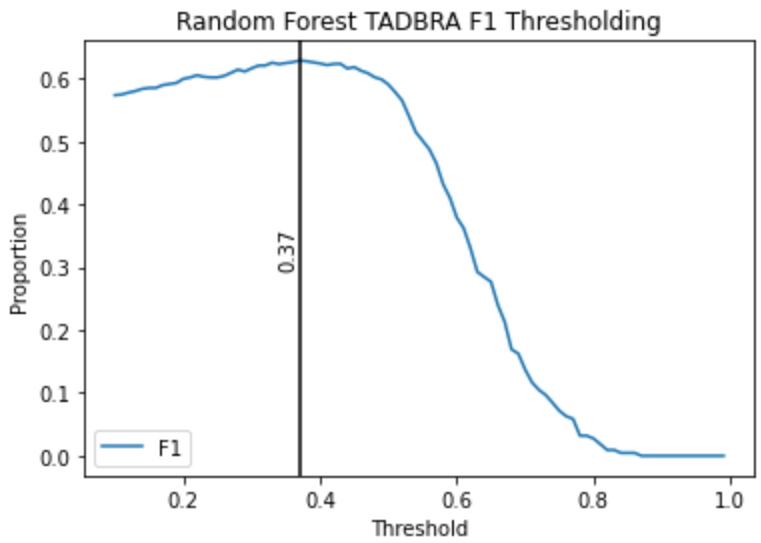
The three species and all species Random Forest models produced very similar F1 scores by match ratio. Kaleidoscope Pro software also produced similar F1 scores for all species and three species. Overall, the Kaleidoscope Pro software out performed the Random Forest models produced in this project. To further examine this, the models were examined by the three most common individual species.

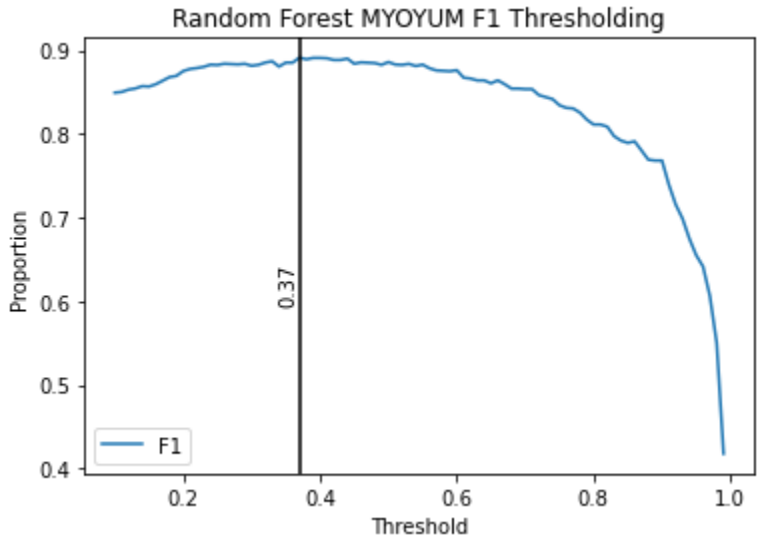
Random Forest models were also created for the three most common species, each individually. Grid search validation was carried out to determine the best hyperparameters for each model and the ROC-AUC score was used to determine the best performing hyperparameters. Data for these models was all unscaled and all features were retained. Results of the models after hyperparameter tuning can be seen in the table below.

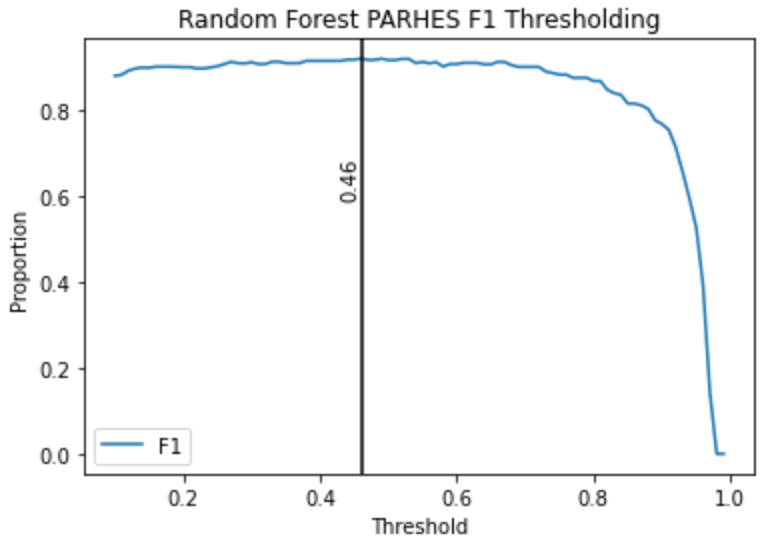
Single Species Random Forest Model	**ROC-AUC Score**	**Accuracy Score**	**Random Forest Hyperparameters**
TADBRA	0.854	0.797	Criterion: gini Max depth:11 Max features: auto N estimators: 400
MYOYUM	0.992	0.965	Criterion: gini Max depth:10 Max features: auto N estimators: 100
PARHES	0.998	0.985	Criterion: gini Max depth:6 Max features: auto N estimators: 600

