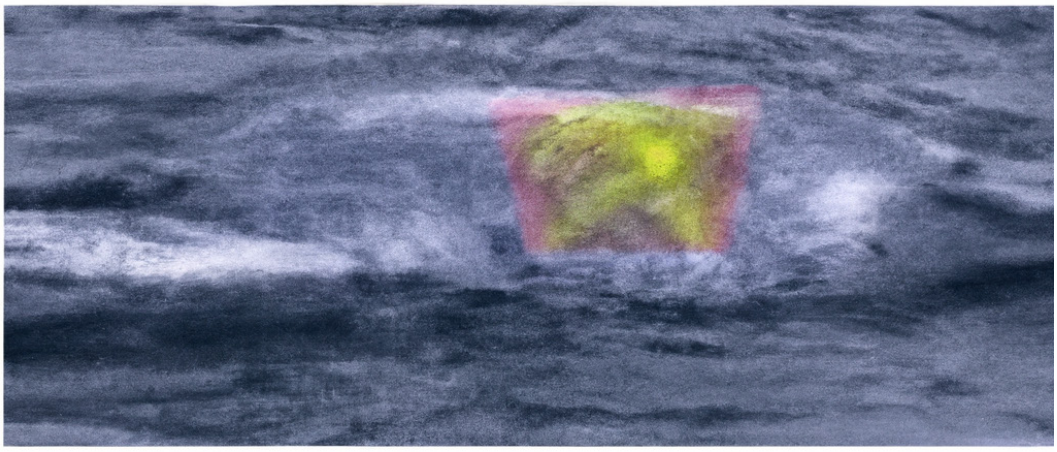
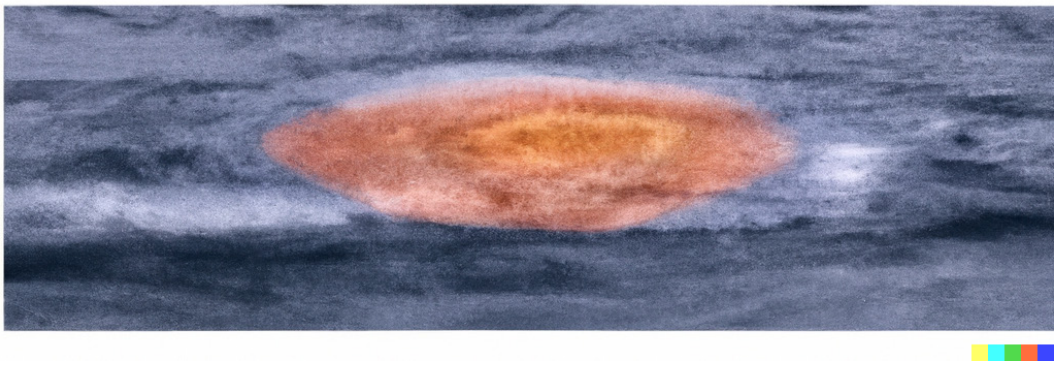
OCS Study
BOEM 2022-064

Phase II of the Gray Whale Study: Development of an Ensemble Algorithm Foundation and a Gray Whale-specific Ensemble Algorithm



$$Score_{combined} = Score_{PIE} * w_{PIE} + Score_{HotSpotter} * (1 - w_{PIE})$$

$$where\ 0 < w_{PIE} < 1$$



U.S. Department of the Interior
Bureau of Ocean Energy Management
Pacific OCS Region, Camarillo, CA

BOEM
Bureau of Ocean Energy Management

OCS Study
BOEM 2022-064

Phase II of the Gray Whale Study: Development of an Ensemble Algorithm Foundation and a Gray Whale-specific Ensemble Algorithm

September 2022

Authors:
Jason A. Holmberg
Andrew Blount

Prepared under Contract 140M0121P0030
By
Wild Me
1726 N Terry Street
Portland, OR 97217

U.S. Department of the Interior
Bureau of Ocean Energy Management
Pacific OCS Region, Camarillo, CA

BOEM
Bureau of Ocean Energy Management

DISCLAIMER

Study concept, oversight, and funding were provided by the U.S. Department of the Interior, Bureau of Ocean Energy Management (BOEM), Pacific OCS Region, Camarillo, CA, under Contract Number 140M0121P0030. This report has been technically reviewed by BOEM, and it has been approved for publication. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of BOEM, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

REPORT AVAILABILITY

To download a PDF file of this report, go to the U.S. Department of the Interior, Bureau of Ocean Energy Management Recently Completed Environmental & Technical Studies – Pacific webpage (https://www.boem.gov/Pacific-Completed-Studies/), and click on the link for 2022-064.

