A Convolutional Neural Network to Classify Phytoplankton Images Along the West Antarctic Peninsula

3 authors:

Schuyler C. Nardelli
United States Geological Survey
8 PUBLICATIONS   131 CITATIONS

SEE PROFILE

Patrick Gray
University of Maine and University of Haifa
28 PUBLICATIONS   528 CITATIONS

SEE PROFILE

Oscar Schofield
Rutgers, The State University of New Jersey
406 PUBLICATIONS   17,908 CITATIONS

SEE PROFILE

A Convolutional Neural Network to Classify Phytoplankton Images Along the West Antarctic Peninsula

AUTHORS

Schuyler C. Nardelli
Rutgers University Center for Ocean Observing Leadership

Patrick C. Gray
Duke University Marine Lab

Oscar Schofield
Rutgers University Center for Ocean Observing Leadership

Introduction

Characterizing the concentration and diversity of marine phytoplankton over relevant ecological temporal and spatial scales has long been the holy grail for aquatic scientists and water quality coastal managers. Diversity and biomass play a central role in aquatic biogeochemical cycles, structuring marine food webs, and driving water quality. The difficulty has been the extremely high diversity of species and morphologies found in the phytoplankton, which reflects a myriad of physical and chemical gradients in nature spanning small scale turbulent mixing to mesoscale circulation processes (Hutchinson, 1961). The development of imaging technologies (Olson & Sosik, 2007) provides for the first time the ability to measure phytoplankton numbers, diversity, and size. These imaging tools offer the potential to document how ocean ecosystems respond to changing environmental conditions.

ABSTRACT

High-resolution optical imaging systems are quickly becoming universal tools to characterize and quantify microbial diversity in marine ecosystems. Automated classification systems such as convolutional neural networks (CNNs) are often developed to identify species within the immense number of images (e.g., millions per month) collected. The goal of our study was to develop a CNN to classify phytoplankton images collected with an Imaging FlowCytobot for the Palmer Antarctica Long-Term Ecological Research project. A relatively small CNN (~2 million parameters) was developed and trained using a subset of manually identified images, resulting in an overall test accuracy, recall, and f1-score of 93.8, 93.7, and 93.7%, respectively, on a balanced dataset. However, the f1-score dropped to 46.5% when tested on a dataset of 10,269 new images drawn from the natural environment without balancing classes. This decrease is likely due to highly imbalanced class distributions dominated by smaller, less differentiable cells, high intraclass variance, and interclass morphological similarities of cells in naturally occurring phytoplankton assemblages. As a case study to illustrate the value of the model, it was used to predict taxonomic classifications (ranging from genus to class) of phytoplankton at Palmer Station, Antarctica, from late austral spring to early autumn in 2017–2018 and 2018–2019. The CNN was generally able to identify important seasonal dynamics such as the shift from large centric diatoms to small pennate diatoms in both years, which is thought to be driven by increases in glacial meltwater from January to March. This shift in particle size distribution has significant implications for the ecology and biogeochemistry of these waters. Moving forward, we hope to further increase the accuracy of our model to better characterize coastal phytoplankton communities threatened by rapidly changing environmental conditions.

Keywords: machine learning, convolutional neural network, polar science, phytoplankton ecology, West Antarctic Peninsula

The West Antarctic Peninsula (WAP) is a highly productive marine ecosystem characterized by large summer phytoplankton blooms that support extensive krill and top predator populations (Ducklow et al., 2013). Rapid warming and melting along the WAP have impacted the phytoplankton community, which has implications for the entire food web.

Midsummer phytoplankton biomass has significantly decreased in the northern WAP, associated with a shift from large-celled diatoms to smaller-celled cryptophytes and mixed flagellates (Montes-Hugo et al., 2009). This shift is concurrent with an increase in low salinity meltwater (Mendes et al., 2013; Moline et al., 2004; Schofield et al., 2017).

The increased spatial coverage of low-salinity surface waters associated with continued glacial and sea ice melt is predicted to increase the prevalence of smaller-celled phytoplankton communities along the WAP, with important implications for food web structure and trophic energy transfer efficiency (Sailley et al., 2013).

The Palmer Antarctica Long-Term Ecological Research project (PAL-LTER; established in 1991) investigates how changes in sea ice along the WAP impact biogeochemistry and pelagic ecosystem dynamics. Phytoplankton taxonomy has previously been characterized using High Performance Liquid Chromatography (HPLC) methods, which use marker pigments to quantify the proportions of the different phytoplankton groups that make up the overall biomass. Molecular analysis such as 16/18S metabarcoding can also be extremely informative and has been used with great success on the WAP (Trefault et al., 2021). However, HPLC lacks more detailed taxonomic resolution (e.g., to genus or species), and both HPLC and molecular analyses lack cell size information that is critical to understanding how warming and melting impacts phytoplankton communities along the WAP. Molecular analyses also require a substantial time investment and cost. Microscopy has been sporadically used in PAL-LTER research (Garibotti et al., 2005), but is extremely time consuming and requires taxonomic experts to manually identify every cell.

To fill this knowledge gap, the PAL-LTER aquired an Imaging FlowCytobot in 2017 (IFCB; McLane Labs, Falmouth, MA, USA). The IFCB is an automated imaging-in-flow submersible cytometer that uses a combination of camera and flow cytometric technology to collect images (see examples in Figure 1) and measure chlorophyll fluorescence and scattered light for each particle (~10–150 µm) in a 5-ml water sample (Olson & Sosik, 2007). These images can be analyzed to determine cell size parameters, and sorted taxonomically to the genus or species level, thus providing more detailed organismal information than HPLC or molecular methods.