Thresholding was carried out to improve for each of the individual species models in order to improve F1 scores. The results of the thresholding can be seen in the table below. By adjusting threshold, F1 scores were able to be optimized between 0.004 and 0.038 for the three species. Results of thresholding can be seen in the table and graphs below.

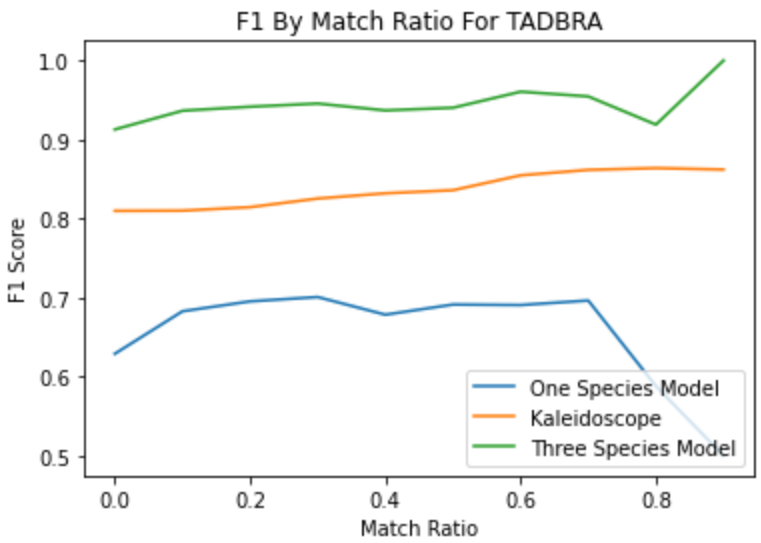
Random Forest Model	**Default F1 Score**	**Optimized F1 Score**	**Optimized Threshold**
TADBRA	0.591	0.629	0.37
MYOYUM	0.886	0.891	0.37
PARHES	0.918	0.922	0.46

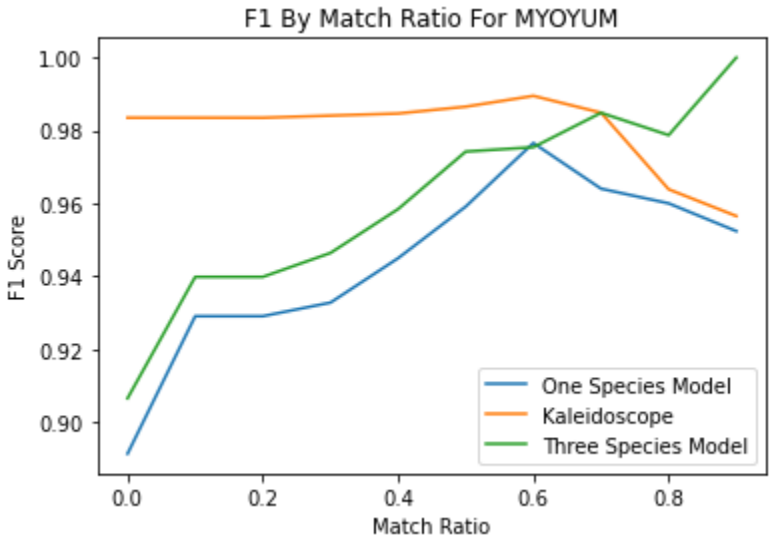

Random Forest TADBRA F1 Thresholding
Proportion
Threshold
0.37
F1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.2
0.4
0.6
0.8
1.0