CITATION

Holmberg JA and Blount A. 2022. Phase II of the gray whale study: development of an ensemble algorithm foundation and a gray whale-specific ensemble algorithm. Camarillo (CA): U.S. Department of the Interior, Bureau of Ocean Energy Management, Pacific OCS Region. OCS Study BOEM 2022-064. 14 p.

ABOUT THE COVER

Our cover photos display example visualizations of corresponding texture patterns in lateral photographs of a previously matched gray whale resighted several years later, as imagined by the DALL-E AI-based image generator. Each image represents a different algorithm (PIE and HotSpotter) but leaves it to the AI to render its interpretation. An AI-generated image is appropriate since the work contracted here built an AI to take the output of two computer vision algorithms and generate an ensemble model using both sets of information to predict the ID of a photographed gray whale. Wild Me added the base formula used by the ensemble.

ACKNOWLEDGMENTS

This report acknowledges additional support from the National Oceanic and Atmospheric Administration's (NOAA) National Marine Fisheries Service, Southwest Fisheries Science Center under Contract Number 1305M321PNFFR0326.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
BOEM	Bureau of Ocean Energy Management
CRC	Cascadia Research Collective
ML	Machine Learning
PIE	Pose Invariant Embeddings. A matching algorithm used by Wild Me.
WBIA	Wildbook Image Analysis

1 Executive Summary

Wild Me (wildme.org) completed all tasks for BOEM Contract 140M0121P0030 (the "AI for Gray Whales" project) and is submitting this final report to complete the project.

"Ensembling" is a group of machine learning (ML) techniques that combines several base models to produce one optimal predictive model. Wild Me created a machine learning ensemble model from two distinct computer vision approaches to reliably reidentify gray whales (*Eschrichtius robustus*) from lateral photos. The ensemble evaluates the scored output of each algorithm as trained on gray whales and uses a learned weight to create a single, more accurate prediction based on their outputs. This single result reduces the need for human interpretation of multiple results and can instead suggest the best ID from a gray whale catalog using all available information. Ultimately, the developed ensemble model will assist in improving the accuracy of population studies.

BOEM contract 140M0121P0030 was implemented in tandem with NOAA contract 1305M321PNFFR0326. In the NOAA side of this joint exploration, Wild Me discovered that accounting for time in PIE algorithm matching (a machine learning approach that matches individuals) had no significant impact in accuracy, suggesting overall that time between matches (and any change of the patterning involved) is not a significant limiter in matching individuals with machine learning. However, we also exceeded the contract scope and choose to further explore parameter optimization in PIE training, using an optimizer to further explore the solution space during PIE training and achieving a +5% top-12 accuracy overall.

All developed and tested machine learning models and ID algorithms evaluated under these contracts are now available in Flukebook.org for evaluation and use.

1.1 Completed Task Summary Table

The following tasks were completed under BOEM Contract 140M0121P0030.

Table 1. Completed Tasks Summary

Task	Computer Vision Techniques	Status
2.3.1.3 Task 3: Develop an ensemble algorithm foundation in Flukebook and a gray whale-specific ensemble algorithm. • Develop an ensemble technical foundation in the Wildbook open-source platform to allow the ID suggestion results of multiple algorithms to be merged into a single apparent algorithm. • Create ensemble machine learning ID model for gray whales, for example, PIE+HotSpotter ensemble. • Generate Top-N performance graphs and report of PIE, PIE+HotSpotter, and the ensemble technique for evaluation of accuracy.	HotSpotter [C], PIE [D]	**COMPLETE:** Wild Me created a base foundation for ML-based ensembling in the Wildbook open-source computer vision pipeline [A][B] and evaluated its performance on PIE and HotSpotter, the top-performing and most complementary computer vision algorithms reviewed in Phase I of the study (BOEM Contract 140M0120P0023) [E], through application to gray whales.