However, the IFCB can generate more than 10,000 high-quality images every hour, which can become an immense amount of data over the duration of a research cruise or field season. This volume of data makes manual image identification impractical; therefore, these imaging platforms are often complemented by automated detection systems that allow for rapid and precise classification of plankton communities. Currently, there are two typical machine learning approaches for operational IFCB image classification: (1) a support vector machine (SVM) based on a feature selection algorithm (88% overall accuracy with 22 classes; Sosik & Olson, 2007), and (2) random forest (RF) algorithms (~70% overall accuracy depending on the model and number of classes; e.g., Picheral et al., 2017). Both approaches are

FIGURE 1

Examples of IFCB images: (A) centric diatom chain; (B) *Cylindrotheca* (pennate diatom); (C) cryptophyte; (D) ciliate (microzooplankton); (E) *Phaeocystis* colony (haptophyte); (F) unidentified mixed flagellate; (G) *Eucampia antarctica* chain (centric diatom); (H) detritus; (I) *Licmophora* (pennate diatom); (J) dinoflagellate, likely *Gyrodinium*; (K) *Chaetoceros* chain (centric diatom). Scale bar is 10 µm.
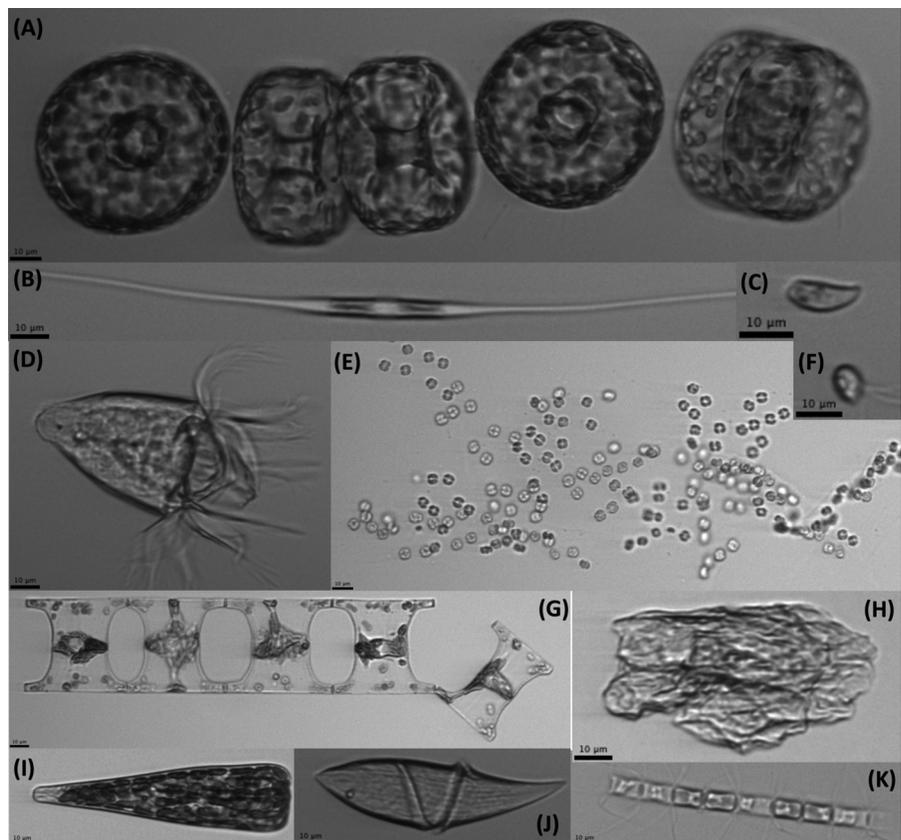
not run on the imagery itself, but rather a set of manually-processed image features such as image texture, cell volume, and gradients.

Striving to increase accuracy beyond the current state of the art and following advancements in the field of computer vision (LeCun et al., 2015), the IFCB community is now transitioning to deep learning methods, including convolutional neural networks (CNNs), due to their improved accuracy in image classification across a broad variety of fields and data domains compared to conventional methods (e.g., RF, SVM). Though often targeting larger zooplankton, there has already been a good deal of success in applying CNNs to planktonic imaging data (Cheng et al., 2019; Orenstein & Beijbom, 2017). There has been mixed success for IFCB data specifically, often challenged by the long-tailed distribution of phytoplankton in the natural environment and the number of cells hovering on the edge of the system resolution, but typically CNNs increase accuracy over conventional methods (Lee et al., 2016). CNNs extract features directly from images. Starting with raw imagery and class labels, the network learns semantically meaningful features as it trains on the data. In theory, extracted features correspond to components of the images relevant to distinguishing between classes, which makes these models highly accurate and well suited for image classification tasks.

