DEVELOPING AND ASSESSING A DIVERSE PLANKTON IMAGERY TRAINING SET FOR MACHINE-LEARNING PLANKTON CLASSIFICATION IN THE NORTH PACIFIC SUBTROPICAL REGION

A THESIS SUBMITTED FOR PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE

IN

GLOBAL ENVIRONMENTAL SCIENCE

MAY 2024

By
Nicole Celine Sulla Mathews

Thesis Advisors

Dr. Angelicque White
Dr. Fernanda Henderikx-Freitas

We certify that we have read this thesis and that, in our opinion, it is satisfactory in scope and quality as a thesis for the degree of Bachelor of Science in Global Environmental Science.

THESIS ADVISORS

Dr. Angelicque White
Department of Oceanography

Dr. Fernanda Henderikx-Freitas
Department of Oceanography

ACKNOWLEDGEMENTS

Thank you to my advisors, Angel, for introducing me to the world of microbial oceanography, and Fernanda, for always answering my countless questions and being my guide throughout this project. Thank you to the annotation team for their contributions to the training set development as well as Dr. Andrew Hirzel for overseeing the CNN development, implementations, and results used for this paper. Thank you to the White/Henderikx-Freitas Lab for providing the time, effort, and resources to support my endeavors. I would also like to thank my professors, advisors, and the Department of Oceanography for sparking my interest in this field and continually providing the opportunities to explore it. Thank you to my reviewers, who took the time and effort to help my thesis to be the best it could be. Last but certainly not least, si yu'os ma'åse' to my family – for always being there for support, inspiration, and guidance – you are the reason I am where I am today.

ABSTRACT

The Imaging FlowCytobot (IFCB) has a continually growing role in oceanographic research, particularly in the exploration of microbial life within the North Pacific Subtropical Gyre (NPSG). However, the vast amount of data generated by the IFCB poses a challenge for manual sorting and taxonomic classification. This study addresses this challenge by developing a Convolutional Neural Network (CNN) training set to efficiently categorize IFCB images into taxonomic groups. Specifically focusing on the diatom *Hemiaulus* and ciliate phylum Ciliphora during a research cruise within the NPSG in the summer of 2021, the study aims to quantify the CNN's performance compared to manual annotations of IFCB images taken on this cruise, providing insights into the CNN’s accuracy and precision over time. Statistical analyses of the CNN’s machine learning-based classifications indicate a high accuracy in the automated identification of *Hemiaulus* and Ciliophora. Analysis of biovolume and particle number concentration reveals trends in taxonomic abundance over the course of the cruise. Despite morphological changes of *Hemiaulus* as it loses structure over time, the CNN demonstrates an overall improvement in accuracy as the cruise progresses, particularly for *Hemiaulus*. This study highlights the development of a robust training set of roughly 76,000 images, allowing the CNN to accurately classify images collected within the NPSG.

Keywords: Taxonomic sorting, machine learning, zooplankton, Imaging FlowCytoBot (IFCB), North Pacific Subtropical Gyre (NPSG), ocean microbiology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 Significance of research

The exploration of plankton in the ocean serves as a fundamental aspect of understanding marine ecosystems, offering insights into biodiversity, ecological dynamics, and biogeochemical processes. Plankton, despite their microscopic size, play pivotal roles in marine food webs, carbon cycling, and global climate regulation (Falkowski, 2012). Therefore, accurate quantification and characterization of planktonic organisms are essential for comprehending and predicting oceanic processes.

The introduction of automated imaging tools such as the Imaging FlowCytobot (IFCB) marks a significant advancement in ocean research that has allowed for a more thorough investigation of the diversity and ecology of microbial life in the global ocean. We have utilized this tool in the North Pacific Subtropical gyre (NPSG), where low concentrations of nutrients limit planktonic growth and the standing stocks of phytoplankton are relatively low. In this ecosystem, the IFCB can collect around 45,000 images per day of particles ranging between ~4-100 µm from surface waters when run in near-continuous mode. However, the sheer volume of data generated by the IFCB presents a daunting challenge—it can be overwhelming, difficult to process, and impractical to manually sort through. This project aims to address this issue by developing an accurate, manually annotated training set to guide a machine learning approach (specifically a Convolutional Neural Network, CNN) to classification of IFCB images into distinct taxonomic categories. Therefore, the overarching goal is to quantify the performance of the CNN compared to manual annotations through the analysis of the abundance and sizes of two taxa, the diatom *Hemiaulus* and ciliate phylum Ciliophora, collected over the course of a 14-day research cruise in the NPSG in July of 2021.

Any machine learning approach requires a high-quality, curated training set of images. The ultimate accuracy of machine learning categorization is directly dependent of the strength and breadth of the training sets in representing the diversity of morphology for distinct classes. Additionally, given *Hemiaulus'* change in morphology over time, as further discussed in this paper, it is hypothesized that these changes may impact the CNN's ability to accurately classify *Hemiaulus.* This paper aims to (1) describe the development of a robust training set of images for the NPSG and (2) assess the accuracy of CNN-based categorization of two distinct taxa of plankton imaged over the course of a process cruise conducted in the summer of 2021. In addressing these aims, this paper will also broadly present the resulting data on abundance and sizes of these organisms, thereby encompassing a comprehensive evaluation of both the classification performance and ecological characteristics of the NPSG plankton community.

1.2 The North Pacific Subtropical Gyre

Being the largest ecosystem on the planet, the North Pacific Subtropical Gyre (NPSG) is a prime example of what is often referred to as an "ocean desert" due to its unique characteristics that contribute to a relatively low level of biological productivity (Dai *et al*., 2023). This phenomenon stems from the establishment of a permanent lowdensity cap within the NPSG, which hampers vertical nutrient mixing and limits essential nutrient availability to surface waters (Karl & Church, 2017). Climate change has the potential to exacerbate these

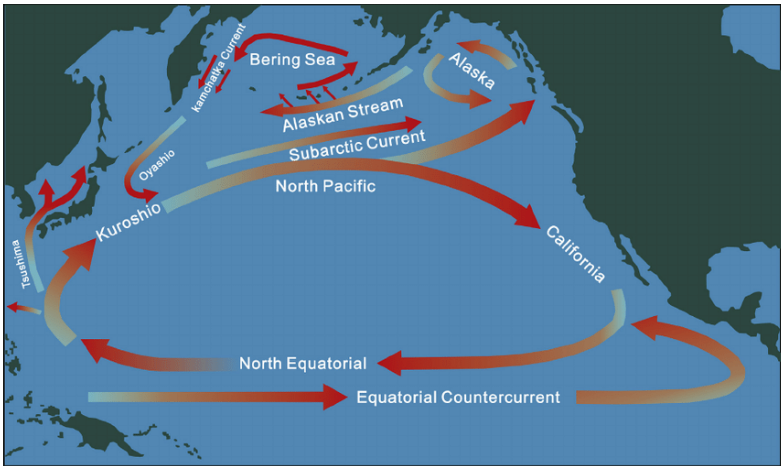


Figure 1.2.1 The North Pacific Subtropical Gyre and its surrounding currents. *(ResearchGate, 2023)*

conditions, with rising sea surface temperatures which can intensify stratification and concomitantly reduce nutrient upwelling from deeper waters (Karl *et al.*, 2021; Gregg *et al.*, 2005; Boyce *et al*., 2010).

The data collected and analyzed here are a component of a larger research expedition to investigate the ecology and biogeochemistry of a region of the NPSG marked by the presence of an anticyclonic mesoscale eddy and enhanced chlorophyll biomass observable from satellite-based remote sensing. Initial in situ observations of this feature revealed high silicate levels in surface waters and elevated concentrations of N_2 associated-diatoms known to be associated with symbiotic N_2 fixing microorganisms. Blooms of diatom-diazotroph associations (termed DDA's) are not uncommon for the region in summer-fall months when waters are strongly stratified (Villareal *et al*., 2012 and Grabowski *et al*., 2019). The growth of diazotrophs is a source of 'new' nitrogen to the upper euphotic zone, which is persistently nitrogen starved. This newly fixed nitrogen can fuel the growth of non-diazotrophic species and contribute to the enhancement of nutrient export processes (Karl *et al*., 2021). Therefore, these blooms play a pivotal role in nitrogen cycling and carbon regulation, significantly impacting the biogeochemical dynamics of the NPSG ecosystem.

1.3 Imaging FlowCytoBot

In 2018, the White/Henderikx-Freitas lab began making underway measurements on regular research cruises in the NPSG using an Imaging FlowCytoBot (IFCB) to better understand the taxonomic makeup of the NPSG and surrounding oceanic region. Further explained in the Methods section of this paper, the IFCB is an in-flow submersible cytometer that is used in near-continuous mode to sample plankton populations from a shipboard uncontaminated flow-through system. With the help of a red diode laser, the IFCB is able to detect and capture all particulates ranging in size between ~4-100 µm that flow through. The IFCB is triggered to capture images through both light scattering and red fluorescence, emitted by chlorophyll and phycocyanin pigments, to allow for the documentation of both phytoplankton and zooplankton (Dugenne *et al.*, 2020 and Dugenne *et al.*, 2023) While all particles scatter the laser's light, those containing chlorophyll will emit a red fluorescence, allowing the machine to also document the fluorescence of each passing particle (Olson and Sosik, 2007). The IFCB has the ability to capture images of particles from ~4 to 100 μm in size with the upper limit determined by the mesh size of screening used at the intake port. The sample volume is ~ 5 ml and the sample frequency is ~ 20 min. (Olson and Sosik, 2007). Through a combination of video and flow cytometric technology, this machine captures images of various microbes that 'trigger' the camera based on scattering or fluorescence thresholds. This allows morphological descriptions of imaged particles along with their relative chlorophyll fluorescence and scattering signals. This data can then be used for taxonomic identification or cell size dynamics (Cael & White, 2020, Dugenne *et al.*, 2020).