1.2 Ensembling Algorithms for Gray Whale Individual ID

In our previous BOEM contract for Phase I , Wild Me evaluated four distinct computer vision approaches to reliably reidentify gray whales (*Eschrichtius robustus*) from lateral photos. Among the evaluated techniques, the HotSpotter [C] and PIE [D] algorithms provided the most overall matching power with an additive performance of top-1 rank of 70% and top-12 of 92%, depending on their chosen configuration and the selection of test data. All developed and tested machine learning models and ID algorithms evaluated under these contracts were deployed in Flukebook.org for evaluation and use.

While the application of two or more computer vision approaches, like HotSpotter and PIE, can help researchers ultimately arrive at the best answer for the question "Which individual animal is in this photo?", multiple approaches also mean multiple ID predictions for human review, with each algorithm offering a potentially different answer to the question resulting in the need for different interpretations of differing scores and lists.

"Ensembling" is a machine learning approach that combines several base models to produce one optimal predictive model. Ensembles can work best if each member approaches the problem differently, allowing each to succeed and fail in different measures. Overall, a good ensemble creates a system that succeeds more than any individual model.

Despite both being considered pattern matchers, HotSpotter and PIE function extremely differently. HotSpotter is not a machine learning algorithm but a geometry-based computer vision algorithm that finds patches (or "hot spots") of local contrast to make matches between photos; the algorithm is not trained on new species but has a fixed logic for generic pattern identification.

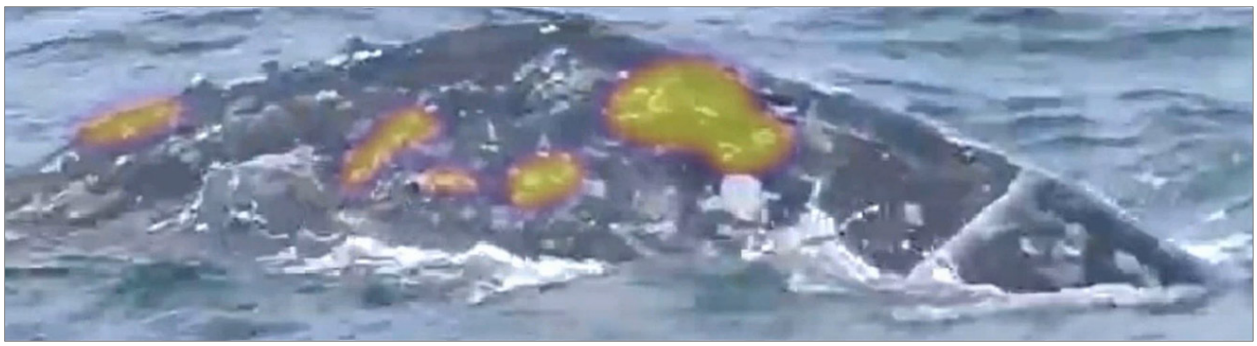


Figure 1. Visualization of HotSpotter "hot spots" used to show textural similarity found in matching.

PIE meanwhile is a classical deep learning approach, using convolutional neural networks trained on species-specific data to extract features that are useful for the particular task of gray whale identification. This architectural difference between PIE and HotSpotter results in very different match behavior, and experience using HotSpotter and PIE for gray whales (see Phase I report [E]) and other species shows that each algorithm is able to find matches that the other misses. This is evident simply by combining the results of both algorithms: whether PIE or HotSpotter was more accurate on its own, looking at the results of both of them would have a higher accuracy still, as shown by the pink line in Figure 2.