Since 2017, the PAL-LTER has collected over 10 million images spanning four field seasons. The goals of our study were (1) to develop a CNN to sort WAP phytoplankton images into taxonomic groups and (2) to apply the CNN to two PAL-LTER

field seasons of IFCB data to demonstrate its utility for characterizing ecological processes along the WAP. A CNN will allow for taxonomic classification of an entire season of collected phytoplankton data in less than a day (~3 million images per PAL-LTER field season, classified at a rate of 101 images per second on a machine with 32 GB of RAM and 12-GB GPU, takes ~8.25 hr to classify). Additionally, the CNN could be used as a tool to characterize phytoplankton communities in the field in near-real time to inform opportunistic sampling strategies (one PAL-LTER sampling station with ~35,000 images takes ~5.75 min to classify). The combination of the IFCB and a high-accuracy automated classification system will allow the PAL-LTER to learn more about shifts in phytoplankton community and size dynamics associated with rapidly changing environmental conditions.

Method
Phytoplankton Image Collection and Processing

IFCB data were collected as part of the PAL-LTER along the WAP over three austral summer field seasons: 2017–2018, 2018–2019, and 2019–2020. For each field season, whole water samples were collected from both the annual January cruise along the WAP and from seasonal sampling at Palmer Station, Antarctica. The annual January cruise samples a fixed grid of stations that extends from Palmer Station in the north (64.77° S, 64.05°W) to ~700 km south near Charcot Island (69.45°S, 75.15°W), and from coastal to slope waters ~200 km offshore (see Figure 1 in Steinberg et al., 2015). At each grid station, five depths are sampled rang-

ing from the surface to hundreds of meters deep depending on oceanographic water column features. Seasonal sampling occurs twice per week from November through March at two stations within 15 km of Palmer Station: Station B (an inshore station with bottom depths of ~75 m), and Station E (an offshore station with bottom depths of ~200 m; see Figure 1 in Schofield et al., 2017). Seven depths are sampled at each station: 0, 5, 10, 20, 30, 40, and 50 m at Station B, and 0, 5, 10, 20, 35, 50, and 65 m at Station E. An IFCB was used to analyze 5 ml from each whole water sample, acquiring one image for each phytoplankton cell/chain in the sample (Figure 1). Samples were passed through a 150-µm Nitex screen prior to analysis to prevent large cells from clogging the IFCB's flow cell. Cells with a major axis length <25 pixels (7.35 µm) were eliminated from the analysis as the image resolution was insufficient to provide clear identification.

Images were processed using default methods and software from Sosik & Olson (2007; https://github.com/hsosik/ifcb-analysis/wiki). Image processing results in a set of 233 features describing each image including fluorescence, scattering intensity, equivalent spherical diameter, area, volume, and other morphometric parameters such as image texture and histogram of oriented gradients. Processed images, metadata, and their associated features were uploaded to the web application EcoTaxa (Picheral et al., 2017; https://ecotaxa.obs-vlfr.fr). Using EcoTaxa, a subset of 18,699 images was visually inspected and manually classified into 38 living groups (taxonomic resolution ranging from genus to class) and

two nonliving groups (detritus and bubbles), with at least 100 images per group. Manual identification of individual cells was performed to the highest possible taxonomic resolution (e.g., most diatoms were identified to the genus level and most phytoflagellates such as cryptophytes, haptophytes, and prasinophytes were identified to class level), with guidance from Hasle et al. (1997) and Scott et al. (2005).

Model Development

The model used in this work is a relatively small CNN (compared to common state-of-the-art architectures) with eight convolutional layers ending in three dense layers and a total of 2 million parameters (https://github.com/patrickcgray/deep_ifcb). The IFCB processing software outputs images clipped to the cell extent and thus object detection is not needed, just image classification. All images are grayscale and can vary in size depending on the cell size. Each image was resampled to 150 × 150 pixels. Since most images are not square, the image is resized so the larger dimension is 150 pixels and then the smaller dimension is filled with black pixels to reach 150 pixels so as not to distort the cell morphology.

The 18,699 manually validated images were divided into training and validation subsets via an 80/20 split, resulting in 14,959 images for training and 3,740 images for validation. The minimum number of images in a class was 116 with an average of 400 for the 38 classes. Training samples (images + features) were augmented to increase training sample size via image rotations, flips, Gaussian noise, and contrast changes. Features were also randomly multiplied by a factor between 0.8 and 1.2 to help augment them and make

the model less sensitive to natural noise in these inputs. After augmentation, a training dataset of 40,000 samples with 1,000 in each class was used to train the CNN. The 3,740 unaugmented images from the validation set, approximately evenly split across classes, were to report validation metrics and choose the stopping point in training. These samples were not seen by the model during training time.

Model precision, recall, and f1-score were calculated for the 40 taxonomic groups (henceforth called "unmerged data") and for eight broader taxonomic groupings (henceforth called "merged data" and includes pennate and centric diatoms, cryptophytes, prasinophytes, mixed flagellates, haptophytes, microzooplankton, and other). The "other" group includes primarily detritus with some bubbles. Precision is defined as true positives divided by the sum of true positives and false positives; it is the proportion of positive identifications that are correct. Recall is defined as true positives divided by the sum of true positives and false negatives; it is the proportion of actual positives that are identified correctly. The f1-score is the harmonic mean of precision and recall. Confusion matrices were also generated showing the percent of manually validated images predicted in each category by the CNN.

Model Evaluation

To further evaluate the model, we analyzed a new subset of 10,269 images selected randomly from the natural distribution of cells in the WAP and filtered by cell major axis length >25 pixels. As a baseline comparison, we used EcoTaxa's RF algorithm to analyze the same images, using a max-

imum of 500 images per group to prevent extreme imbalance in training. Predictions from both models were compared to manual identification of the images. Model precision, recall, and f1-score were calculated for unmerged and merged data for both the CNN and RF models, and a confusion matrix was generated for the CNN.

