
# Passive Acoustic Monitoring of Blue and Fin Whales through Machine Learning

**Daniel De Leon, Cabrillo College**

***Mentors: Danelle Cline, Dr. John Ryan***

***Summer 2017***

**Keywords: convolutional neural network, machine learning, transfer learning, fin whale pulse call, blue whale b call, hydrophone, Raven, spectrogram**

**Abstract:**

With an abundance of audio data collected from Monterey Bay Aquarium Research Institute's hydrophone, we plan to implement a convolutional neural network (CNN) machine-learning algorithm that can effectively classify blue whale b-calls and fin whale pulse-calls. Using Raven, a bioacoustics analysis software program, we have extracted a total of 16,886 false-positive and true-positive blue whale b calls and 5,927 false-positive and true-positive fin whale pulse calls from the hydrophone's audio recordings. These short sound files were converted to spectrograms and be pre-processed through a series of image enhancements prior to training and testing the CNN. 70% of the data was used to train the network, 10% was used for validation and 20% for testing. Analysis on the effectiveness of the classifier was based on quantitative results presented from ROC curves, confusion matrices, and accuracy and precision calculations. Qualitative analysis was also done by a long-term comparison of the CNN classifier outputs with respect to results from long-term spectral average graphs that illustrate whale call abundance throughout the first two years of recordings. The blue whale b classifier results show an overall 98.3% prediction accuracy and the fin whale pulse classifier results show an overall 97.8% prediction

accuracy. Future work will be to add in detection methods as a precursor to the CNN classifier. Together, this detection and classification workflow can be used to monitor whale calls synchronously with the hydrophone to give scientists valuable data regarding whale migratory behavior. With ocean events such as rises in sea temperature that have caused an ecological disruption, listening to the soundscape could bring insight into the effects of physical disturbances.

**Introduction:**

The rise in ocean water temperature over the past few years in the North East Pacific, known as the 'Warm Blob,' has caused an ecological disruption amongst the inhabitants (Peterson et al. 2015). Irregular sightings of both blue and fin whales have been recorded over the past couple years at the Monterey Bay Aquarium Research Institute (MBARI) and there is reason to believe that this could be due to the displacement of phytoplankton.

To further discover the effects of this 'Warm Blob' on the marine ecology, much effort has been put into monitoring population density. Even though visual data collection methods are reliable sources of evidence, they can be especially difficult to obtain when dealing with illusive migratory whale patterns and behaviors. A different approach of monitoring whale population density and pod migration is through passive acoustic techniques. As of July 28th, 2015, MBARI has been powering and continuously recording from a hydrophone that is connected to their Monterey Accelerated Research System cabled observatory in the central canyon region of the Monterey Bay Marine Sanctuary (Ryan et al. 2016). With a bandwidth ranging from 10 Hz to 200 kHz and a sample rate of 256 kHz, MBARI's hydrophone was deployed to begin analyzing the soundscape and how anthropophony may interfere with the specific frequency bands that whales and dolphins occupy for the channeling of communication and echolocation.

The blue whale has a vocalization known as a B call that is a emitted lower the 20 Hz and can reach a duration of at least 15 seconds. The second harmonic of this call, however, carries much more energy than the fundamental frequency and exists between 41 and 48 Hz (see figure 1). The fin whale emits a short pulse between 15 and 30 Hz and can last a duration of at least 1.5

seconds (see figure 2)(Debich et al. 2015). This makes the two species great candidates to acoustically monitor given the physical fact that lower frequency sound can travel through media much further. With the North East Pacific Ocean being home to the largest population of blue whales, MBARI's continuously powered hydrophone is advantageously positioned to be a useful tool in observing whale communication and activity. Since the hydrophone's installation off the coast of Moss Landing, Dr. John Ryan generated the figure 3.b that plots acoustic index VS time in months, where acoustic index represents a monthly average sound energy at a given frequency. The blue and fin lines represent the middle frequencies of the blue whale b (BWB) call and fin whale pulse (FWP) call. Figure 3.c (also generated by Dr. John Ryan) shows depth in meters VS the same time axis as figure 3.b. The color represents a change in ocean temperature at the M1 mooring located in the Monterey Bay (36.75 N, -122 W). In comparing the two plots, it can be noted that in the same months that there is a large rise in sea temperature, there is a peak in acoustic energies at the FWP and BWB frequencies. It can also be noted that there is a two-month offset between FWP and BWB peeks. It would be insightful to be able to quantify the average sound energy by having an actual count of these calls.

With the large amount of live-stream data that the hydrophone collects in real time, however, manually examining all of the audio would prove to be too cumbersome of a task. Such an obstacle then begs the question, how effective is a machine-learning algorithm in classifying blue and fin whale calls? There have been many recent developments on approaches of detection-classification (DC) systems of whale calls. Popular approaches that have proven to be successful are those which involve pattern recognition of processed imagery such as the spectrograms seen in figures 1 and 2. Kaggle, an international competition platform for predictive modeling in data science, held a contest that asked the machine learning community for the best algorithm to predict if there is a right whale upsweep call in a given sound clip ("The Marinexplore and Cornell University Whale Detection Challenge | Kaggle", 2017). Of the 245 teams in the competition, those who used convolutional neural networks (CNN's) were seen to have placed amongst the highest (Dugan, A. J., et al. 2015). This gave us enough reason to believe that a CNN would prove to be an effective model in classifying spectrograms images of both BWB calls and FWP calls.