Over the past six years of using the IFCB, the lab has accumulated over 9 million images across multiple research cruises across the Pacific Ocean. With research cruises such as a 2021 process cruise (termed PARAGON 1) accumulating roughly 650,000 images, manual annotations of IFCB images become time-consuming and prohibitively expensive from a labor perspective. As a result, machine learning classification pipelines are typically used to facilitate automated classification of imaged. There are currently two major automated learning approaches for IFCB images: support vector machine based on a feature selection algorithm and random forest (RF) algorithms (Juranek *et al*., 2020 and Nardelli *et al*., 2022). Following recent advances in deep learning techniques for computer vision, the IFCB community has been transitioning to CNN's to enhance image classification accuracy. CNNs directly extract features from images and learn semantically meaningful features as they train on labeled images. These features correspond to relevant components of the image related to the labels, making CNNs highly accurate for image classification tasks (Nardelli *et al*. 2022). Additionally, CNNs offer advantages over traditional methods beyond improved accuracy. They can automatically learn and adapt to complex patterns within the images, eliminating the need for manual feature extraction and selection (González *et al*., 2019). This inherent adaptability enhances the efficiency and scalability of the classification process, making CNNs a promising tool for large-scale analysis of microbial communities in the marine environment.

This study aims to address the pressing challenges in the field of oceanographic research, particularly concerning the exploration of microbial life in the NPSG. While the IFCB's images have the potential to provide great insights into the dynamics of microbial life in the NPSG, the large quantities of images must be accurately classified. The overarching goal of this research is to develop a robust training set for machine-learning-based plankton classification, utilizing data collected from the IFCB during a research cruise in the NPSG. Specifically, this study seeks to quantify the accuracy of a CNN in categorizing two distinct taxa of plankton, *Hemiaulus* and Ciliophora, imaged during the cruise. By doing so, this study aims to contribute to a better understanding of the ecological dynamics and taxonomic composition of plankton communities in the NPSG. To achieve these goals, a diverse training set of images was meticulously curated to be representative of the NPSG's plankton diversity. Subsequently, a CNN-based classification approach was then implemented and assessed for its performance in accurately categorizing plankton images. This study aims to advance the field by providing insights into the efficacy of machine learning approaches in plankton classification and their potential applications in oceanographic research.

2.0 METHODS

2.1 Overview

The primary objective of this study was twofold:

1. *Development of a Robust Training Set:* The first goal was to create a comprehensive training set of images specifically tailored for the NPSG . This involved curating a diverse collection of high-quality images obtained from multiple research cruises, including the 2021 PARAGON 1 expedition. The training set aimed to encompass the wide range of plankton taxa and environmental conditions characteristic of the NPSG, serving as a foundation for training the CNN.
2. *Assessment of CNN-Based Categorization Accuracy:* The second goal was to evaluate the accuracy of CNN-based categorization, particularly focusing on two distinct taxa of plankton imaged during the 2021 PARAGON 1 cruise: *Hemiaulus* and Ciliophora. This assessment involved comparing the CNN's classifications with manual annotations conducted using the Ecotaxa web application. By examining the performance of the CNN in accurately categorizing these taxa, insights were gained into the efficacy and reliability of the CNN for taxonomic classification within the NPSG ecosystem.

2.2 Data Collection

Data collection for this study involved the utilization of the IFCB. The IFCB operates by continuously sampling water from a depth of 7 meters (via the ships uncontaminated seawater intake) as the vessel travels underway. The dataset used for training set development comprised images collected from multiple research cruises, including the 2021 PARAGON 1 cruise, as well as other expeditions conducted within the NPSG region. These images provided a diverse representation of plankton taxa and environmental conditions in the NPSG. Through the 2021 PARAGON 1 cruise, the IFCB collected an extensive dataset comprising approximately 650,000 images.

Using the additional data collected by the IFCB, the total volume concentration of each particle captured could be calculated. The total volume concentration per class (μL L^{-1}) was calculated by accounting for the volume sampled and individual particle biovolume estimates using the algorithm developed by Moberg and Sosik (2012), which are both standard outputs of the IFCB raw data. The biovolume algorithm operates on a two-dimensional image processed to identify organism boundaries, computing the distance of each interior pixel to the nearest boundary. These distances are then assumed to apply for projection in the third dimension, with resulting volumes adjusted by a multiplicative factor, assuming locally circular cross-sections (Moberg and Sosik, 2012).

2.3 Approach used and model development

In order to properly assess IFCB images through manual annotations, it was important to first become familiar with plankton taxonomy via literature review and study the types of images that the instrument detects. During this period, an annotation guide was developed aimed at helping define the general number of classes. As the training set was morphological-based, unique characteristics for each major grouping were outlined to aid in classification. Where taxa were difficult to differentiate, groupings were aggregated to increase accuracy. These morphological groupings and their determining characteristics can be seen through the flowchart shown in the Appendix. Following the preliminary period, manual annotations were conducted through the taxonomy categorizing website, Ecotaxa (Picheral *et al*., 2024).

Ecotaxa is a web application tailored for taxonomic sorting which was used to facilitate the organization and validation of IFCB images. For the purposes of building the training set, images were manually sorted into folders based on their taxonomic groups. During manual validation, each image was carefully reviewed and categorized either as belonging to the predicted taxonomic group or as "dubious" if there were uncertainties, ensuring meticulous validation and accuracy in taxonomic classifications.

Through Ecotaxa, 654,312 images were manually annotated and classified into 12 living groups and 6 non-living/detrital groups. As seen in the Appendix, these 18 groups were largely morphologically based, defined by visual characteristics such as centric, pennate, or crescent appearance. Following manual annotations of PARAGON 1, samples of each group were placed into a training set for the CNN (see section 2.4 *Convolutional Neural Network* for detailed information). This training set encompassed a diverse array of IFCB images collected not only during the PARAGON 1 expedition but also from several other research cruises (PARAGON 2, G4, G5, PFIX) conducted within the region. Each taxonomic group within the training set was represented by varying numbers of images, ranging from approximately 200 to 10,000 images per group. While this training set aimed for equal representation from all contributing cruise data, this goal was limited by the availability of high-quality images from our instrument and sampling region. As certain taxa are more abundant than others within this region, more images are available for selection into the training set. To address this imbalance, classes with a larger abundance were under sampled in the training set to ensure a more balanced representation of taxa.

For more easily identifiable classes such as *Arthopoda*, Ciliophora, and *Radiozoa*, less images were needed for the training set as there was more confidence in these images when placed in the training set. The more heterogenous, or aggregated the class, such as *Dinophyceae* and *Haptophyta*, the more images were needed to accurately represent the range in morphologies across the large grouping. An in-depth breakdown of the number of images in each classifying group can be found in the Appendix.

For the training set discussed in this paper, PARAGON 1 contributed 16,755 out of 75,945 images, representing 22% of the training set. All image sources and their percentage represented within the training set can be seen in Table 2.3.1.

Image source	ZIP	PARAGON 1	PARAGON 2	G4	G5	PFIX
Image amount	18,267	16,775	7,516	14,265	11,170	7,952
Percent represented in training set	24%	22%	10%	19%	15%	10%
Total training set images	75,945					

Table 2.3.1. Breakdown of image source representation for the training set utilized by the CNN. Image source names refer to the names of specific cruises or expeditions. Images from these cruises can be browsed here: http://ifcbdb.soest.hawaii.edu/dashboard

2.4 Convolutional Neural Network

The CNN employs a hold-out cross validation approach, where 20% of the training set is withheld for validation, while the remaining 80% is utilized for training. During training, the CNN iteratively refines its classification capabilities by comparing features extracted from the training set with those from the withheld validation set. This process is repeated multiple times (epochs), with the classifier adapting its parameters based on the performance observed on the validation set. It's important to note that the same split of training and validation sets is maintained for each epoch, ensuring consistency in evaluation. Additionally, the order in which images are presented to the CNN varies based on random seeds, resulting in each classifier learning slightly differently.

To ensure robustness and reliability, the CNN model was run four times, each with a different random seed. The results from these four instances are summarized in Table 3.1.1. It is crucial to note that these instances provide insights into the variability in CNN performance due to random initialization.

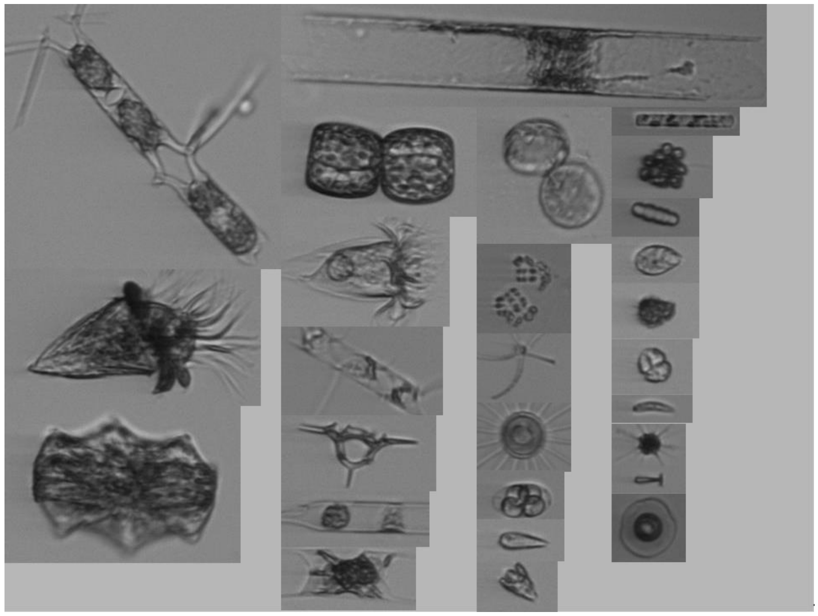


Figure 2.4.1 Mosaic of a portion of the images included in the training set utilized for the CNN

2.4.1 Confusion matrices: role and purpose

To evaluate the performance of the CNN classification model, a confusion matrix was used. By examining the distribution of predictions across the matrix, one can identify patterns of misclassification and areas for improvement. At its base, a confusion matrix organizes the model's predictions into four categories:

1. True Positives (TP): Instances where the model correctly predicts a positive morphological grouping.
2. False Positives (FP): Instances where the model incorrectly predicts a positive morphological grouping.
3. True Negatives (TN): Instances where the model correctly predicts a negative morphological grouping.
4. False Negatives (FN): Instances where the model incorrectly predicts a negative morphological grouping.

These four categories then form the foundation for calculating various performance metrics such as:

- Precision: The proportion of true positive predictions among all positive predictions made by the model, calculated as *TP / (TP + FP).*
- Recall (also known as sensitivity or true positive rate): The proportion of true positive predictions among all actual positive instances, calculated as *TP / (TP + FN).*
- F-1 Score: A measurement of model accuracy, calculated as the harmonic mean of precision and recall, calculated as *2 * (precision * recall) / (precision + recall).*

It's worth noting that similar performance metrics were also applied to the White/Henderikx-Freitas Lab's previous classification tool, Random Forest, which was utilized for annotating IFCB images collected by the lab. This approach aimed to ensure consistency and enable meaningful comparisons between different classification methods. By evaluating these metrics across various classification approaches, a comprehensive understanding of the accuracy and precision of the CNN in comparison to alternative methods can be obtained.