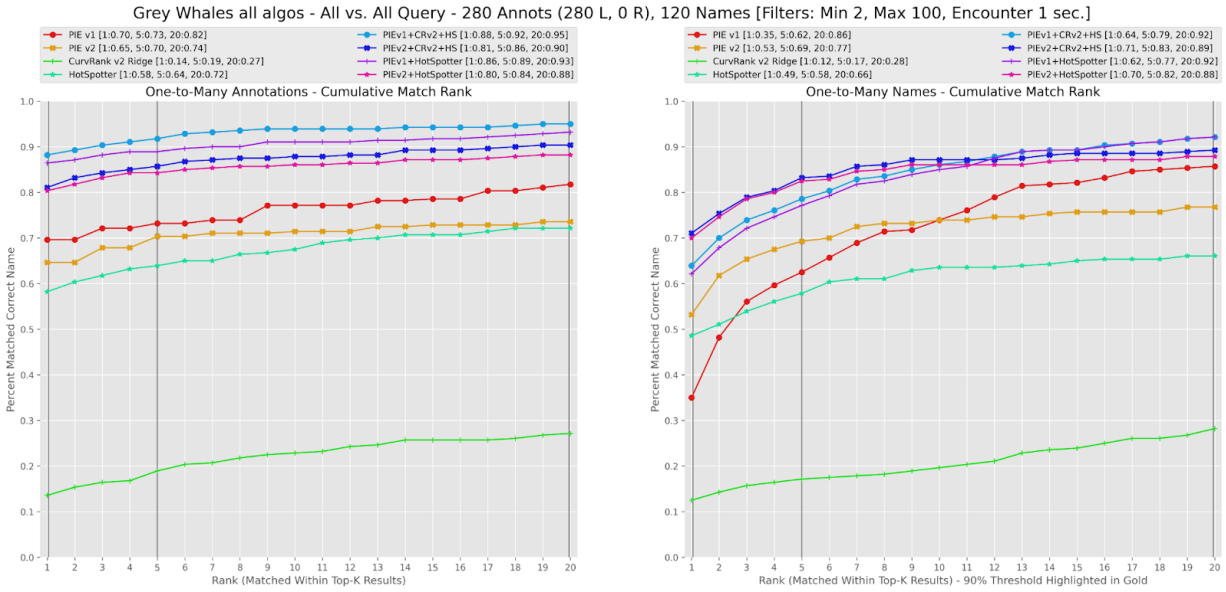


Figure 2. Top-n accuracy for all ID matching algorithms evaluated as well as their combined performance.

In practice, that means users query both algorithms simultaneously on Flukebook, and see each set of results on the same page. The user considers both result/score lists when making the final identification decision. One thing that can confuse users during this process is that PIE tends to have much lower scores (the number indicating algorithm confidence for each candidate match) than HotSpotter, for essentially arbitrary reasons relating to how each algorithm is implemented. A strongly matched HotSpotter score might be 26 while the same PIE match score might be 0.33.

The additional time spent reviewing multiple results, and the fact that algorithms can fail and succeed in ways that allow them to catch each other's mistakes, suggests that blending (also known as "ensembling") the two algorithms into a single unified result would save time for users and be more accurate than either algorithm on its own. There are many approaches to ensembling algorithms, ranging from a simple weighted average of each component score to deep-learning-based approaches that intelligently weigh each algorithm based on the features of each specific query image.

From a deep dive into the behavior of PIE and HotSpotter on gray whales, we determined that a linear combination of the two scores (a weighted average) would be a strong solution to the ensemble problem. We walked through the execution of each algorithm on a number of queries and found that they each tend to find different best-scoring annotations, and each gives all annotations below a certain rank a score of zero. So, for a query matching a gray whale photo against 400 example annotations, each algorithm would return a score vector with about 50 non-zero scores and the remaining 350 scores being zero. This is known as a "sparse vector (or matrix)", one that is mostly zeroes with a few values. Most importantly, the algorithms would tend to have very little overlap between these two sets of 50 non-zero scores. Even though oftentimes both algorithms find the right individual, they look at different annotations to make that decision.