Ultimately, the purpose of the project is to see if an autonomous DC system can be implemented to successfully quantify BWB and FWP calls. Our goal was to develop and train an accurate CNN classifier so that the detector parameters could be vague and lenient enough to pick up as many potential calls as possible. Having an effective DC system would then lead to implementation of a pipeline that connects the live-streaming, decimated acoustic data from MBARI's hydrophone to the detector to extract sound clips that represent potential FWP and BWB calls. These sound clips would be converted to spectrograms and enhanced to then be classified by the CNN autonomously and continuously as long as the hydrophone is recording.

**Methods:**

There were three stages that made up the entirety of the project. The first stage consisted of collecting sound clips of manually classified BWB and FWP calls through the detector. The second stage involved converting the sound clips into clear spectrogram images. The third and final stage was dedicated to choosing the CNN model and training and testing on the BWB and FWP spectrogram images.

1. Data Collection

In order to begin collecting any calls, we first had to decimate the audio from 250 kHz down to 250 Hz to reduce computational complexity when processing it. We chose to decimate all the audio recorded in the months of November 2015 and 2016. A couple weeks in September 2015 were also decimated.

We chose to use Raven's Band Limited Energy Detector (BLED) to detect the FWP and BWB calls throughout this decimated audio. Raven is a sound analysis software program developed at the Cornell Lab of Ornithology to study bioacoustics. The BLED is a sliding window-box detector that has a set of parameters that describe the targeted signal. Among the many parameters, the ones we focused on setting and tuning were:

- Min. and max frequency
- Min. and max duration
- Min. separation

- Min. Occupancy (%)
- Signal to noise ratio (SNR) threshold (db)
- Block size
- Hop size
- Percentile

Description of each parameter can be found in Raven's user manual (Charif et al. 2010). The parameters that were set for detecting FWP and BWB calls can be seen in table 1. We also changed the window size spectrogram parameter that adjusted the fast-Fourier transform operation within Raven. This made the spectrograms clearer to see depending on certain frequencies. A lower sample per window rate resulted in a smearing of the frequency axis and was seen to be advantageous when dealing with the FWP calls because of their long frequency band and short duration. A higher sample per window rate showed a smearing in the time axis and made BWB calls much clearer given their long duration and smaller frequency band. The windows size for the BWB call was set to 512 samples per window and the FWP call was set to 128 samples per window. After configuring the BLED and running it across the decimated audio with a 50% overlapping sliding window, we manually labeled every single detection as either bt (true) or bf (false) for BWB calls and ft (true) or ff (false) for FWP calls. An example of the labeling for the FWP calls can be seen in figure 4. For every day that the BLED ran on, a table was generated that had the time, duration and true or false label for every detection. For that same day, sound clips of each detection were generated as wave files along with a text file that was a list of every filename of every sound clip. All BWB calls were temporally padded by 5 seconds in case the detector failed to encompass the entire call. For the same reason, all of the FWP calls were padded by 3 seconds. A total of around 16,886 and 5,927 BWB and FWP sound clips were extracted, respectively.

## 2. Pre-Processing

Once we felt that we had enough sound clips we began to write the necessary code to generate the spectrograms. Most of the code we used to create the spectrograms was inspired from the winning team from the Kaggle competition mentioned earlier (Kridler et al. 2013). We followed a tutorial of their code via Jupyter Notebook (Helgren et al. 2016). Python3 was the language used in conjunction with all of the libraries listed in the requirements.txt file found in our PAM

repository (Cline et al. 2017). A few different interpolation and normalization methods were attempted to enhance the signal appearance. Subjective inspections led to use a bi-linear interpolation operation provided by the matplotlib library. A filter array was also used to convolve through the whole image to create a smoother spectrogram. Much emphasis was put into the appearance of the spectrograms because of the insight shared by former Kaggle winner, Nicholas Kridler, who focused on enhancing the training images rather than changing his CNN algorithm. This pre-processing step ultimately converted the wave sound clip files, provided by Raven, into 300x300 jpeg images as seen in figures 1 and 2. All false and true spectrograms that were labeled from Raven were separated into different directories.

## 3. The CNN

With the short amount of time left before the deadline, we decided to apply transfer learning which involves using a pre-existing network that has already been previously trained on different image data. By adding a series of extra layers to the end of the network and training only those layers, we can redirect the network to classify BWB and FWP spectrograms. This can be advantageous if the network that is being transferred is highly trained because it doesn't require as much training. The CNN we decided to transfer is Google's Inception V3 deep learning network (Szegedy et al 2016) in conjunction with the TensorFlow toolkit. The transfer learning was inspired by a TensorFlow tutorial that walks through training and testing the last layers ("How to Retrain Inception's Final Layer for New Categories | TensorFlow ", 2017). Code for all implementation of Inception V3 can be found on our PAM repository.