Case Study: Phytoplankton Seasonal Succession at Palmer Station

After training and evaluation, the model was applied to two austral summer field seasons (2017–2018 and 2018–2019) of IFCB images from Palmer Station (surface samples from Station B) to conduct a preliminary assessment of its utility for characterizing WAP ecological processes. CNN predictions were compared to manual validation of the images by calculating the difference between the two methods to determine the accuracy of the predicted seasonal trends. Additionally, centric and pennate diatoms were separated into size classes (<10 μm, 10–15 μm, 15–20 μm, 20–50 μm, and >50 μm) to help elucidate seasonal diatom diversity trends.

Because phytoplankton phenology is tightly linked to seasonal sea ice dynamics (Vernet et al., 2008), general seasonal succession patterns and diatom diversity data were compared to the timing and concentration of sea ice in the Palmer region. Sea ice metrics were calculated from satellite-derived daily sea ice concentration (%) using the Goddard Space Flight Center Bootstrap algorithm version 3.1. Sea ice duration is the time elapsed between day of advance and day of retreat. All sea ice metrics use the 200-km area south and west of Palmer Station. See

TABLE 1

Confusion matrix for merged taxonomic groups using 10,269 new, random images. Rows are normalized such that each row totals to 100% and the value on the major diagonal represents the recall (true positives divided by true positives and false negatives) in each class. See https://github.com/patrickcgray/deep_ifcb/blob/master/supplemental_figs.md for unmerged confusion matrices.

True label*		Pennate diatoms	Centric diatoms	Cryptophytes	Prasinophytes	Mixed flagellates	Haptophytes	Microzooplankton	Other
	Pennate diatoms	**92.9**	0.8	0.3	0.6	4.8	0.0	0.0	0.7
	Centric diatoms	2.8	**64.3**	5.2	2.4	15.3	0.0	0.0	10.0
	Cryptophytes	9.4	1.0	**65.0**	4.4	19.8	0.0	0.0	0.5
	Prasinophytes	2.6	1.6	0.4	**39.6**	28.0	0.0	0.0	27.8
	Mixed flagellates	11.6	1.5	3.9	7.4	**66.2**	0.0	0.3	9.2
	Haptophytes	0.0	0.0	0.0	0.0	0.0	**100**	0.0	0.0
	Microzooplankton	0.0	16.7	0.0	0.0	16.7	0.0	**66.7**	0.0
	Other	26.9	10.9	4.9	18.4	23.8	0.0	0.3	**14.9**
		Pennate diatoms	Centric diatoms	Cryptophytes	Prasinophytes	Mixed flagellates	Haptophytes	Microzooplankton	Other
	Predicted label**								

*Sample size for True Label groups: Pennate diatoms $n = 1,577$; Centric diatoms $n = 249$; Cryptophytes $n = 2,565$; Prasinophytes $n = 493$; Mixed flagellates $n = 1,085$; Haptophytes $n = 1$; Microzooplankton $n = 6$; and Other $n = 3,475$.

**Sample size for Predicted Label groups: Pennate diatoms $n = 2,788$; Centric diatoms $n = 601$; Cryptophytes $n = 1,898$; Prasinophytes $n = 1,040$; Mixed flagellates $n = 2,304$; Haptophytes $n = 1$; Microzooplankton $n = 16$; and Other $n = 803$.

FIGURE 2

Methods comparison of phytoplankton seasonal succession for the (A–C) 2017–2018 and (D–F) 2018–2019 field seasons. For each merged group, (A and D) show manually validated biovolume data, (B and E) show CNN-predicted biovolume data, and (C and F) show the biovolume difference between CNN-predicted and manually validated methods, with positive values indicating that the CNN overestimated biovolume and negative values indicating that the CNN underestimated biovolume.
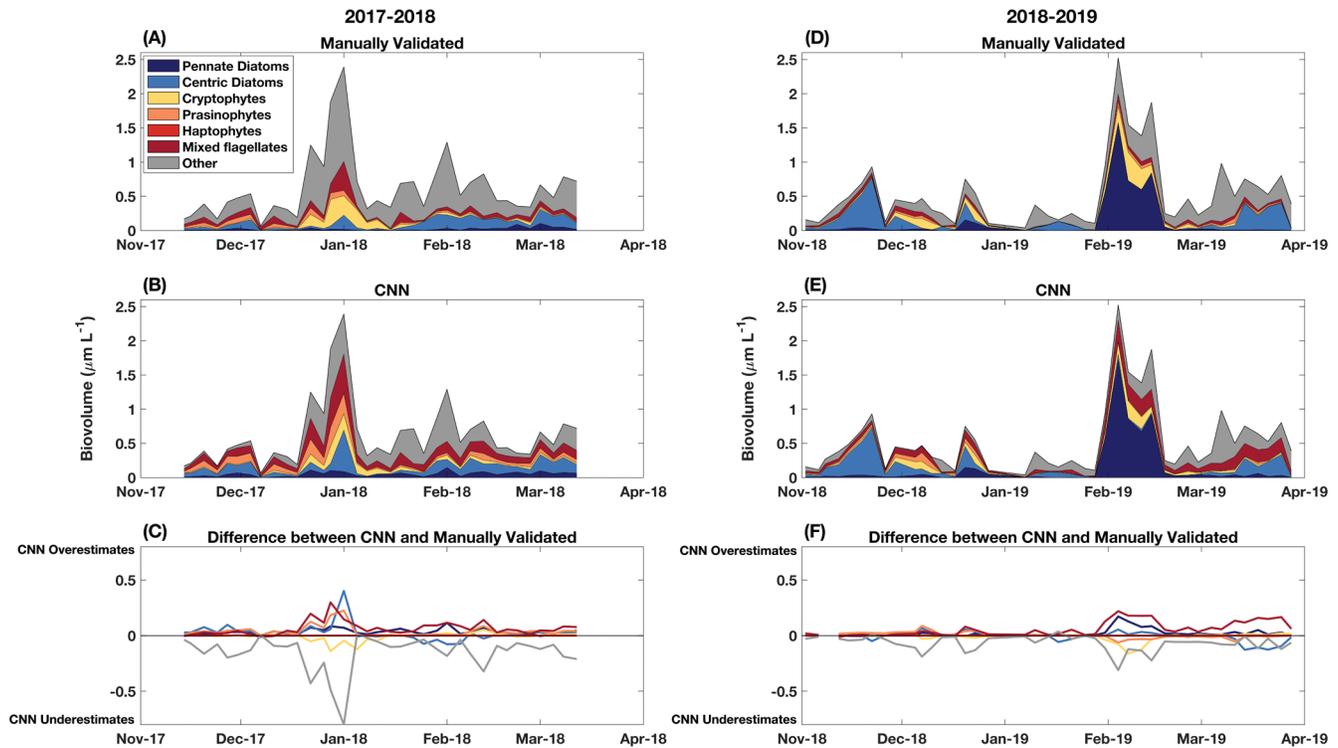
Stammerjohn et al. (2008) for more information.