2.5 <u>Manual vs. CNN: statistical analyses on observational data of abundances and biomass</u>

To compare the performance of the CNN and manual annotations, statiscual analyses of observational data related to abundance and biomass are conducted. The analyses conducted within this paper assume that all manual annotations accurately represent the true value, providing a reference point for evaluating the CNN annotations.

To assess the CNN's accuracy in classifying particles, various statistical methods are employed to measure the level of agreement or discrepancy between CNN annotations and manual annotations. One key metric used in this comparison is the absolute difference, which quantifies the magnitude of variation between variables derived from the CNN annotations (experimental values) and those from the manual annotations (theoretical values). Specifically, variables such as volume concentrations and total particle counts of specific taxa of interest obtained from the IFCB are analyzed. The absolute difference enables us to understand the extent of disparity between the two annotation sets, offering insights into the CNN's classification accuracy.

Moreover, additional statistical tests are utilized to further examine the agreement between CNN and manual annotations. The t-test is employed to compare paired observations derived from the CNN and manual datasets, with a particular focus on temporal variations. Through these analyses, we aim to comprehensively evaluate the performance of the CNN in comparison to manual annotations, shedding light on its accuracy and precision in classifying planktonic particles.

3.0 RESULTS

3.1 Overview

As previously outlined, this work aimed to (1) describe the development of a robust training set of images for the NPSG and (2) assess the accuracy of CNN-based categorization of two distinct taxa of plankton imaged over the course of the 2021 PARAGON 1 cruise. In doing so, this paper also aims to (3) conduct a statistical comparison on the abundance and sizes of these organisms.

This analysis focuses on comparing the accuracy of the CNN against manual classifications of two taxa: the chain-forming diatom genus *Hemiaulus* which is often associated with symbiotic diazotrophs of the genus *Richelia,* and a general classifier for the phylum Ciliophora, providing insights into the efficacy and reliability of the CNN in discerning taxonomic distinctions within the NPSG region.

Based on the F1 scores for the *Hemiaulus* and Ciliophora as depicted in Table 3.1.1, it is evident that the CNN was highly accurate in classifying these two specific taxa across a combined dataset. The F1 score, interpreted as a measurement of model accuracy, represents the harmonic mean of precision and recall, providing a comprehensive assessment of clarification performance. Specifically, F1 scores for *Hemiaulus* ranged from 98% to 99%, while F1 scores for Ciliophora ranged from 87% to 93%. Notably, the CNN exhibits a difference in prediction accuracy between the two taxa, with *Hemiaulus* demonstrating a higher range in F1 scores across all trials, or instances.

Hemiaulus		**Instance**	**Amount of PARAGON 1 images used in withheld CNN group**	**F1**
	Withheld CNN	1	5075	0.9789
		2		0.9901
		3		0.9827
		4		0.9882
Ciliophora	**Withheld CNN**	1	1881	0.9294
		2		0.8682
		3		0.9251
		4		0.9084
All classes	**Withheld CNN**	1	75945	0.9146
		2		0.8869
		3		0.9331
		4		0.9283

Table 3.1.1. *Hemiaulus* and Ciliophora comparison of F1 scores across the 4 epochs/trials of the CNN

Figures 3.1.1 and 3.1.2a focus on the particle amount and total concentration of Ciliophora. Table 3.1.1 highlights a significantly lower number of images within the CNN training set for the Ciliophora morphological group, totaling 1,881 images. This lower volume of images may explain the significantly noisier data points observed in Figure 3.1.3 for Ciliophora in contrast to the smoother plot for *Hemiaulus* (3.1.4).

Seen in Figure 3.1.1 and 3.1.2a, both the total biovolume concentration and particle amount for Ciliophora appear to remain consistently low. For a closer examination of Figure 3.1.2a, a direct comparison between CNN and manually validated particle amounts on a log10 scale was conducted and is illustrated in Figure 3.1.2b, showing that the relationship is significant ($p < 0.0001$) but the variance explained (27%) is relatively low.

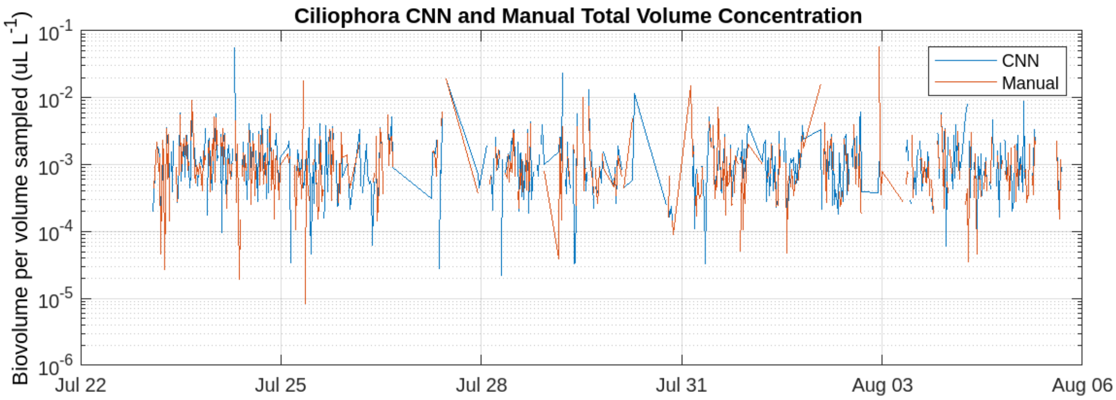


Figure 3.1.1 Ciliophora CNN vs. manual annotations: total volume concentration on a log10 scale

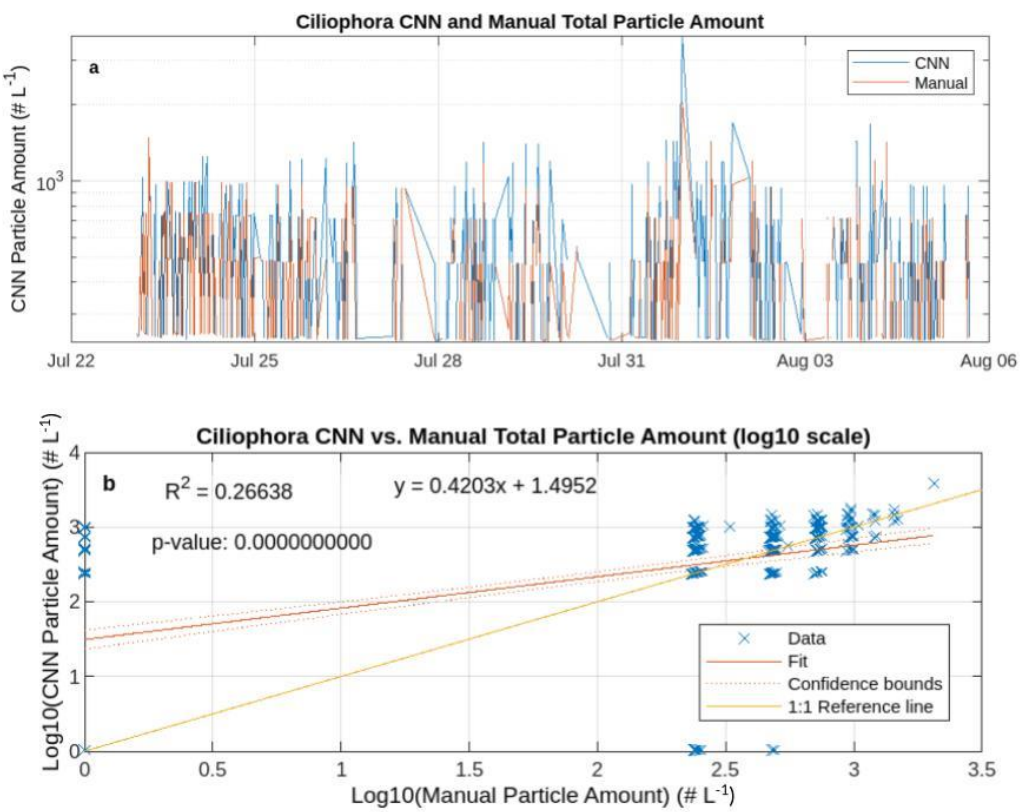


Figure 3.1.2 Ciliophora CNN vs. manual annotations on a log10 scale: particle amount over time (a) and linear regression model (b)

Shifting focus to *Hemiaulus*, as depicted in Figures 3.1.3 and 3.1.4, a discernible trend in the taxon's abundance emerges as the cruise progresses. Notably, this group encompasses approximately five times the number of withheld CNN images compared to the Ciliophora folder, totaling 5,075 images (Table 3.1.1). This increase in image quantity is reflected in the reduced noise observed in Figures 3.1.3 and 3.1.4. Additionally, *Hemiaulus* exhibits an R-squared value roughly tripled that of the Ciliophora equivalent with a value of 0.79. This significant difference underscores a much stronger agreement between manual and CNN validated particles for *Hemiaulus* compared to Ciliophora. The R-squared value of 0.78 indicates that approximately 78% of the variance in the number of CNN particles validated for *Hemiaulus* can be explained by the number of manual particles validated. To further validate this observation, the f-test for Figure 3.2.5b shows a $p < 00001$. Such a high level of agreement suggests a robust consistency between the two methods in classifying *Hemiaulus*, affirming the reliability of the CNN in accurately identifying this taxon.

Figures 3.1.3 and 3.1.4 unveil two significant time periods in *Hemiaulus* abundance throughout the cruise duration: an increase between July 22 and July 29, 2021, followed by a decrease or plateau between July 29 and August 6, 2021. This trend aligns with PARAGON 1's timing, as this cruise captured the decline of an algal bloom.