Training and testing of the final layers of the network was split into 70% training data, 10% validation data and 20% testing data. Because of the fact that the FWP and BWB calls had different samples per window it was required to separate them during training and testing so that two independent classifiers could be developed. We initially started with two classes for the network to train on (true and false). After running the model the first time we realized that we needed to increase the number of classes in order to be more specific so that the network could recognize the patterns more effectively. We split all the BWB spectrograms into three classes: bt, bf_unk, bf_lines (see figure 5). We split all the FWP spectrograms into four classes: ft, ff_unk,

ff_lines, ff_blue (see figure 6). Creating more classes made the data more specific but we had fewer images for each class. Thus, we had to revisit Raven and generate more sound clips (we specifically needed to find more false positives detections). To do this we ran the BLED on new decimated audio and lowered SNR parameter down to 4.5 and 2.0 for the BWB and FWP calls, respectively. We only augmented the false classes to avoid creating a lopsided network. We added around 1,030 bf_unk spectrograms and 80 bf_lines spectrograms. We also added 1,627 ff_unk and 119 bf_lines spectrograms.

Once we completed training the model we then ran the model on newly decimated data between the months of August and December to compare the number of calls found by our new DC system with figure 1.

**Results:**

After the first training of the CNN classifier with only two classes (true and false) for both BWB and FWP calls, overall accuracy and precision are plotted in 7.a and 8.a. Specific accuracy and precision for each class are plotted in 7.b and 8.b. Accuracy and precision were calculated for the classes as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \cdot 100\%$$

$$Precision = \frac{TP}{TP \quad + FP} \cdot 100\%$$

After adding more classes to the classifier and augmenting data, figure 9.a and 9.b show overall accuracy/precision and specific accuracy/precision for each a class for the BWB call. Figure 10.a and 10.b show the ROC curve and confusion matrix for the BWB call as well. Figures 11 and 12 show the same as figures 9 and 10 except they represent results for the FWP call.

The ROC curves show true positive rate (TPR) vs false positive rate (FPR) where:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The confusion matrices plot the number of misclassifications and where they were misclassified. The y axis represents the correct classification and x axis represents the predictions. The color weights show how many were predicted for each classification. A diagonal line of dark blue through the middle as seen in figure 12 shows that the FWP classifier was not confused.

Figure 13 shows the results of the final model run over all the decimated audio of the months of August through December. The y axis represents the monthly average of calls and the x axis represents each month. This graph was used to compare with Dr. John Ryan's graph in figure 3.b.

Table 2 includes the amount of spectrograms collected for each individual class.

**Discussion:**

The results show that the application of a CNN to classify BWB and FWP calls can be a viable option when dealing with such big data from the hydrophone. It also shows that transfer learning of a deep CNN such as Inception V3 is a great option for spectrogram pattern recognition. In comparing figures 7 and 8 to figures 9 and 10, it can be seen that augmenting the data and creating more classes greatly improves the accuracy and precision of the classifier. Qualitative analysis comparing figure 13 and 3.b shows that counting calls with Raven's BLED combined with the CNN classifier is working. The average number of calls per month reported by the DC system is proportional the monthly acoustic index. This supports the possibility of a relationship between the rise in both sea temperature and whale vocalizations in the Monterey Bay.

Nonetheless, there is still room for improvement. If there was more time to augment training data, the network could have greater precision and accuracy in each individual class. Because the data is heavily lopsided toward having more true calls for both BWB and FWP calls, the overall accuracy and precision graphs do not appropriately portray the actual behavior of the network very well. In order to prevent this in the future, the CNN needs to have more examples of false calls. To do this using the BLED, lowering the SNR threshold parameter would increase the

number false positives detected by the BLED and would give the CNN more examples to recognize noise patterns.

Other future work that can be built off of these findings could include understanding where the background noise is coming from. Many of the misclassifications from the CNN were from having multiple classes in one spectrogram. For example, a line of noise could appear right next to quite BWB call and the CNN would classify it as a ff_lines class. We initially thought these lines were coming from ship noise but further inspection revealed that these lines were appearing during parts of the day where the hydrophone did not pick up any signs of ship activity. With the DC system in place, finding when and how many of these lines show up in the audio data could give insight into where they are coming from.

There was also a noise band that existed around 30 Hz that was interrupting many of the FWP calls. It made it more difficult to detect the FWP's because the SNR had to be increased. This explains why we collected less FWP training data all together. Further investigation of this noise could lead to either an effective band filter or to be able to silence it if it were to be a mechanical sound.

Once a fine tuned DC system can be set-up to begin counting individual whale calls over continuous live-streaming data, more research can be dedicated to investigating whale call separation time. Finding temporal patterns within a series of BWB calls, for example, could help behavioral ecologists translate what the largest animals on the planet are communicating to each other.

## Figures

Figure 1.