Results
Model Accuracy

When assessed using the validation dataset, the overall precision, recall, and f1-score of the model were 93.8, 93.7, and 93.7%, respectively. Using the same metrics on training, data led to values just 1%–2% higher, not indicative of any kind of overfitting. After merging the initial set of 40 classes into the eight broader taxonomic groups, the precision, recall, and f1-score of the model all increased to 96.5%. Accuracy per group was >95% for all groups except for microzooplankton (>80%), mixed flagellates (>90%), and other (>90%).

Using the model to predict on the 10,269 randomly selected images resulted in unmerged and merged f1-scores of 46.5% and 47.6%, respectively. While a major decrease compared to the balanced validation metrics, this is a 12% increase in the unmerged f1-score over EcoTaxa's RF model (46.5% vs. 41.5%, respectively), which was trained on the same training data and evaluated on the same validation set. The CNN predicted most accurately for pennate diatoms (92.9%) and performed worse for microzooplankton (66.7%), mixed flagellates (66.2%), cryptophytes (65.0%), and centric diatoms (64.3%; Table 1). Our model was least precise predicting prasinophytes (39.6%) and other cells (14.9%; Table 1). Only one haptophyte was

manually identified in the test dataset but was predicted correctly.

Case Study: Phytoplankton Seasonal Succession at Palmer Station

Overall, the CNN identified important seasonal trends in phytoplankton dynamics. In both 2017–2018 and 2018–2019, peak phytoplankton biovolume occurred midsummer (1 January 2018 and 4 February 2019; Figures 2A–2B and 2D–2E). In 2017–2018, the peak was dominated by a mix of cryptophytes and mixed flagellates, while in 2017–2018, the peak was dominated by pennate diatoms. The CNN also identified late spring and early autumn increases in centric diatoms in 2018–2019 (Figures 2D–2E).

However, there were several discrepancies between the two methods. Figures 2C and 2F show the difference in biovolume between CNN prediction and manual validation. In both years, but particularly 2017–2018, there were many cells manually identified as "other" (e.g., detritus and bubbles) that were classified as mixed flagellates, prasinophytes, and centric diatoms by the CNN (Figure 2 and Table 2). In this manner, the CNN appeared to overestimate the abundance of these groups, and to overestimate total biovolume attributed to living phytoplankton (e.g., the summer phytoplankton peak in 2017; Figures 2A–2C and Table 2). The CNN also underestimated the abundance of cryptophytes (Table 2), especially during peak biovolume in both years (Figures 2C and 2F).

The CNN identified interesting seasonal trends in the diatom community. There was less total diatom biovolume in 2017–2018 compared to 2018–2019 (Figures 3A and 3D). In both years, centric diatoms shifted from a dominance of >20-µm cells in November and December, to a dominance of <20-µm cells in February and March (Figures 3B and 3E). Pennate diatoms were consistently dominated by cells <10 µm, with an increase in biovolume during February and March, especially in 2018–2019 (Figures 3C and 3F). Both years were primarily dominated by centric diatoms, with the notable exception of a large peak in pennate diatom biovolume in 2018 (Figures 3A and 3D).

The winter of 2017 had lower maximum sea ice coverage and shorter sea ice duration than 2018, but a later sea ice retreat (Table 3 and Figure 4). Sea ice cleared the region rapidly in 2017, dropping from 52% coverage in November, to 12% in December, and 3% in January (Figure 4). In 2018, the sea ice retreated earlier but coverage stayed higher in the region into the summer, with 24% coverage in November, 17% coverage in December, and 10% coverage in January (Figure 4).