To fit the weighted average ensemble, we designed our experiments to evaluate a training set of left- and right-side gray whale lateral photos. These are the same photos PIE (both v1 and v2 models) was trained on in the Phase I study, provided by Cascadia Research Collective. On a set of 280 right-side photos of 120 individuals with at least two left-side photos each, we queried each photo against the remaining 279

using both algorithms. And on a set of 360 right-side photos of individuals with at least two right-side photos each, we queried each photo against the remaining 359 using both algorithms. We saved the score vector for each query, containing the similarity scores between the query image and every other image. This resulted in a 360x360 similarity score matrix for each algorithm. We built the infrastructure within the Wildbook Image Analysis software [B] to construct these matrices, to sum them together with any defined weights, and then compute the match accuracy of that summed similarity matrix. The fitting task was then to find weights that produce the highest accuracy.

Once the score similarities are computed, weights can be found with linear regression or other automated fitting techniques. We tried a number of these to choose the weights but ultimately searched for them manually based on 1) the fact that really, we are looking for only one single weight number to express the relative weight of each algorithm:

$$Score_{combined} = Score_{PIE} * w_{PIE} + Score_{HotSpotter} * (1 - w_{PIE})$$

$$where\ 0 < w_{PIE} < 1$$

And 2) we were already extremely familiar with the scoring behavior of these algorithms from years of use and our detailed exploration on gray whales.

We ultimately found a weight of 0.9795918367346939 for variable w_{PIE} to have the highest accuracy. Naively, this looks like the ensemble weighs PIE much more heavily than HotSpotter, nearly ignoring the latter. However, PIE weights are generally much lower to begin with. Since each algorithm's score vector is relatively sparse and non-overlapping with the other's, we think of this value as (approximating) the average ratio of a HotSpotter score to a PIE score.

This weighted sum of sparse score vectors results in a nearly-optimal blend, as shown in Figures 3 (left-side images) and 4 (right-side images), where the new ensemble accuracy (blue line PIE-HS) is quite close to the either-or accuracy (PIE +HS) that represents human review of each algorithm's results and subsequent perfect ID judgment. These figures include the independent HotSpotter (green) and PIE (red) plots as well for comparison of accuracy and contribution to matchability. We are pleased with this ensemble result, as it has extremely low computation overhead beyond computing PIE and HotSpotter scores first. The solution is also elegant and can be expressed by a single equation and weight, and it achieves near-optimal performance. As is true of most of our open-source development, we have built the infrastructure to be able to replicate this process with future species of concern to BOEM and other combinations of algorithms.