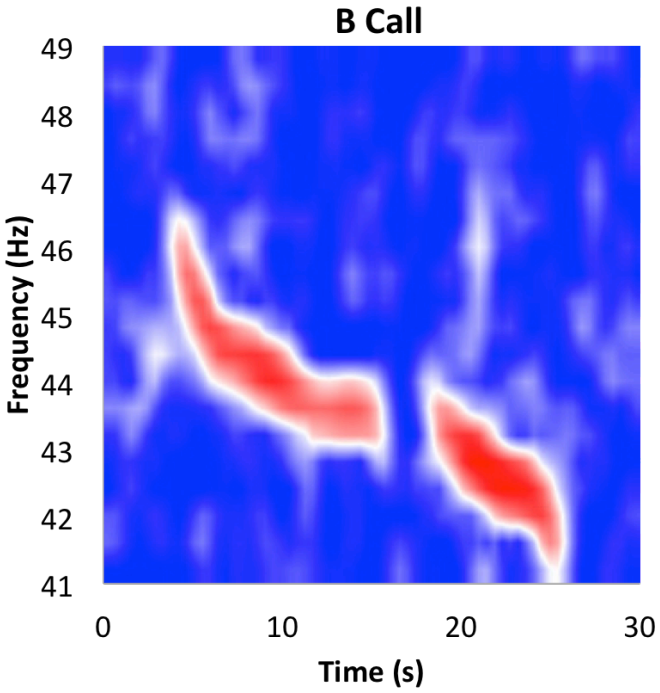



Figure 2.

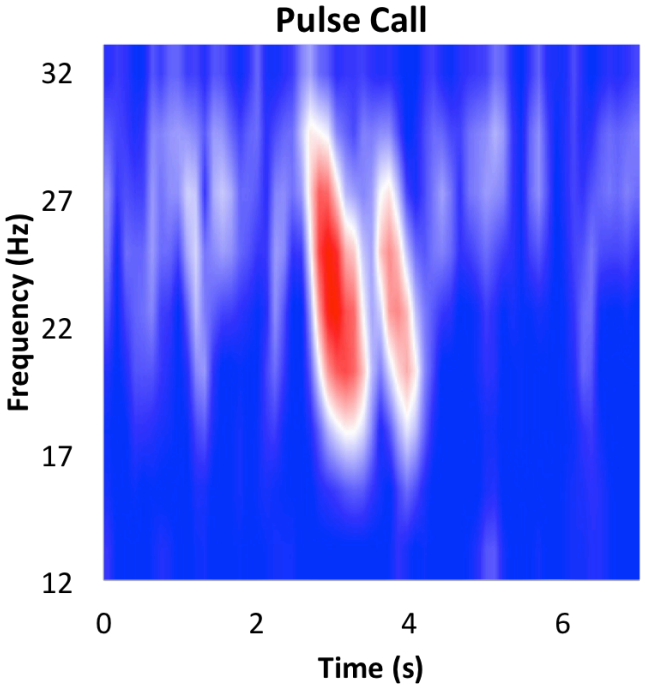


Figure 3.

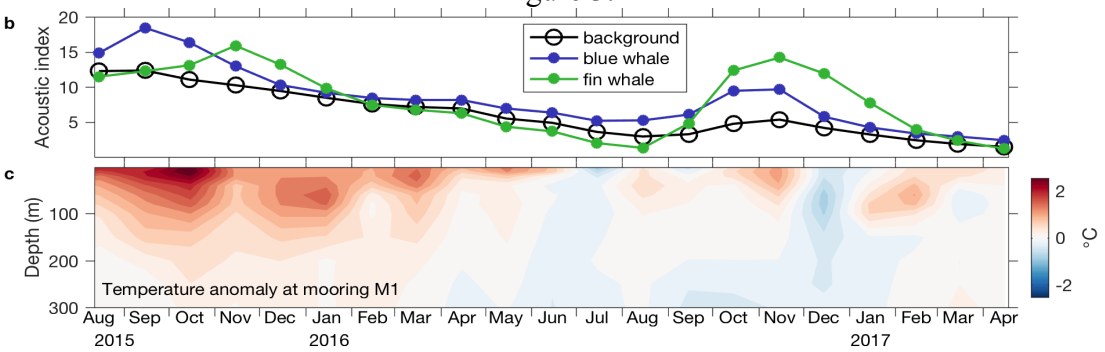



Figure 4.