Discussion
Model Development: Successes and Challenges

In this work, we used a fairly simple CNN instead of common state-of-the-art architectures. While we did initially test a few networks such as Xception, DenseNet, and VGG16, we found they did not perform as well as a smaller CNN with fewer parameters. Our assumption is that this stems from the limited training sample size and the lack of transferability in the features from typical large Red Green Blue (RGB) training datasets (e.g., COCO or ImageNet) for IFCB images. Additionally, we tested a range of simple CNN architectures and found the validation score generally insensitive to minor changes in structure, and thus selected the simplest architecture before validation score decreased.

Overall, we achieved the primary goal of our study: to create a CNN to accurately sort WAP phytoplankton into taxonomic categories. On our balanced validation dataset, the

TABLE 2

Difference (mean, minimum absolute, and maximum absolute) between CNN-predicted biovolume and manually validated biovolume for each merged taxonomic group in each summer field season. Positive values indicate that the CNN overestimates biovolume, and negative values indicate that the CNN underestimates biovolume. Dates where there were <10 cells detected for a particular group via either method were removed from this analysis. Biovolume units are µL L^{-1}.

Merged Groups	2017–2018				2018–2019			
	n	Mean Difference	Minimum Absolute Difference	Maximum Absolute Difference	n	Mean Difference	Minimum Absolute Difference	Maximum Absolute Difference
Pennate diatoms	34	0.03	2.5×10^{-6}	0.11	38	0.02	1.87×10^{-4}	0.17
Centric diatoms	34	0.02	1.8×10^{-5}	0.40	26	−0.006	8.77×10^{-5}	0.13
Cryptophytes	35	−0.007	9.0×10^{-5}	0.14	34	−0.009	5.76×10^{-4}	0.16
Prasinophytes	35	0.04	3.3×10^{-4}	0.22	33	0.003	2.21×10^{-4}	0.09
Haptophytes	1	NA	NA	NA	5	8.67×10^{-5}	0	6.88×10^{-4}
Mixed flagellates	35	0.06	0.003	0.23	39	0.06	1.21×10^{-4}	0.22
Other	35	−0.15	0.005	0.80	41	−0.07	0.002	0.31

Diatom seasonal diversity as predicted by the CNN for the (A–C) 2017–2018 and (D–F) 2018–2019 field seasons. (A and D) Total biovolume attributed to pennate and centric diatoms. (B and E) Total biovolume attributed to different size classes of centric diatoms. (C and F) Total biovolume attributed to different size classes of pennate diatoms.
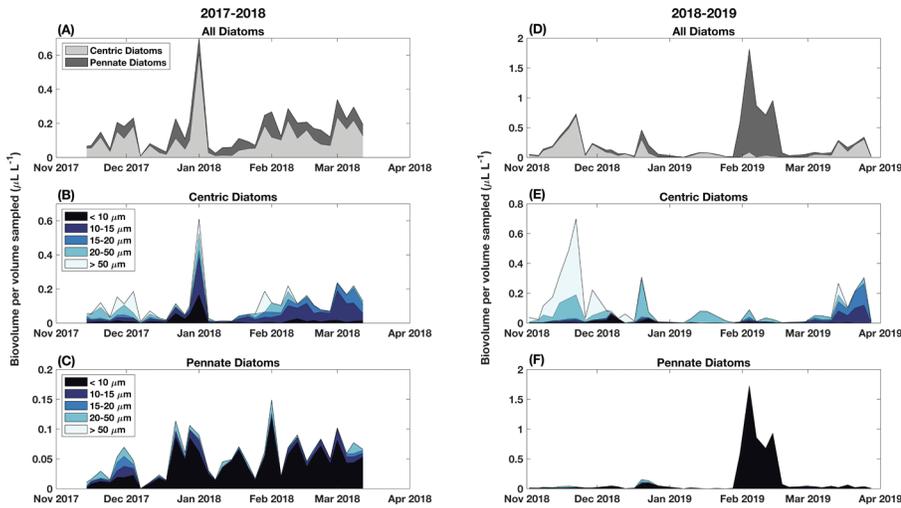
CNN achieved an f1-score of 93.7% with an increase to 96.5% for merged taxonomic groupings. This indicates that our phytoplankton imagery data can be successfully and accurately sorted with machine learning techniques, greatly reducing the time spent classifying these images manually. Absolute comparisons to classification algorithms from other studies in the literature is challenging given different numbers of classes, data filtering schemes, and methods for determining what constitutes test data, but in general, these metrics compare very favorably to other models. The develop-

ment of regional and global phytoplankton-classifying CNNs presents an opportunity to greatly advance our understanding of plankton diversity and ecology.

However, our model f1-score dropped dramatically from 93.7% on the validation set to 46.5% during model evaluation on a new, random dataset with a class distribution representative of that found in natural waters. This low model accuracy when classifying planktonic imagery from the natural environment is a key challenge that plagues many other ecological studies (Culhane et al., 2020; Kerr et al., 2020; Lee et al., 2016) and that must be addressed in plankton classification. One reason for this decrease is the highly imbalanced class distributions of naturally occurring phytoplankton assemblages compared to our model testing dataset (e.g., see *n* values in Table 1). Model categories such as detritus are highly abundant in our dataset, often composing up to 50% of the biovolume in a sample,

while other ecologically important groups, such as large, morphologically distinct diatoms including *Corethron pennatum* and *Eucampia antarctica*, are encountered sporadically in our dataset. A small percentage of detritus misclassified as one of these rare classes can easily overwhelm that category.