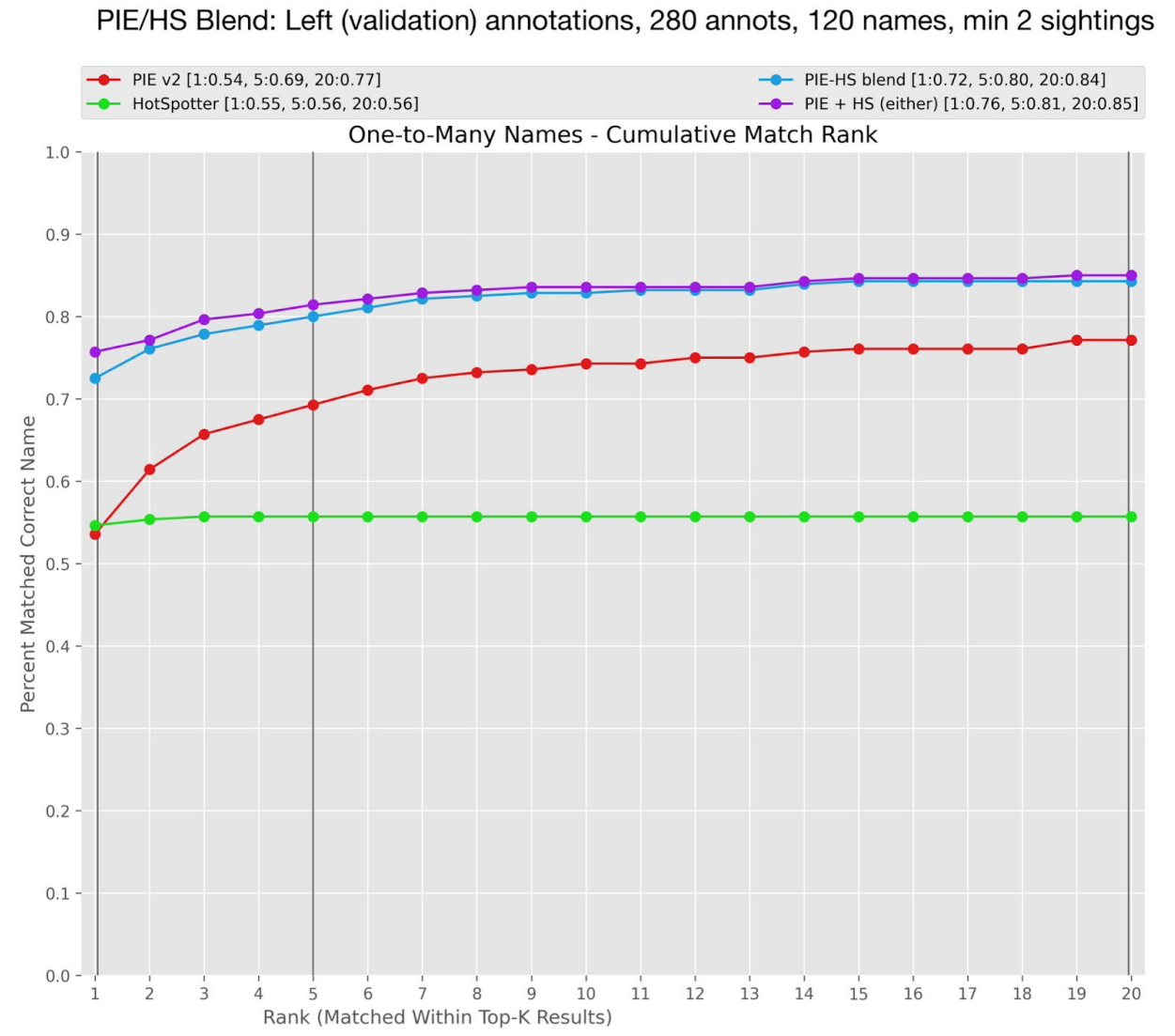


Figure 3. Final results of ensemble accuracy on held-out (not-trained-on) left-side data. "top-k accuracy" is defined as the fraction of queries where the correct result was returned in the top k results, for k values 1-20. Note that every algorithm's accuracy is highly dependent on the exact data being queried.