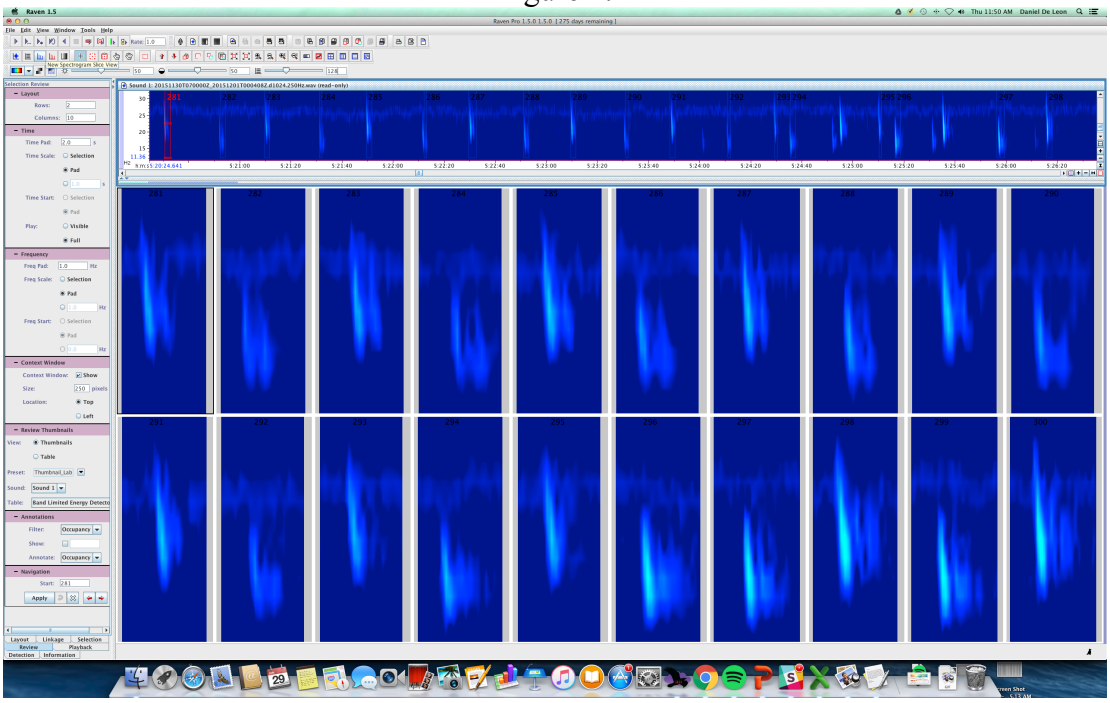


Figure 5.

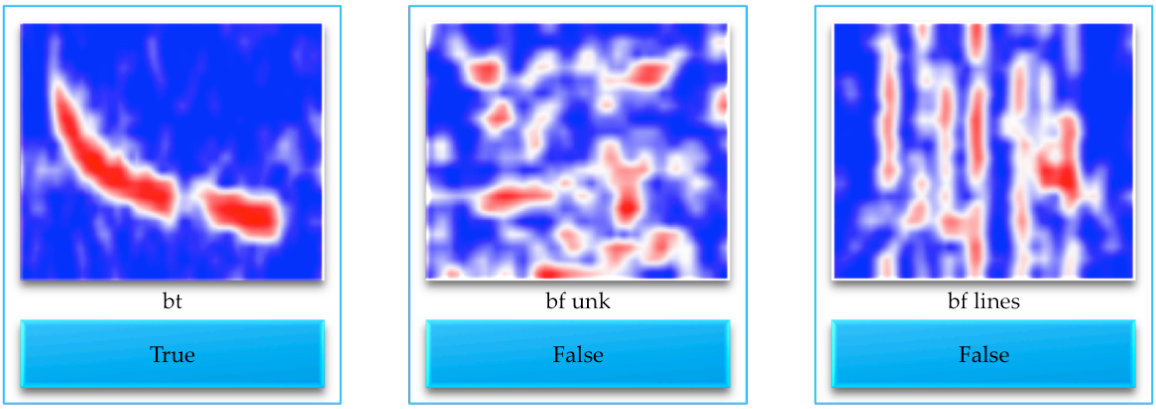



Figure 6.

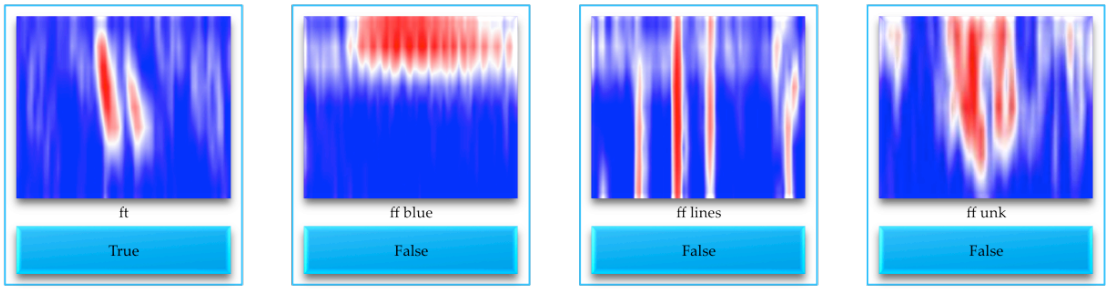
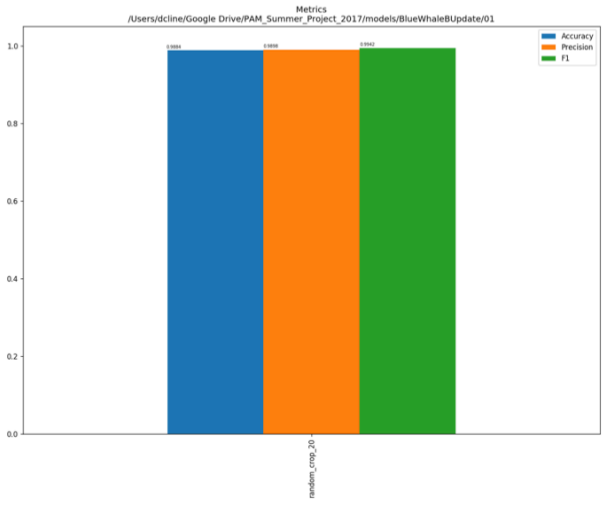



Figure 7.

a.




b.