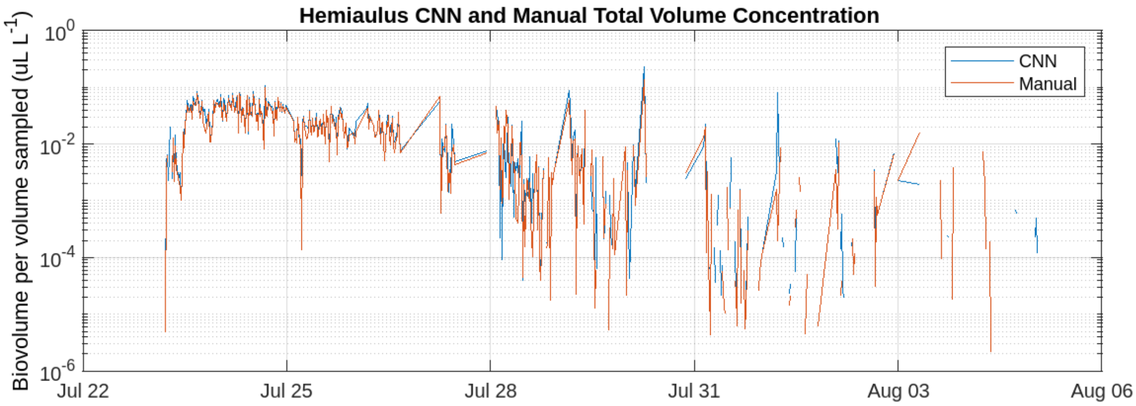


Figure 3.1.3 *Hemiaulus* CNN vs. manual annotations: total volume concentration

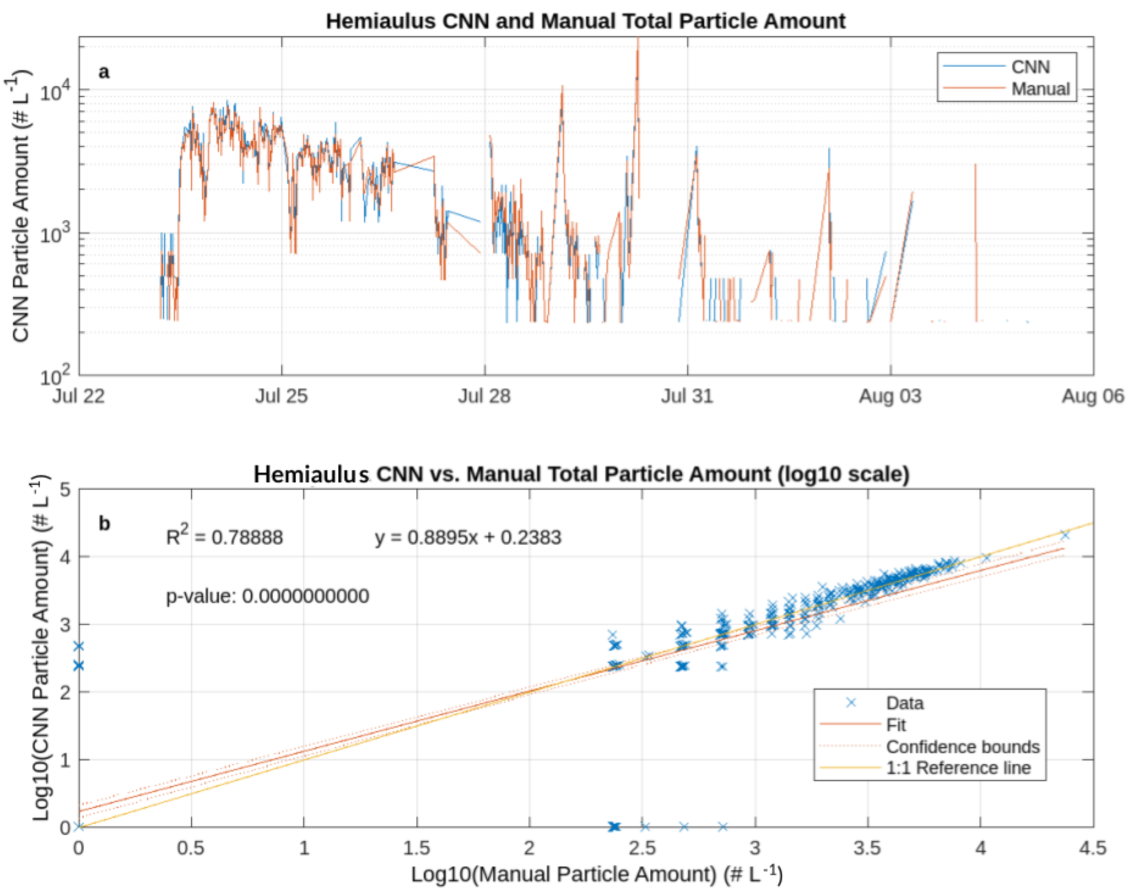


Figure 3.1.4 *Hemiaulus* CNN vs. manual annotations: particle amount over time (a) and linear regression model (b)

A summative perspective of this analysis can first be seen through proportional difference plots, which illustrate the difference between two variables over time. The proportional difference was calculated by:

$$\frac{Theoretical\ value\ (manual) - Experimental\ value\ (CNN)}{Theoretical\ value\ (manual)}$$

Figures 3.1.5a and 3.1.6a reveal that during the initial phase of the cruise (July 22-29, 2021), both Ciliophora and *Hemiaulus* exhibit relatively small differences in total volume concentration. Similarly in Figure 3.1.5b, depicting Ciliophora's proportional difference in particle amount, the difference value remains relatively stable until an increase following July 31, 2021. Figure 3.1.6b shows a contrasting perspective as the proportional difference for *Hemiaulus*' particle amount appears to decrease in the second phase of the cruise (July 29-August 6, 2021). It should be noted that while the F1 results provide insight into the accuracy and recall of the CNN's classifications, the absolute proportional difference plots compare the CNN's classifications with manual classifications in the context of field concentration measurements. These analyses serve different purposes and should not be conflated.

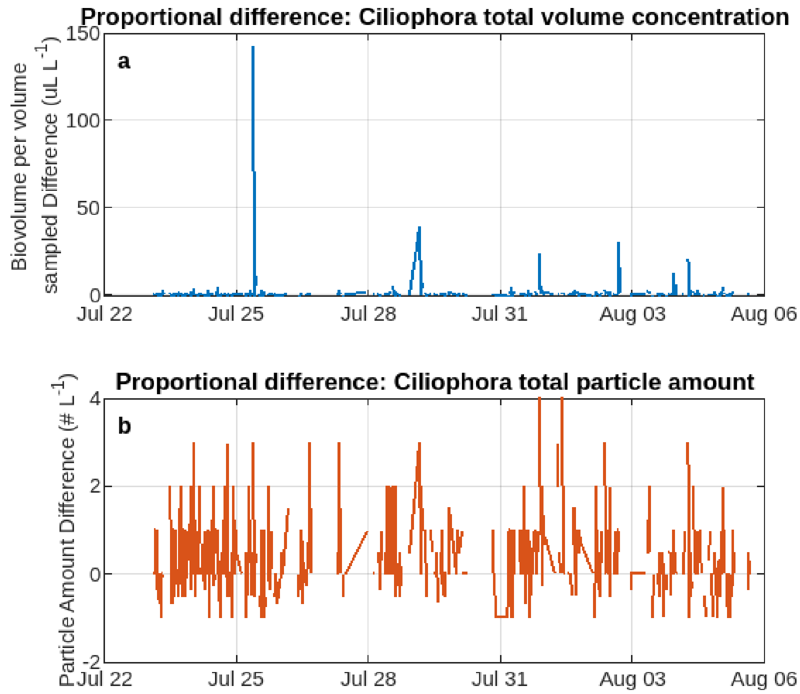


Figure 3.1.5 Ciliophora proportional difference from actual: total volume concentration (a) and particle amount (b).

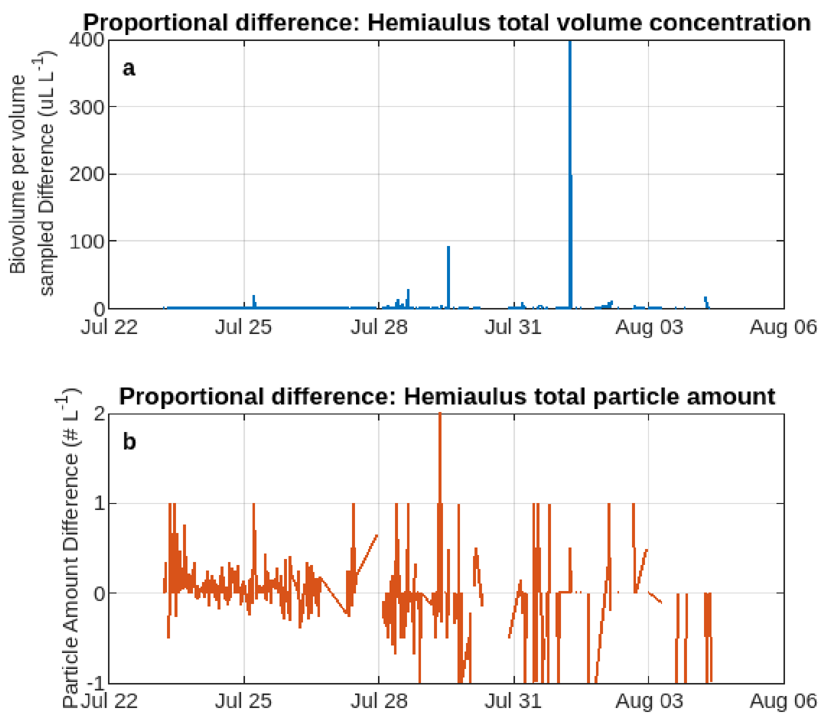


Figure 3.1.6 *Hemiaulus* proportional difference from actual: total volume concentration (a) and particle amount (b).

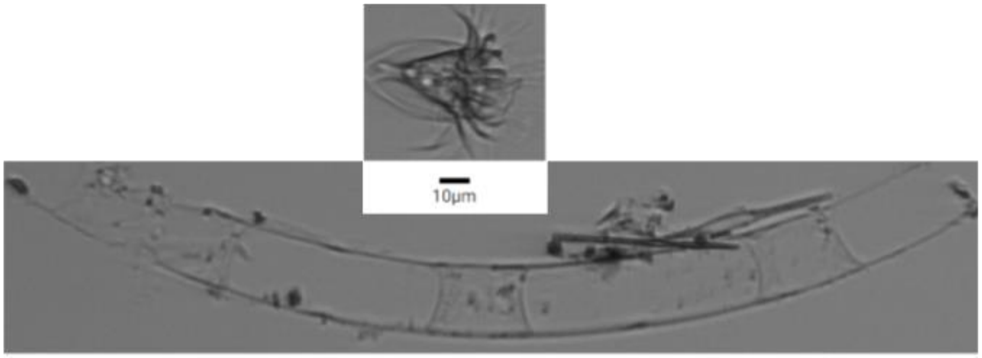


Figure 3.1.7 Size comparison of a *Hemiaulus* and Ciliophora (*Tintinnids*)
(http://ifcbdb.soest.hawaii.edu/timeline?dataset=IFCB_KM2112_PARAGON)

Given *Hemiaulus's* change in morphology over time, as depicted in Figure 3.2.8, it is then hypothesized that these changes may impact the CNN's ability to accurately classify *Hemiaulus*. While highly-degraded cells as seen in in Figure 3.2.8 were not used within the training set, “empty” cells, or those without fluorescence, were included. Due to the clear trends exhibited by *Hemiaulus* between July 22-29 and July 29-August 6, 2021, further analysis can be conducted to assess the agreement between manual and CNN annotations.