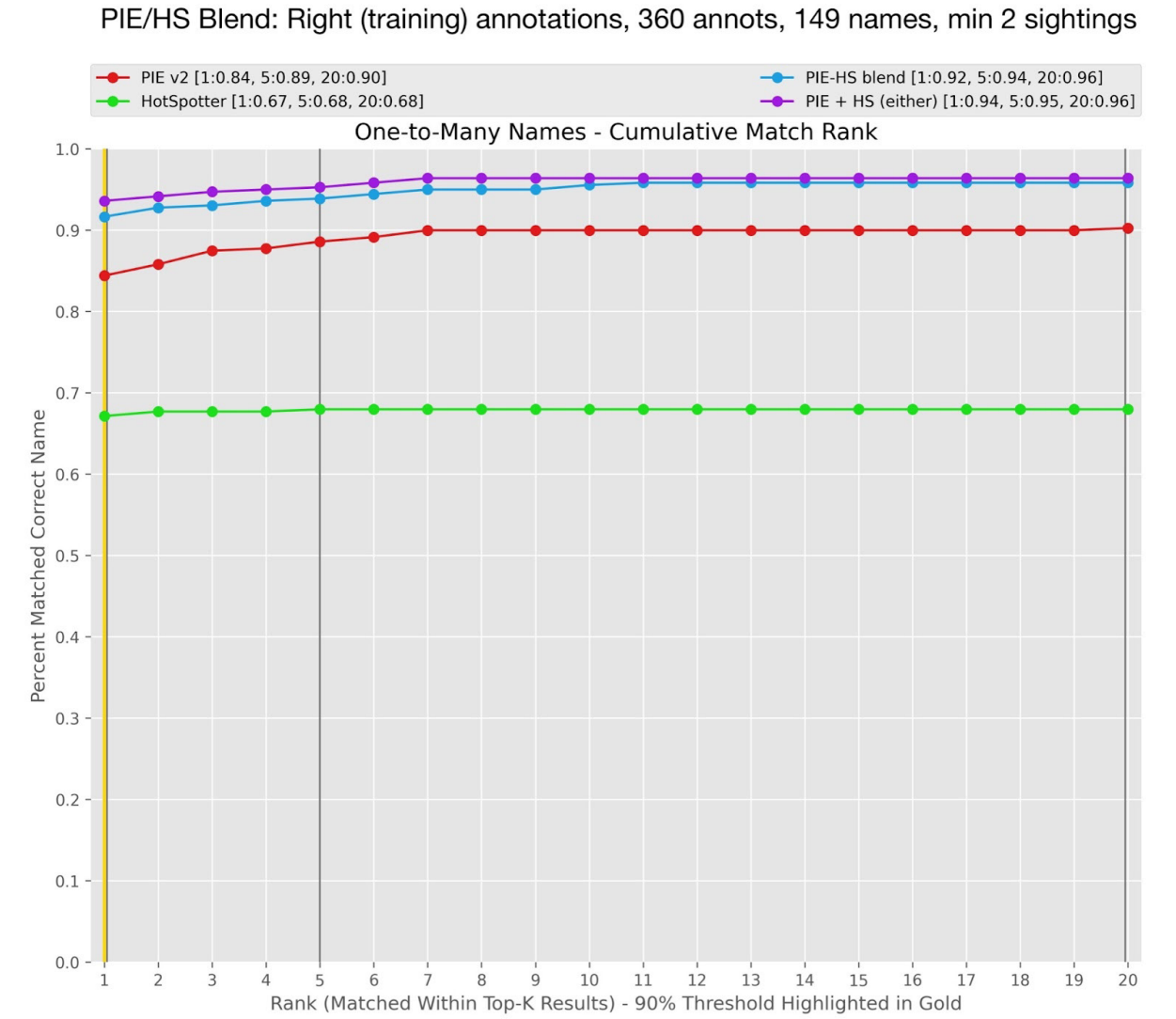


Figure 4. Training accuracy on right-side data. Accuracies are higher than Figure 3 because this is data that PIE was trained on. HotSpotter difference vs. Figure 3 is simply due to different data and small data sizes.

Ensembles do include the limitations of their component parts:

- The ensemble predictor must run after it has both the completed HotSpotter and PIE match results to consider. Given that users of Flukebook.org are likely to run both anyway as the default algorithms, this is trivial overhead.
- Unlike HotSpotter (but like PIE), the ensemble ID algorithm cannot explain "Why" it believes two individuals are matched. Its inputs are numeric, it is trained per species, and its output is a ranked list but not a visualization to guide the eye.

1.2.1 Implementation in Flukebook

The ensemble algorithm is now available for immediate use in Flukebook. It is selectable for gray whales via the “start match” menu option, and it has been set as a default algorithm.

Figure 5. Selection of the Ensemble algorithm is now available in Flukebook.org.