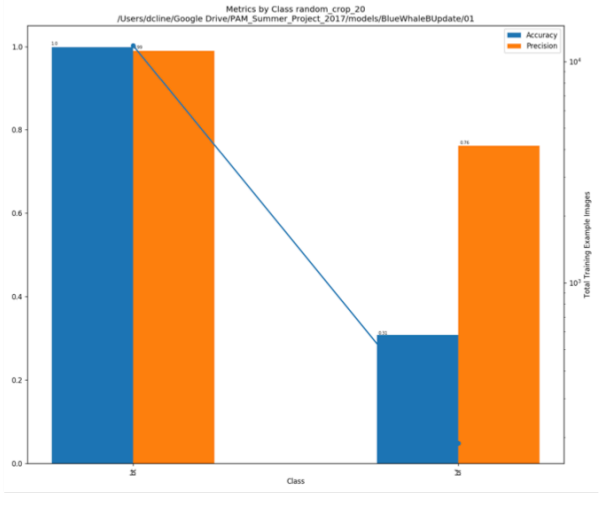
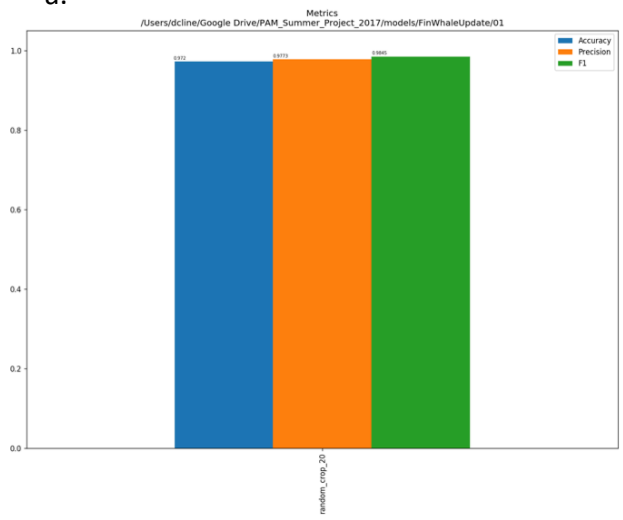


Figure 8.

a.




b.

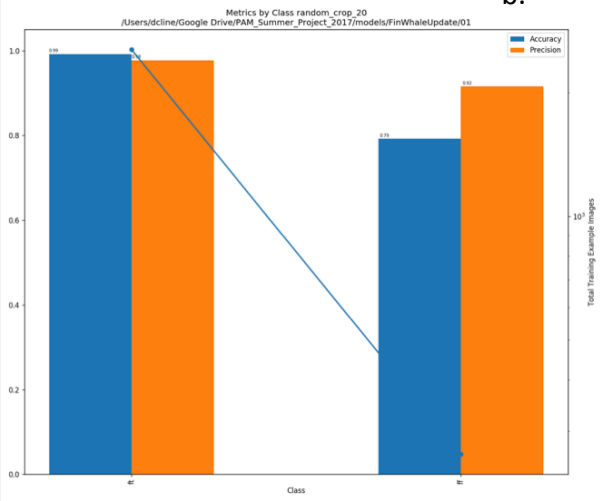
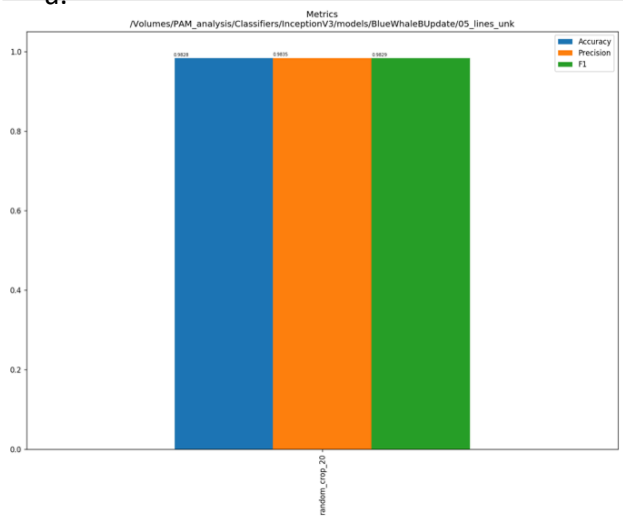



Figure 9.

a.




b.