Nearly all previous studies report accuracy for a balanced and curated test dataset rather than a random sample of natural waters. Though alternate approaches exist to ameliorate the problem (Tan et al., 2020), typically during model development, a balanced class distribution is necessary to ensure the model equally weights each category during training. For example, if during model development a single class composed 90% of the training dataset, the model could classify every sample as that class, ignoring all others, and be 90% accurate. The gradient descent optimization algorithm used to train the network would be unlikely to learn any other classes. The few studies that do report accuracy in natural samples show a drop-off similar to this study (see Table 2 in Sosik & Olson, 2007; Lee et al., 2016).

Highly unbalanced classes in the natural environment create several model development choices, including whether to exclude, up-sample, or augment low incidence classes, and how specific model classes should be (e.g., high-level classes like diatoms, dinoflagellates, etc., or genus-level classes like *Thalassiosira* and *Gyrodinium*). We tried to strike a balance in our model setup by eliminating extremely rare groups or merging them into broader groups while keeping them morphologically distinct to prevent model confusion. However, there remains a degree of high intraclass variance and interclass similarity

TABLE 3

Sea ice characterization.

Year	Sea Ice Duration (Days)	Date of Sea Ice Retreat
2017	132	December 3
2018	153	November 27

FIGURE 4

Percent sea ice coverage in the 200-km area south and west of Palmer Station during the 2017–2018 season (black) and the 2018–2019 season (blue).

identification, as a CNN can learn relationships between a broad range of image attributes and potential classes. One way we attempted to eliminate a portion of these ambiguous cells was to exclude all cells with a major axis length <25 pixels (7.35 μm) prior to model training. These small cells are below the quantifiable limit of detection based on instrument resolution and, thus, have a high probability of being incorrectly identified. Accurately classifying these smaller cells will likely require techniques other than imaging. Previously described issues of class imbalance can also magnify labelling errors, especially when these errors are within abundant classes such as "detritus."

Case Study: Phytoplankton Seasonal Succession at Palmer Station

Tables 1-2 and Figure 2 showed that there are sometimes major inaccuracies in the CNN predictions compared to manual validation. Thus, it is important to note that the predictions from our model in its current state should not be used without manual validation or verification by other methods (e.g., HPLC). However, Figure 2 showed that the model still captured important seasonal trends in phytoplankton taxonomy, at much higher resolution than the PAL-LTER has previously had, making it a useful tool to obtain a preliminary prediction of species composition. This could be very helpful in determining near real-time sampling strategies, choosing dates/stations/samples of interest to perform a more comprehensive analysis on, or investigating preliminary ecological patterns that could be expanded upon the in future. This case study attempts to explore this last option.

in morphology that was impossible to eliminate (e.g., 14.9% classification accuracy for "other" in Table 1). This challenge can be addressed on the other end of model development, by filtering samples where model uncertainty is high. The CNN outputs a confidence score (from the Softmax classification layer) for each prediction from 0 to 1 that can be used to filter samples below a certain threshold. While potentially increasing the model accuracy, this is not always a reliable metric of uncertainty (Nguyen et al., 2015) and could bias the system against certain classes that are challenging to classify, and thus was not implemented in this work. The heavy long-tailed distribution of phytoplankton species makes it extremely challenging to incorporate all possible classes that could be encountered in the training set. Lessons from work in domain generalization and out-of-distribution clas-

sification (Zhou et al., 2022) need to be incorporated into future phytoplankton classifiers.

Another potential cause of reduced model accuracy is data labelling errors. Theoretically, manual identification of images should be close to perfect. However, in this work and most others, there is often a bias for training and test data that is easily identifiable by manual validation, which prevents test metrics from translating exactly to the wild. There are also many images with conglomerations of cells including detritus and multiple living species. While these may be manually sorted into a category labeled "multiple" or discarded from the analysis, a CNN might sort these images into the most prominent class present within each image. Additionally, morphologically ambiguous cells may be sorted more accurately by a CNN than by manual

Similar to other studies (Saba et al., 2014; Schofield et al., 2017), we found that following a winter with lower sea ice (2017), the phytoplankton community had less diatoms and more mixed flagellates and cryptophytes, and following a winter with higher sea ice (2018), the community was dominated by diatoms. Following trends found in previous years at Palmer Station (Schofield et al., 2017), diatoms dominated in the late spring and early autumn, and there were higher cryptophyte concentrations in December and January.

Along the WAP, the timing and duration of phytoplankton blooms is linked to light availability and sea ice retreat. The growing season initiates in austral spring as daylength becomes longer, solar warming increases, and sea ice melts, combining to stratify the upper water column providing a stable environment for phytoplankton growth (Venables et al., 2013; Vernet et al., 2008). This promotes the development of a large, diatom-dominated spring bloom as we saw in 2018 (Mitchell & Holm-Hansen, 1991; Prézelin et al., 2000). In 2017, there was 52% sea ice coverage in November, likely inhibiting light penetration and subsequent phytoplankton growth. The large drop in sea ice coverage from November to December suggests that sea ice was advected from the region by wind, reducing local melting near Palmer Station and potentially reducing the stability of the upper mixed layer. In 2018, although sea ice retreat was 6 days later than in 2017, November sea ice coverage was only 24%, allowing adequate light for phytoplankton growth. Additionally, sea ice coverage (and subsequent local ice melt) lingered into December and January (17 and 10%, respectively), providing a stable environment for

growth well into the summer. Matching our results, Annett et al. (2010) found that rapid sea ice retreat was associated with lower proportions of centric diatoms during the spring in Ryder Bay, Antarctica.