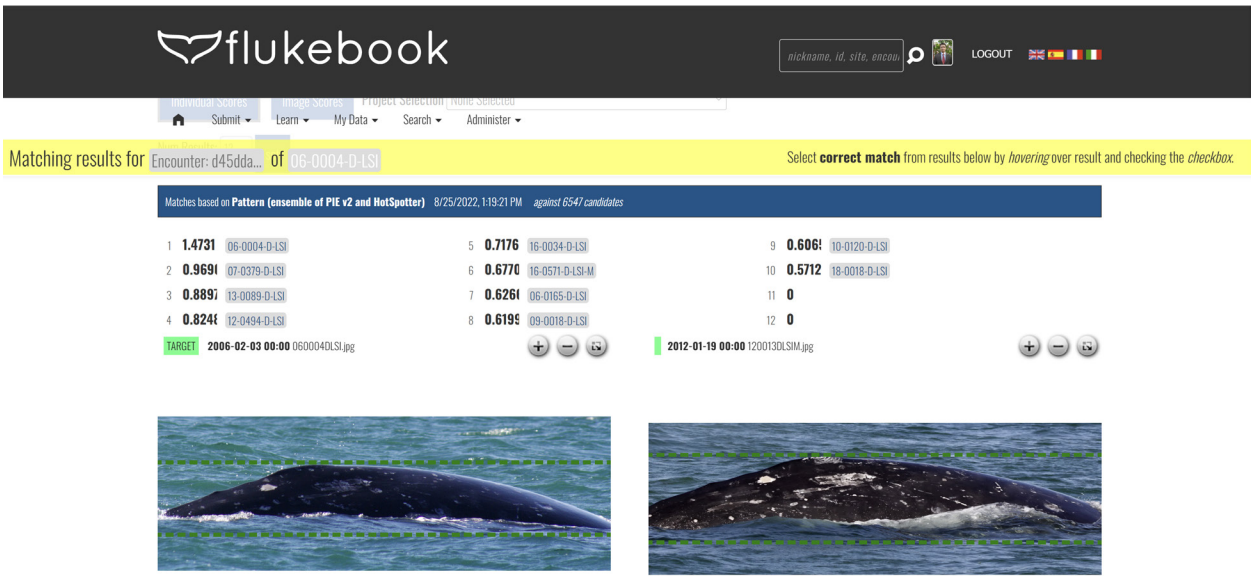


Figure 6. Ensemble match results displayed in Flukebook.org.

1.3 Opportunities for Further Development

1.3.1 Additional Algorithms for Ensembles

Our work suggests new avenues for exploration since our approach can allow for the ensembling of *n*-algorithms per species in Wildbook. These might include:

1. Adding PIE v1 to the ensemble. Ongoing matching work by CRC with PIE v1 shows it has strong matchability that may be more significant than its impact shown in Figure 2.
2. Adding an algorithm resulting from ongoing exploration by ML contractor Jaime Thompson to develop an independent algorithm for gray whale matching in Flukebook, which was recently approved under NOAA Contract 1305M322PNFFR0505. Jaime's work is not likely to be derivative of PIE or HotSpotter, and therefore its independent approach may succeed and fail differently, allowing for a third algorithm to improve the overall accuracy of the ensemble.
3. Including a third-party algorithm developed from an independent Kaggle competition recently completed. This multi-species competition may bring new improvements to gray whale matchability as competitors worked to optimize ID matching for multiple species, and Wild Me has been awarded funding to pursue this evaluation under BOEM IDIQ #140M0121D0004 - Task Order #140M0122F0006. Any resulting, chosen, and implemented algorithm is likely to be distinctly different and offer additional power to a trained ensemble.

2 Works Cited

[A] Parham J, Stewart C, Crall JP, Rubenstein D, Holmberg J, and Berger-Wolf T. 2018. An Animal Detection Pipeline for Identification. 1075-1083. 10.1109/WACV.2018.00123.

[B] Wildbook Image Analysis (WBIA) Pipeline: https://docs.wildme.org/docs/researchers/ia_pipeline

[C] Crall JP, Stewart CV, Berger-Wolf TY, Rubenstein DI, and Sundaresan SR. 2013. HotSpotter-Patterned species instance recognition. In 2013 IEEE Workshop on Applications of Computer Vision, WACV 2013 (p. 230-237). [6475023] (Proceedings of IEEE Workshop on Applications of Computer Vision). https://doi.org/10.1109/WACV.2013.6475023

[D] Moskvyak O. et al. 2019 Robust Re-identification of Manta Rays from Natural Markings by Learning Pose Invariant Embeddings. https://arxiv.org/pdf/1902.10847.pdf

[E] Holmberg JA, Parham JR, Blount A. 2021. Feasibility Analysis: Using Artificial Intelligence to Match Photographed Lateral Ridges of Gray Whales. Camarillo (CA): US Department of the Interior, Bureau of Ocean Energy Management, Pacific OCS Region. OCS Study BOEM 2021-059. 29 p.

U.S. Department of the Interior (DOI)
DOI protects and manages the Nation's natural resources and cultural heritage; provides scientific and other information about those resources; and honors the Nation's trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated island communities.

Bureau of Ocean Energy Management (BOEM)
BOEM's mission is to manage development of U.S. Outer Continental Shelf energy and mineral resources in an environmentally and economically responsible way.