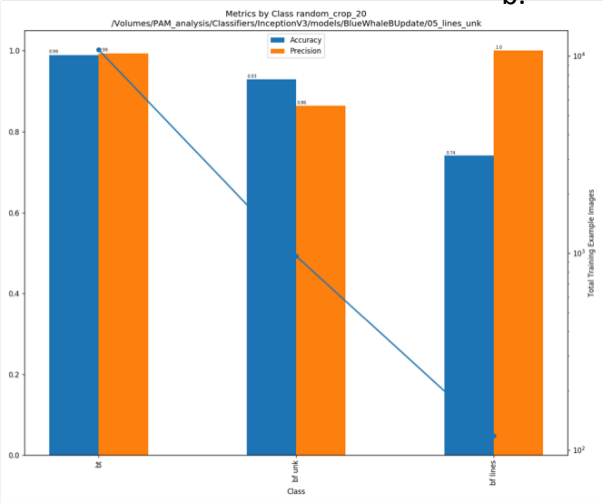
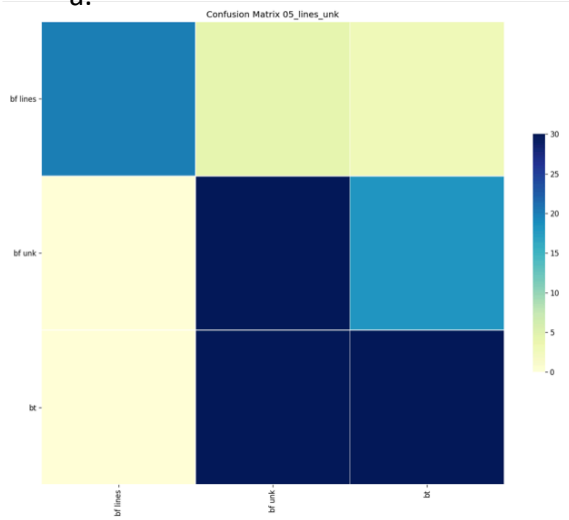


Figure 10.

a.




b.

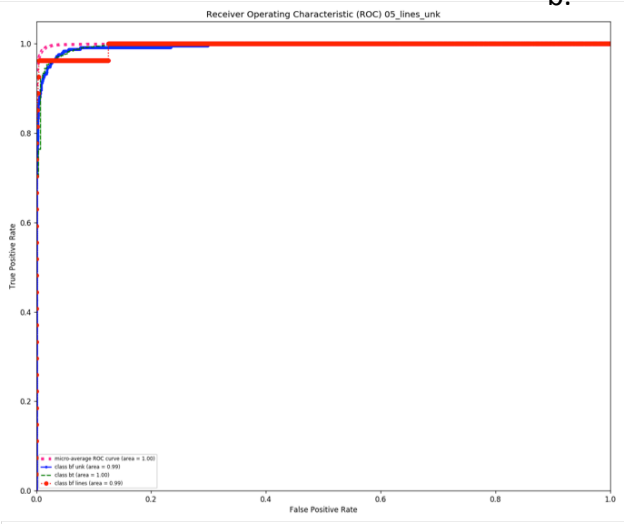
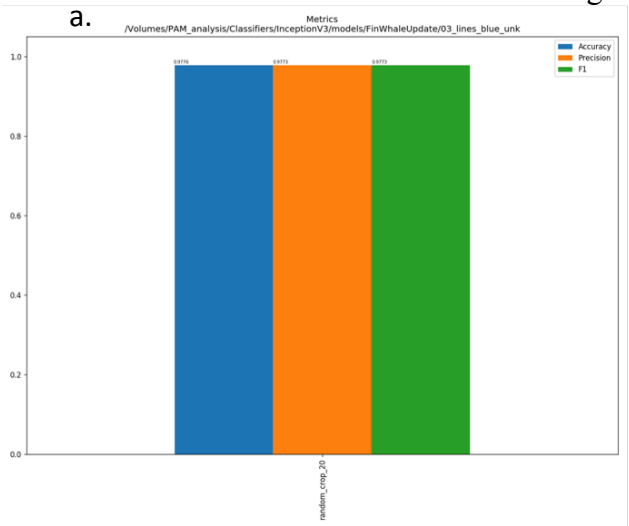
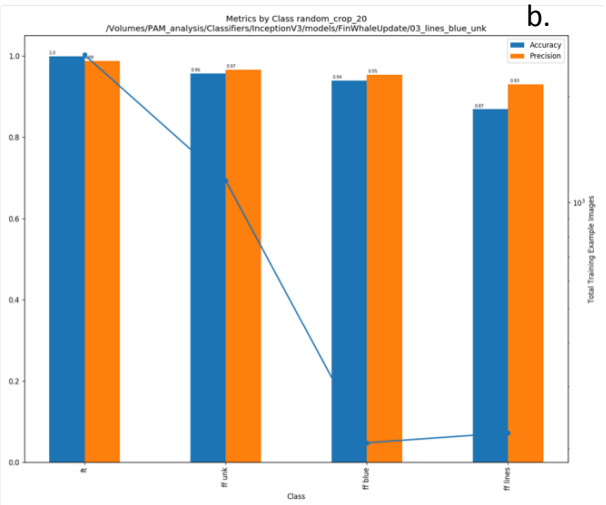



Figure 11.