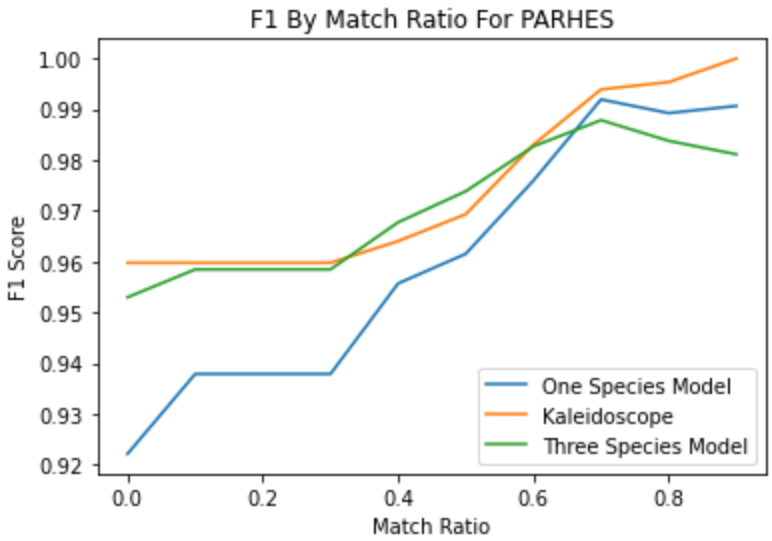

Random Forest MYOYUM F1 Thresholding
Proportion
Threshold
0.37
F1
0.4
0.5
0.6
0.7
0.8
0.9
0.2
0.4
0.6
0.8
1.0



Below F1 scores for different match ratios were graphed for models developed for the three most common species in the data. These models were plotted in comparison to the Kaleidoscope Pro software identification. It was found that Kaleidoscope Pro software was more accurate for both PARHES and MYOYUM. However, the three species Random Forest model had higher F1 scores across all match ratios for TADBRA. While it is not exactly known why the three species model performed so much better than the Kaleidoscope model, it can be noted that F1 scores for Kaleidoscope were much lower compared to the other two species, allowing for significantly more improvement in the model. Additionally, TADBRA has a variable call that could be difficult to classify, especially if there are a smaller number of calls in the dataframe. With over 1500 calls in the training data for TADBRA, it is possible that this number of calls had better coverage of the variability this species produces in call features. For all three species, the one species models had overall the lowest F1 scores. Kaleidoscope and the three species model performed similarly for all but the highest match ratios for PARHES, possibly because of the low number of calls with very high match ratios in the database from with the three species model was produced. It can also be noted that F1 scores dropped slightly for the highest match ratios for a number of the models. While it is not known exactly why this happened, it is a possible result of there being relatively few calls which a very high frequency of match ratios.


F1 By Match Ratio For TADBRA
F1 Score
Match Ratio
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
One Species Model
Kaleidoscope
Three Species Model


F1 By Match Ratio For MYOYUM
F1 Score
Match Ratio
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
One Species Model
Kaleidoscope
Three Species Model



Summary and Conclusion

Bat echolocation research produces a large amount of data that takes a significant amount of time to process for downstream analysis. Doing so manually can take many hours to days and weeks of work. However, the development machine learning classifiers can reduce the amount of time required for processing data down to a matter of minutes. However, machine learning classifiers are generally not as accurate as manual processing. In this project it was determined that Kaleidoscope Pro software was overall 68% accurate for the species examined, with certain species having overall higher accuracy than others. The goal of this project was to produce a machine learning classifier model that would improve on this accuracy. Random Forest, KNN, XGBoost, and multinomial Logistic Regression models were developed for all of the species, as well as the three most common species, in the dataset. Random Forest was determined to produce the model with the highest accuracy as well as F1 score. However, this model did not improve on the overall accuracy and F1 score of Kaleidoscope. Accuracy and F1 scores were also examined for the three most common individual species for the three species Random Forest model. Random Forest models were also produced for each individual species of the three most common species. TADBRA from the three species Random Forest model was the only species that outperformed Kaleidoscope. Individual species Random Forest models consistently underperformed all other models. TADBRA is the most common species in the southwestern United States and plays an extremely significant role ecologically. While the best performing model produced in this project did not outperform Kaleidoscope, the Random Forest model produced here can be utilized in conjunction with Kaleidoscope inorder to better identify this particular species. The practice of utilizing mu

Future work

1. Future work on this project should involve the development of a more balanced dataset for the purpose of developing models. Only three species in the dataset had over 500 samples, with the remaining 14 species having less than 500 samples. This likely had a significant impact on how well a model could be developed.

2. Models could be developed by removing NoID from the dataset. Then, by thresholding the individual species in the model, if a certain threshold is not met then the model would ID the call as NoID.
3. Feature development for this project was dependent on Kaleidoscope Pro software. In the future it would likely be beneficial to utilize Python sound analysis software to develop our own features.
4. Models determined to be useful could be deployed as web apps.