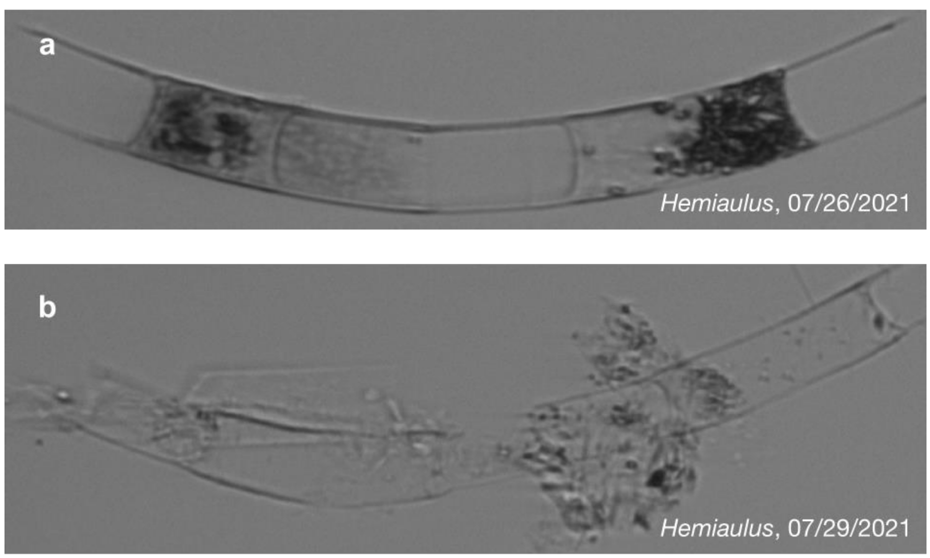


Figure 3.1.8 Comparison of *Hemiaulus* cells depicting the loss of fluorescence and structure over time.

To further understand the differences between manual and CNN annotations, the distributions of the differences between the two sets of measurements at different timepoints, must be analyzed. This analysis involves comparing whether the means of these distributions differ between the two timepoints, essentially asking if there is a significant difference in the patterns of differences observed between the manual and CNN annotations across different phases of the cruise. While this comparison shares similarities with a paired t-test, it represents a more nuanced approach, focusing on the distribution of differences rather than directly assessing paired measurements.

Figure 3.2.9 illustrate box plots for both particle amount and total concentration corresponding to the two time periods of interest for *Hemiaulus*, utilizing data collected from both the CNN and manual annotations. Examination of these figures reveals that both selected time periods' values demonstrate a median close to 1. Additionally, Figure 3.2.9a (July 29 – August 6, 2021) demonstrates a significantly smaller interquartile range than Figure 3.2.9b. This difference arises because when both the CNN and manual annotations did not detect or classify any *Hemiaulus*, the resulting value was zero. Therefore, Figure 3.2.9b is not a useful determinant of the CNN's accuracy. This is further demonstrated in Figure 3.2.10, which depicts the difference between the CNN and manual annotations over time. Through this time series, a notably large amount of zero values are observed, further emphasizing the need to modify the dataset to accurately determine CNN accuracy.

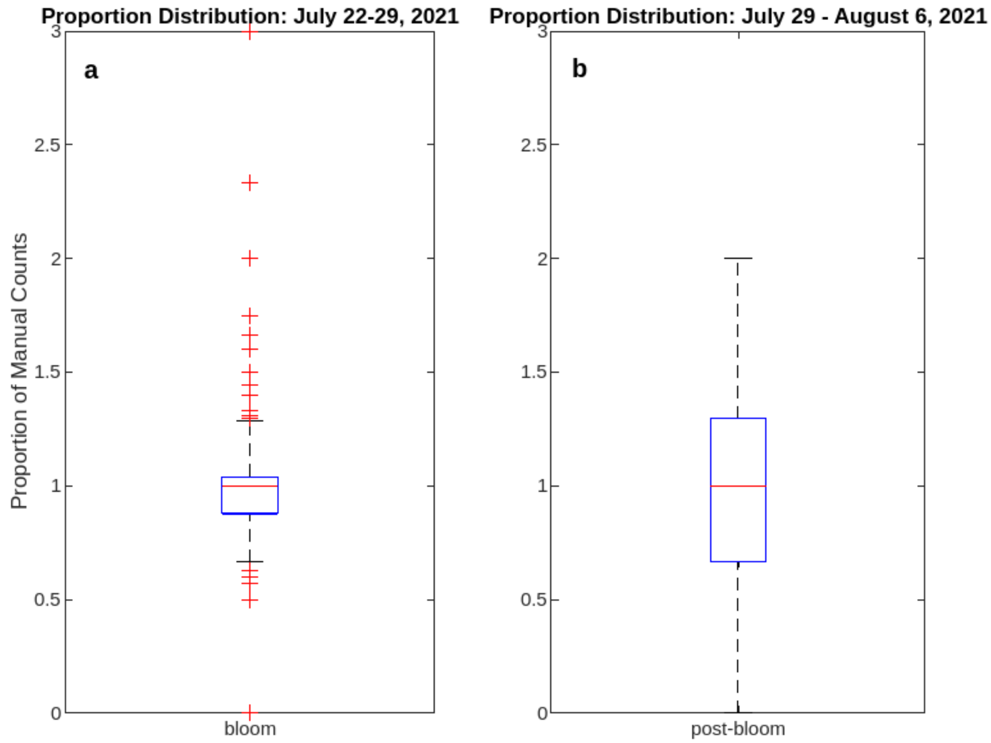


Figure 3.1.9 *Hemiaulus* concentration amount differences proportion distribution box plot for the two time periods of: July 22-29, 2021 (a) and July 20-August 6, 2021 (b)

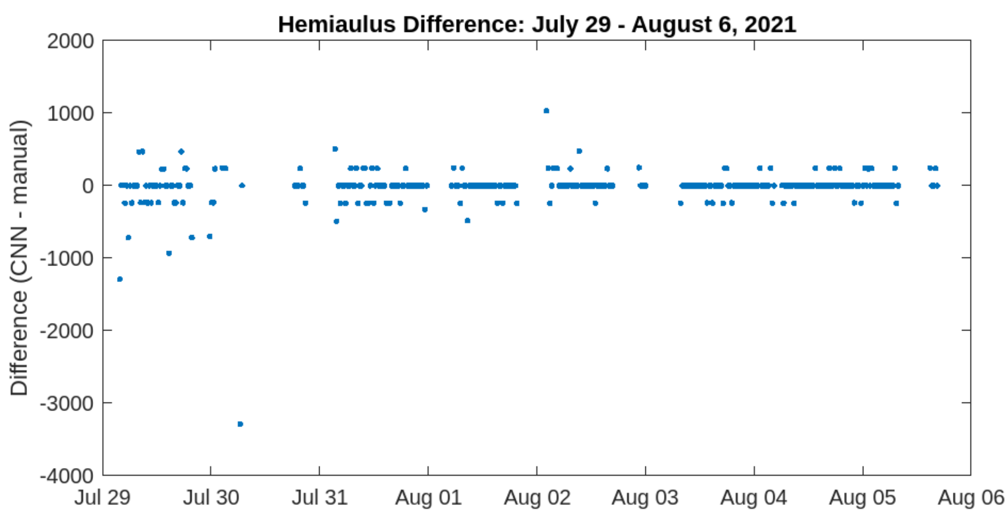


Figure 3.1.10 *Hemiaulus* particle amount differences over time (July 29- August 6, 2021)

To better examine Figure 3.1.9b, all the data points where the manual annotations equaled zero were removed. In doing this, this prevented the plotting of any points that were being classified as zero due to the lack of particle presence at any given time. The resulting values are then depicted in Figure 3.1.11. This then shows that the second half of the cruise had much less outlying data points while still retaining a consistent proportion of 1, indicating a more consistent performance of the CNN during that period. Moreover, this reduction in outlying points suggests reduced variability in the CNN's classifications during this time frame.

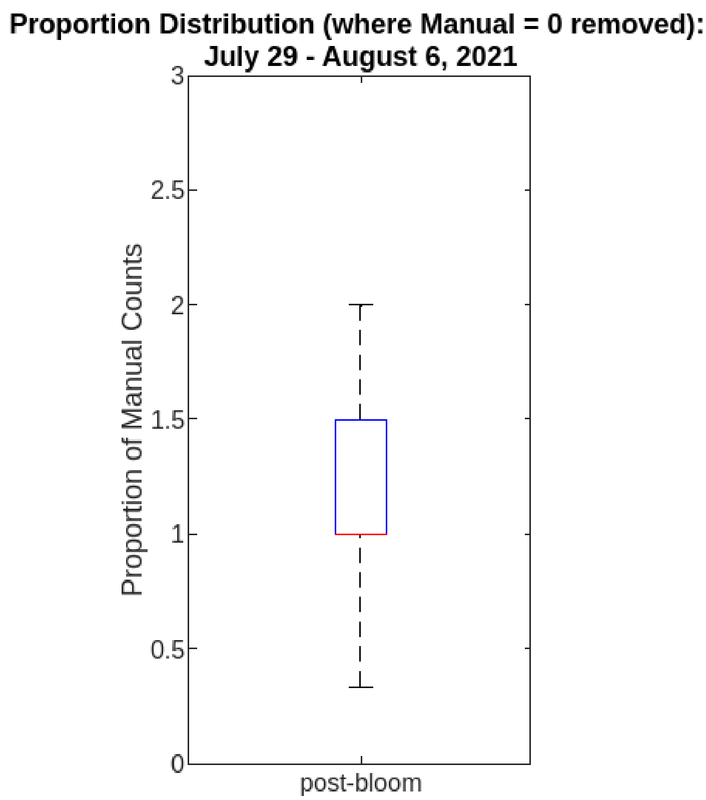


Figure 3.1.11 *Hemiaulus* concentration amount differences proportion distribution box plot for the time period of: July 20-August 6, 2021 where Manual = 0 is removed

Examining Figures 3.2.11a and 3.2.12a, which focus on the peak abundance of *Hemiaulus* during the cruise, both the total concentration and particle amount exhibit p-values < 0.0001, signifying a statistically significant difference between the CNN and manual annotations. Conversely, Figures 3.2.11b and 3.2.12b, which highlight the decline in *Hemiaulus* abundance, demonstrate higher p-values compared to Figures 3.2.11a and

3.2.12a. Despite Figure 3.2.11a displaying a higher p-value than Figure 3.2.11b, it was determined that the difference between the CNN and manual annotation values remained significant.