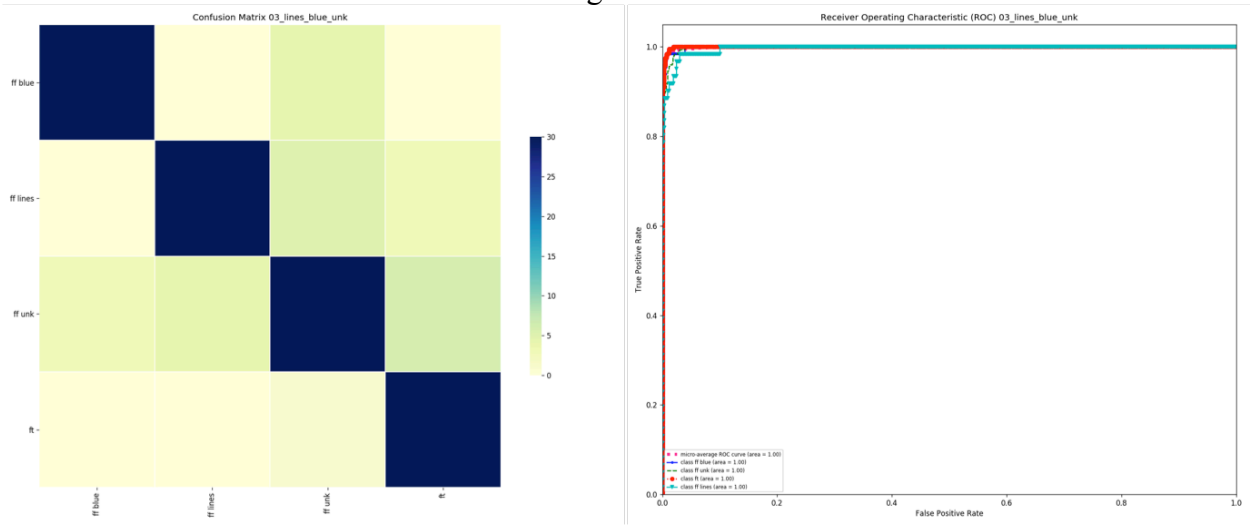
a.




b.



a. Figure 12. b.




## Tables

Table 1.

| Parameters | BWB | FWP |
|---|---|---|
| Min. frequency | 41 Hz | 12 |
| Max. frequency | 48 Hz | 33 |
| Min. duration | 7.168 s | 1.024 s |
| Max. duration | 27.648 s | 4.096 s |
| Min. separation | 5.12 s | 5.12 s |
| Min. occupancy | 60% | 40% |
| SNR Threshold | Above 7.0 | Above 10.0 |
| Block size | 50.176 s | 6.144 s |
| Hop size | 21.504 s | .512 s |
| Percentile | 10.0 | 20.0 |

Table 2.

| Call Type | Xt | Xf_unk | Xf_lines | ff_blue |
|---|---|---|---|---|
| BWB | 15,364 | 1,362 | 160 | NA |
| FWP | 3,721 | 1,607 | 305 | 294 |

## References


Charif, RA, AM Waack, and LM Strickman. 2010. Raven Pro 1.4 User's Manual. Cornell Lab of Ornithology, Ithaca, NY.

Cline, D., De Leon, D. (2015). PAM. GitHub Repository: https://github.com/daniel-deleon/PAM

Debich, A. J., et al. (2015). Passive Acoustic Monitoring for Marine Mammals in the SOCAL Naval Training Area Dec 2012-Jan 2014. Scripps Institution of Oceanography, Marine Physical Laboratory, La Jolla, CA. 7–10. Retrieved from https://www.navymarinespeciesmonitoring.us/files/6614/2767/1542/Debich_et_al._2015_PAM_in_SOCAL_Jan-July_2014_26Feb2015.pdf

Dugan, P. J., et al. (2015). DCL System Using Deep Learning Approaches for Land-Based or Ship-Based Real Time Recognition and Localization of Marine Mammals. Bioacoustics Research Program, Cornell University Ithaca United States. Retrieved from https://arxiv.org/pdf/1605.00982.pdf

Helgren, J., Pastor, J., Singh, A. (2016). whale-sound-classification. GitHub Repository: https://github.com/jaimeps/whale-sound-classification

How to Retrain Inception's Final Layer for New Categories | TensorFlow. (2017). TensorFlow. Retrieved 10 August 2017, from https://www.tensorflow.org/tutorials/image_retraining

Kridler, N., Dobson, S. (2013). moby. GitHub Repository: https://github.com/nmkridler/moby

The Marinexplore and Cornell University Whale Detection Challenge | Kaggle. (2017). Kaggle.com. Retrieved 8 August 2017, from https://www.kaggle.com/c/whale-detection-challenge#description

Peterson, W., Robert, M., & Bond, N. (2015). The warm blob continues to dominate the ecosystem of the northern california current. *PICES Press, 23*(2), 44-46. Retrieved from https://search.proquest.com/docview/1705538895?accountid=10355

Ryan, J., et al. (2016). New Passive Acoustic Monitoring in Monterey Bay National Marine Sanctuary. OCEANS MTS/IEEE Monterey, Monterey, CA, 2016, pp. 1-8. doi: 10.1109/OCEANS.2016.7761363

Szegedy, C., et al. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826. Retrieved from https://arxiv.org/pdf/1512.00567.pdf.