Among the four plots within Figures 3.2.11 and 3.2.12, only Figure 3.2.12b exhibits a p-value higher than 0.05 of 0.2291, indicating that there is no statistically significant difference between the CNN and manual annotations in this specific case.

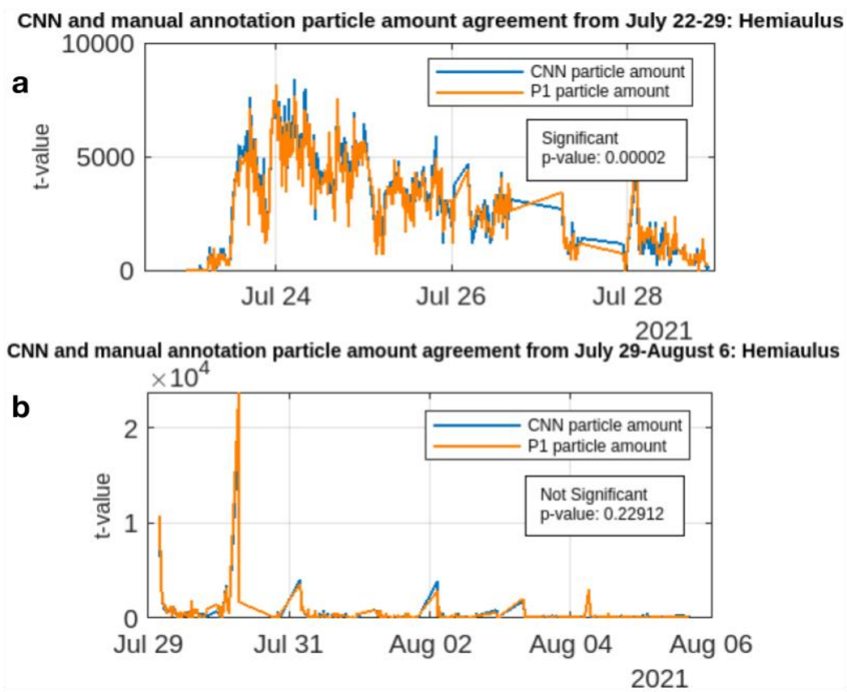


Figure 3.1.12 *Hemiaulus* particle amount agreement t-test for: July 29, 2021 (a) and July 20-August 6, 2021 (b)

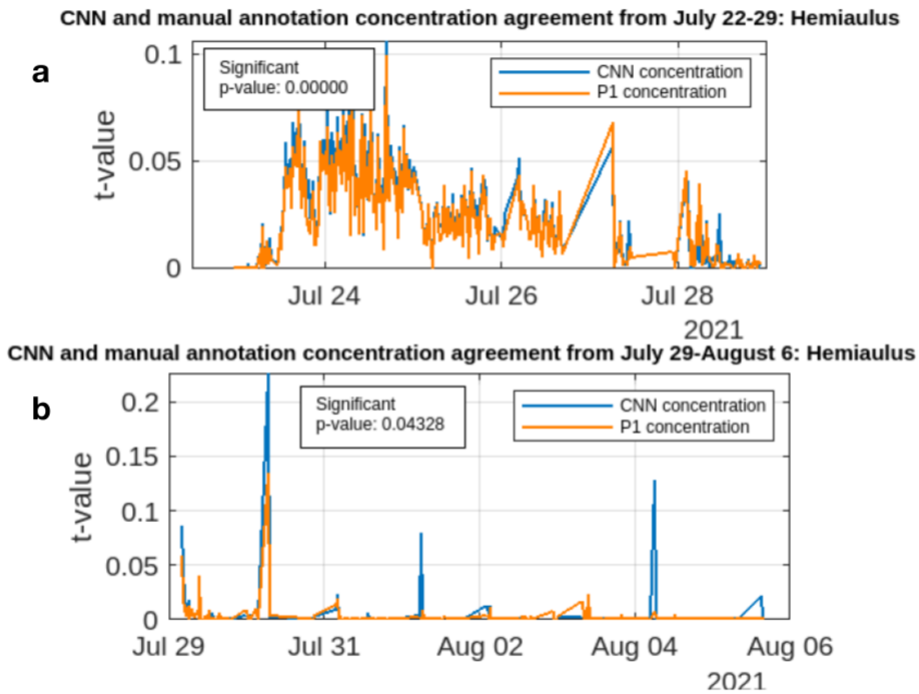


Figure 3.1.13 *Hemiaulus* concentration agreement t-test for: July 29, 2021 (a) and July 20-August 6, 2021 (b)

3.2 Discrepancies

There appears to be a consistent spike in biovolume concentration and particle amount between July 28 and July 31, 2021, as observed in Figures 3.1.2 and 3.2.6. Upon closer examination of the images within this time period (Figure 3.3.1), it was found that they predominantly consisted of clusters of *Hemiaulus* and detrital material. Notably, one cluster documented a diameter of 109 microns. In images such as those seen in the *Hemiaulus* mosaic in Figure 3.3.1, detritus accounts for a significant portion of the image. However, these images are still classified as *Hemiaulus* within manual annotations and training set as it is the dominant taxa within these clusters, affirming that the CNN is indeed classifying these Hemiaulus and detritus clusters as Hemiaulus.

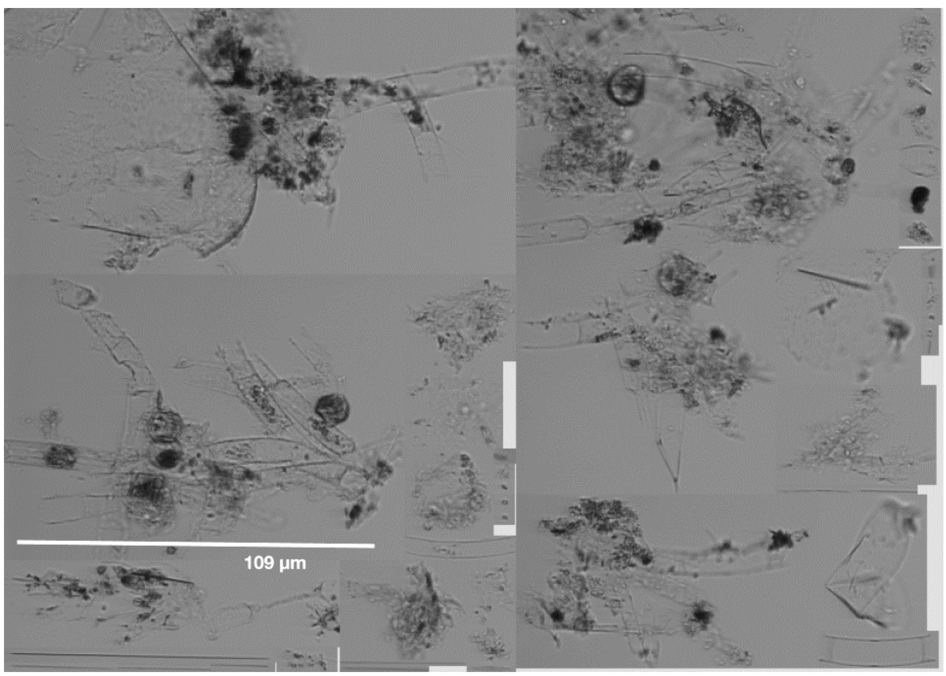


Figure 3.3.1 *Hemiaulus* mosaic from July 30, 2021
(http://ifcbdb.soest.hawaii.edu/timeline?dataset=IFCB_KM2112_PARAGON)

4.0 DISCUSSION

The PARAGON 1 cruise was conducted at the end of an algal bloom, where an abundance of detrital material was expected. As time advanced, *Hemiaulus* followed a consistent pattern of decreasing abundance (Figures 3.2.7 and 3.2.8). Additionally, as discussed in section 3.3, *Hemiaulus* was frequently found within detrital clusters. As the IFCB was only sampling at a consistent depth of 7 meters throughout the cruise, it is unclear whether these *Hemiaulus* clusters sank or aggregated with other material to account for the decrease in biovolume and particle amount over time. However, this observation aligns with the trend of decreasing *Hemiaulus* abundance over time, as the taxa likely sank along with the detrital material (Farnelid *et al*., 2019).

This hypothesis can be further analyzed through a comparison of Particulate Carbon (PC) export rates from PARAGON 1 and the average summer season at Station ALOHA. It has been consistently observed that PC export occurs at Station ALOHA, with the 30-year mean for PC export during summer recorded at 33.5 mg C m^{-2} d^{-1} (Karl *et al.*, 2021). In comparison, the recorded PC export for PARAGON 1 was 49.5 mg C m^{-2} d^{-1}, showing an increased PC export rate from the average summer value.

In contrast, Ciliophora did not exhibit a similar trend of decreasing abundance over time. Although we have no reason to expect this taxa to follow the same temporal trends as Hemiaulus, this discrepancy could be attributed to the lower abundance of this grouping compared to *Hemiaulus*, coupled with the potential impact of limited image quantities on the CNN's accuracy.

It was initially hypothesized that as time progressed, the morphological changes of *Hemiaulus* would impede the CNN's capability to accurately classify the taxon.

However, the analysis conducted in the Results section revealed slight differences in results between the two stages of *Hemiaulus*, contrary to expectations. Despite the structural decline of *Hemiaulus* cells over time and their association with detrital material, which was expected to pose challenges for accurate classification by the CNN, the analyses demonstrated a slight increase in CNN accuracy as the cruise progressed. This unexpected finding suggests several possibilities: the training set provided a reasonable amount of *Hemiaulus* and detrital clusters for the CNN to accurately classify this unique grouping, the decline in taxa variability and abundance at the end of the cruise allowed for easier identification of *Hemiaulus*, or that there are factors beyond morphological changes may have influenced the CNN's performance, warranting further investigation. The observed discrepancy in model performance between *Hemiaulus* and Ciliophora, especially in post-bloom phase, prompts and inquiry into the underlying factors driving these differences. Notably, while the *Hemiaulus* model maintains accuracy even after the bloom (July 29 – August 6, 2021), the performance of the Ciliophora model diminishes. Despite the lower abundance of Ciliophora compared to *Hemiaulus* post-bloom, the CNN's accuracy appears to plateau. This observation suggests the existence of a potential threshold at which the CNN's accuracy diminishes. Exploring this threshold and its implications could provide valuable insights into the factors influencing the CNN's classification accuracy and offer a nuanced understanding of the observed variations in model performance.

Regarding the development of a training set for the CNN, our findings highlight the critical role of incorporating diverse morphological characteristics and environmental conditions representative of the study area. While acknowledging the dynamic nature of certain taxa, such as *Hemiaulus*, and their ability to undergo morphological changes over time, it is important to ensure adequate representation of these images within the training set. This facilitates accurate CNN classification across all stages of NPSG bloom events. Notable, the CNN’s ability to consistent identify *Hemiaulus* throughout its morphological variations underscores the significance of abundance in driving model performance.

Iterative training of the IFCB with updated datasets could enhance the CNN's ability to accurately classify taxa, particularly in dynamic environments such as post-bloom conditions encountered during the PARAGON 1 cruise. By continuously refining the training set based on new data and insights gained from ongoing analyses, the CNN's performance and robustness can improve over time.

5.0 CONCLUSION

This study aimed to develop a precise training set for a CNN to effectively classify IFCB data obtained from the NPSG. The significance of this research lies in addressing the challenge posed by the vast volume of IFCB data, which can be laborious to manually sort and classify. Through harnessing the capabilities of the CNN, this study has demonstrated the potential to expedite the sorting process for future IFCB datasets, particularly in the context of ocean microbiology research. The labor undertaken to generate the training set was extensive, detailed, and requiring cross-validation by multiple experts (the author included), reflecting the necessity for such datasets to be highly curated.

The process of generating a training set for the CNN involved meticulous annotation and classification of thousands of images across several research cruises conducted in the NPSG. For the CNN to produce accurate results, large quantities of high-quality images were manually selected from each of the various cruises in order to best represent the region's particle diversity. Notably, our dataset may suggest that species prone to blooming are more suitable candidates for these models, while those persisting at low abundances may present challenges for classification. As a related point, capturing bloom events effectively trains the models, a concept that this study demonstrates. Furthermore, the necessity for this training set to be highly curated is underscored by the unique characteristics of the NPSG ecosystem. Algal blooms, particularly diatom blooms, are a prominent feature of this region's ecological dynamics (Villareal *et al.*, 2012). Understanding and accurately classifying these blooms are crucial for comprehending the biogeochemical processes occurring in the NPSG. By creating a training set tailored to the specificities of the NPSG, this study enhances the ability to monitor and analyze algal blooms in this region, thereby advancing understanding of its ecological dynamics.

The results revealed successful classification of taxa, particularly *Hemiaulus* and Ciliophora, with varying degrees of accuracy across different time periods of the cruise. Despite the challenges posed by morphological changes and the presence of detrital material, the CNN demonstrated an overall improvement in accuracy as the cruise progressed, particularly in the context of *Hemiaulus*.

In summary, this study has demonstrated the efficacy of CNNs in taxonomic classification of IFCB data from the NPSG region. By providing a streamlined approach to data sorting, the CNN has the potential to significantly enhance the efficiency and scalability of ocean microbiology research, paving the way for future advancements in our understanding of marine ecosystems.

5.1 Future research

Moving forward, the CNN has demonstrated remarkable effectiveness and accuracy, as evidenced by its latest F-1 scores averaging 0.916 (Table 3.1.1). With this significant improvement in the classifier's capabilities within the North Pacific Subtropical region, the analysis of IFCB data from this area can be conducted faster and with greater accuracy. However, it's essential to recognize that the true value of the training set lies not only in its performance compared to the PARAGON 1 dataset but also in its role as the foundation for future classification efforts within the NPSG.

While the high F-1 scores reflect the CNN's proficiency in classifying particles, it is essential to juxtapose these scores with the results of the comparison between manual and CNN outputs. This comparison provides valuable insights into the degree of agreement or divergence between the CNN's annotations and manual annotations, thereby enhancing the interpretability of CNN-generated classifications. Addressing the discrepancies observed between manual and CNN outputs should be a priority for future research efforts, involving more detailed analyses to identify specific types of particles or environmental conditions that pose challenges for accurate CNN classification.

The training set developed along with this study serves as the cornerstone for building classifiers that will be used to categorize all other cruises within the NPSG region. This dataset, meticulously curated and tailored to the unique characteristics of the NPSG ecosystem, provides the framework upon which future research can rely.

Although this paper primarily utilizes IFCB data from PARAGON 1, focusing on the end stages of an algal bloom event, there remains a wealth of IFCB data available from various stages of these blooms for classification. By leveraging the CNN and robust training set for future research cruises, more detailed documentation of the stages of algal blooms in the NPSG can be accounted for, providing valuable insights into the dynamics of these ecosystems.

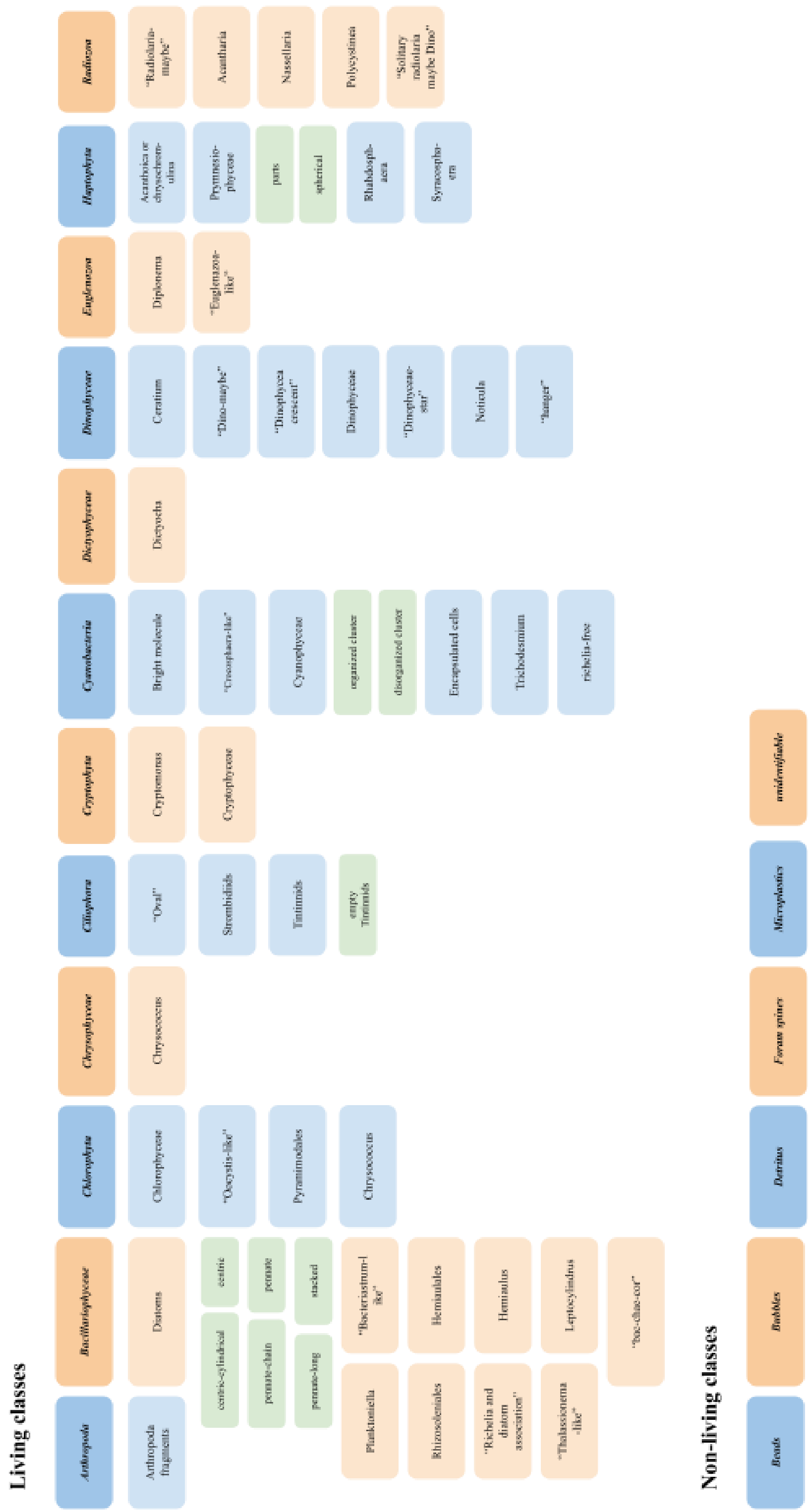
APPENDIX

Final CNN training set and image distribution:

	ZIP	P1	P2	G4	G5	PFIX	Total
Arthropoda_Arthropoda	58	38	10	18	12	26	162
Arthropoda_Arthropoda-fragments	26	114	15	7	5	36	203
Bacillariophyceae_Bacteriastrum-like	0	0	0	58	18	0	76
Bacillariophyceae_Diatom-bac-chae-cor	140	178	45	84	317	376	1140
Bacillariophyceae_Diatom-centric	1494	19	35	82	130	4	1764
Bacillariophyceae_Diatom-centric-cylindrical	104	323	20	4	26	10	487
Bacillariophyceae_Diatom-pennate	1545	558	87	1902	725	1114	5931
Bacillariophyceae_Diatom-pennate-chain	76	0	0	961	169	37	1243
Bacillariophyceae_Diatom-pennate-long	25	0	0	0	0	383	408
Bacillariophyceae_Diatom-stacked	149	21	32	70	187	20	479
Bacillariophyceae_Hemiaulales	40	83	23	3	1	416	566
Bacillariophyceae_Hemiaulus	299	2841	1841	4	0	91	5076
Bacillariophyceae_Leptocylindrus	36	40	9	76	15	5	181
Bacillariophyceae_Planktoniella	2	3	0	28	77	0	110
Bacillariophyceae_Rhizosoleniales	406	483	101	15	52	7	1064
Bacillariophyceae_Richelia-diatom-association	40	130	69	0	0	1	240
Bacillariophyceae_Thalassionema-like	9	20	0	2	2	5	38
Beads	2824	0	0	23	0	63	2910
Bubbles	1739	0	0	123	1	0	1863
Chlorophyta_Chlorophyceae	1191	6	0	11	18	2	1228
Chlorophyta_Oocystis-like	1	157	1	0	0	0	159
Chlorophyta_Pyramimodales	211	58	263	586	450	0	1568
Chrysophyceae_Chrysococcus	20	0	0	6	4	1	31
Ciliophora_Oval	58	36	34	108	73	199	508
Ciliophora_Strombidiids	142	144	113	113	107	40	659
Ciliophora_Tintinnids	87	32	28	46	29	8	230
Ciliophora_Tintinnids-empty	0	45	7	6	2	431	491
Cryptophyta_Cryptomonas	56	831	67	20	17	1	992
Cryptophyta_Cryptophyceae	153	200	418	134	175	26	1106
Cyanobacteria_Bright_molecule	0	0	0	0	0	51	51
Cyanobacteria_Crocosphaera-like	1322	23	58	0	8	4	1415
Cyanobacteria_Cyanophyceae	801	165	8	1	0	1	976
Cyanobacteria_Cyanophyceae-disorganized-clusters	36	27	11	67	70	58	269
Cyanobacteria_Cyanophyceae-organized-clusters	31	2	0	211	142	2	388

Cyanobacteria_Encapsulated-cells	0	0	0	0	0	1000	1000
Cyanobacteria_Trichodesmium	358	165	256	8	2	183	972
Cyanobacteria_richelia-free	405	150	97	1	0	2	655
Dictyophyceae_Dictyocha	79	46	5	159	133	4	426
Dinophyceae_Ceratium	294	30	23	43	68	9	467
Dinophyceae_Dino-maybe	0	1	4	1	0	0	6
Dinophyceae_Dinophycea-crescent	3	4	1	0	0	0	8
Dinophyceae_Dinophyceae	1868	1949	1401	2547	1968	273	10006
Dinophyceae_Dinophyceae-star	0	2	8	3	1	0	14
Dinophyceae_Noctiluca	2	0	0	1	1	0	4
Dinophyceae_hanger	0	0	0	4	5	0	9
Euglenozoa_Diplonema	0	1929	0	2	2	2	1935
Euglenozoa_Euglenozoa-like	0	692	0	0	10	0	702
Haptophyta_Acanthoica-or-chrysochromulina	296	79	122	185	229	1	912
Haptophyta_Prymnesiophyceae	300	729	772	948	665	17	3431
Haptophyta_Prymnesiophyceae-parts	0	164	418	221	220	0	1023
Haptophyta_Prymnesiophyceae-spherical	211	122	43	575	831	72	1854
Haptophyta_Rhabdosphaera	126	150	65	209	500	1	1051
Haptophyta_Syracosphaera	456	0	2	7	8	0	473
Nonliving_Detritus	369	1867	846	1874	1662	898	7516
Nonliving_Foram-spines	0	609	0	0	0	0	609
Nonliving_Microplastics	16	299	35	264	58	409	1081
Radiolaria-maybe	74	0	0	0	0	0	74
Radiozoa_Acantharia	253	190	117	164	247	0	971
Radiozoa_Nassellaria	27	12	0	7	10	2	58
Radiozoa_Polycystinea	7	13	6	2	2	0	30
Radiozoa_Solitary-radiolaria-maybeDino	2	0	0	0	1	140	143
unidentifiable	0	996	0	2271	1715	1521	6503

Training set category breakdown



Appendix Figure 1. Breakdown of training set's 12 living and 6 non-living classes. Subcategories are denoted by a lighter shade than their main category. Subcategories are then further broken down into smaller, morphological subcategories denoted by a green shade.

LITERATURE CITED

Boyce, D.G., Lewis, M.R., Worm, B. (2010). Global phytoplankton decline over the past century. *Nature,* 466 (7306), 591–596. https://doi.org/10.1038/nature09268.

Cael, B. B., & White, A. E. (2020). Sinking versus suspended particle size distributions in the North Pacific Subtropical Gyre. *Geophysical Research Letters*, 47, e2020GL087825. https://doi.org/10.1029/2020GL087825

Chauhan, R., Ghanshala, K. K., & Joshi, R. C. (2018). Convolutional Neural Network (CNN) for Image Detection and Recognition. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC),* 278–282. https://doi.org/10.1109/ICSCCC.2018.8703316

Dai, M., Luo, Y., Achterberg, E. P., Browning, T. J., Cai, Y., Cao, Z., Chai, F., Chen, B., Church, M. J., Ci, D., Du, C., Gao, K., Guo, X., Hu, Z., Kao, S., Laws, E. A., Lee, Z., Lin, H., Liu, Q., … Zhou, K. (2023). Upper Ocean Biogeochemistry of the Oligotrophic North Pacific Subtropical Gyre: From Nutrient Sources to Carbon Export. *Reviews of Geophysics* (1985), 61(3). https://doi.org/10.1029/2022RG000800

Dugenne, M., Freitas, F. H., Wilson, S. T., Karl, D. M., & White, A. E. (2020). Life and death of Crocosphaera sp. in the Pacific Ocean: Fine scale predator–prey dynamics. *Limnology and Oceanography*, *65*(11), 2603–2617. https://doi.org/10.1002/lno.11473

Dugenne, M., Gradoville, M. R., Church, M. J., Wilson, S. T., Sheyn, U., Harke, M. J., Björkman, K. M., Hawco, N. J., Hynes, A. M., Ribalet, F., Karl, D. M., DeLong, E. F., Dyhrman, S. T., Armbrust, E. V., John, S., Eppley, J. M., Harding, K., Stewart, B., Cabello, A. M., … Zehr, J. P. (2023). Nitrogen Fixation in Mesoscale Eddies of the North Pacific Subtropical Gyre: Patterns and Mechanisms. *Global Biogeochemical Cycles*, *37*(4). https://doi.org/10.1029/2022GB007386

Farnelid, H., Turk-Kubo, K., Ploug, H., Ossolinski, J. E., Collins, J. R., Van Mooy, B. A. S., & Zehr, J. P. (2019). Diverse diazotrophs are present on sinking particles in the North Pacific Subtropical Gyre. *The ISME Journal*, 13(1), 170–182. https://doi.org/10.1038/s41396-018-0259-x

Falkowski, P. (2012). Ocean Science: The power of plankton. *Nature* 483, S17–S20. https://doi.org/10.1038/483S17a

González, P., Castaño, A., Peacock, E. E., Díez, J., Del Coz, J. J., & Sosik, H. M. (2019). Automatic plankton quantification using deep features. *Journal of Plankton Research*, 41(4), 449–463. https://doi.org/10.1093/plankt/fbz023

Grabowski, E., Letelier, R. M., Laws, E. A., & Karl, D. M. (2019). Coupling carbon and energy fluxes in the North Pacific Subtropical Gyre. *Nature Communications*, 10(1), 1895–1895. https://doi.org/10.1038/s41467-019-09772-z

Great pacific garbage patch | National Geographic Society. (n.d.). Retrieved December 30, 2022, from https://education.nationalgeographic.org/resource/great-pacific-garbage-patch

Gregg, W.W., Casey, N.W., McClain, C.R. (2005). Recent trends in global ocean chlorophyll. *Geophys. Res. Lett.*, 32 (3), L03606. https://doi.org/10.1029/ 2004GL021808.

Imaging FlowCytobot. (2023). Retrieved from https://mclanelabs.com/imaging-flowcytobot/

Juranek, L. W., White, A. E., Dugenne, M., Henderikx Freitas, F., Dutkiewicz, S., Ribalet, F., Ferrón, S., Armbrust, E. V., & Karl, D. M. (2020). The Importance of the Phytoplankton "Middle Class" to Ocean Net Community Production. *Global Biogeochemical Cycles, 34(12).* https://doi.org/10.1029/2020GB006702

Karl, D. M., & Church, M. J. (2017). Ecosystem Structure and Dynamics in the North Pacific Subtropical Gyre: New Views of an Old Ocean. *Ecosystems (New York),* 20(3), 433–457. https://doi.org/10.1007/s10021-017-0117-0

Karl, D. M., Letelier, R. M., Bidigare, R. R., Björkman, K. M., Church, M. J., Dore, J. E., & White, A. E. (2021). Seasonal-to-decadal scale variability in primary production and particulate matter export at Station ALOHA. *Progress in Oceanography*, 195, 102563. https://doi.org/10.1016/j.pocean.2021.102563

Moberg, E. A., & Sosik, H. M. (2012). Distance maps to estimate cell volume from two-dimensional plankton images. *Limnology and Oceanography, Methods,* 10(4), 278–288. https://doi.org/10.4319/lom.2012.10.278

Nardelli, S. C., Gray, P. C., & Schofield, O. (2022). A Convolutional Neural Network to Classify Phytoplankton Images Along the West Antarctic Peninsula. *Marine Technology Society Journal*, 56(5), 45–57. https://doi.org/10.4031/MTSJ.56.5.8

Olson, R. J., & Sosik, H. M. (2007). A submersible imaging ‐ in ‐ flow instrument to analyze nano ‐ and microplankton: Imaging FlowCytobot. *Limnology and Oceanography, Methods,* 5(6), 195‐203. https://doi.org/10.4319/lom.2007.5.195

PARAGON-1 IFCB Dashboard. (n.d). Retrieved from http://ifcbdb.soest.hawaii.edu/timeline?dataset=IFCB_KM2112_PARAGON

Picheral M, Colin S, Irisson J-O. EcoTaxa, a tool for the taxonomic classification of images. http://ecotaxa.obs-vlfr.fr

Research Cruises : 2021 SCOPE-PARAGON Cruise. (n.d.). Retrieved from https://scope.soest.hawaii.edu/data/scope2021/

Villareal, Tracy A., et al. “Summer Diatom Blooms in the North Pacific Subtropical Gyre: 2008-2009.” *PloS One*, vol. 7, no. 4, 2012, pp. e33109–e33109, https://doi.org/10.1371/journal.pone.0033109.