Despite differences in phytoplankton abundance and community structure between the 2 years, there were similar trends in diatom seasonal succession. The late spring/early summer was dominated by large (>20 µm) centric diatoms timed with sea ice retreat as described above. Moving through summer, centric diatoms became smaller (<20 µm), and the abundance of <10-µm pennate diatoms increased. A potential explanation for this size shift is increased glacial meltwater from January to March (Meredith et al., 2021). Increased meltwater inputs cause stronger surface stratification, inhibiting nutrient injections from deeper waters via mixing and selecting for smaller species with higher surface-area-to-volume ratios to optimize nutrient uptake (Li et al., 2009). Additionally, experimental data from Potter Cove, Antarctica, found that low salinity conditions (30 ppt) shift the diatom community from large centric cells to small pennate cells (Hernando et al., 2015). The authors suggest that this shift is largely driven by differences in tolerance and physiological response to osmotic stress between morphotypes. Thus, increases in glacial meltwater in late summer could cause diatom communities to become smaller and increasingly dominated by pennate cells as we observed.

Conclusions and Next Steps

As illustrated by the case study, our CNN is a step forward for understanding phytoplankton ecology

along the WAP. However, there are still improvements to be made before it becomes a long-term tool for the community. As explained above, an important issue to address is the long-tail distribution of natural phytoplankton assemblages, which makes it challenging to train a model that is accurate across the full range of species it could encounter. This could lead to incorrect ecological takeaways; for example, abundant classes occasionally misclassified as rare ones can entirely change trends for that rare class.

One option to better organize these undifferentiated classes that are challenging to manually discriminate (e.g., "detritus" or "multiple") is to use unsupervised methods (e.g., nonlinear dimensionality reduction, clustering, and manifold learning) to break these classes into several new groups. Defining classes purely via natural clusters in the data, rather than taxonomy, could potentially help models through more easily separable decision boundaries. Semi-supervised classification could also reduce labeled training data needs, and in many cases, unsupervised techniques may help identify anomalous groups (Pastore et al., 2020), or even be sufficient for answering questions about phytoplankton dynamics without any need for supervised classification (Culhane et al., 2020). Another method could be to use a stagewise approach, with a one-class-classifier or binary classification to exclude "detritus" and "multiple" images up front to limit the spread of these issues into the full output range, which is exacerbated by the prevalence of these classes. In tandem to improving the classification itself, per class uncertainty estimates (see Sosik & Olson, 2007) will be critical

to unbiased extrapolation from CNN output to ecological dynamics.

With further increases in model accuracy, we hope our CNN will become a helpful tool for phytoplankton research. Long-term warming and sea ice declines along the WAP contribute to shifts to smaller phytoplankton populations with less biomass (Montes-Hugo et al., 2009), and these trends are likely to continue. Understanding the seasonal and spatial dynamics of phytoplankton diversity is integral to contextualizing how communities will change in the future. Beyond the CNN's ability to complement IFCB-like tools and rapidly classify entire seasons of collected phytoplankton imagery, it can also be used to characterize phytoplankton communities in near-real time (~5 min per PAL-LTER station). Getting a snapshot of species and cell size dynamics soon after collecting a sample would aid in opportunistic sampling while still in the field. This would be invaluable, as research time in Antarctica is both limited and expensive.

The PAL-LTER is not the only group experiencing these challenges: There is a broad phytoplankton imaging user community searching for methods to automate sample classification to reduce the need for manual image validation. Various groups are independently creating phytoplankton CNNs and other models for their study sites of interest. We implore the community to begin reporting their model metrics on data with distributions representative of the natural environment, sharing labeled data openly on freely accessible platforms (e.g., EcoTaxa, IFCB Dashboard), and sharing open and reproducible code for processing and model development. As models im-prove, the community may be able to develop a series of regional models, freely available to download for classification of a dataset, or even a single generalizable model usable for the global ocean. Moving forward toward this vision, it will be critical for oceanographers to collaborate with computer scientists and modelers, applying the best computer vision and classification techniques to these datasets to ultimately better understand phytoplankton dynamics in a changing ocean.

Even beyond oceanography, the challenge of classifying imbalanced, long-tail distributions datasets is a broad issue in applied computer vision, particularly in ecology (Van Horn & Perona, 2017). While there is not yet a consensus on how to resolve this class of problems, there are many promising approaches (Liu et al., 2019). This includes practical iterative human–machine workflows where low confidence classes from a model are continually validated, and the model is continually retrained and improved over time (Miao et al., 2021). Improved loss functions designed for class distributions found in natural environments could be a key step forward (Cao et al., 2019). Improvements are needed to identify unknown classes never seen in training (often model confidence scores are used for this purpose but are not reliable)—there is promise in using the distance in embedding spaces or novel unsupervised approaches (Geng et al., 2021). These technical advances may in time resolve this issue, or they may simply supplement efforts to accurately label more data that will almost certainly improve model performance. Within the IFCB user community, a large, consistent, accurately labeled dataset pooled from all IFCB users is likely one of the best solutions to these issues.

Lastly, a broader challenge the phytoplankton community faces is an increasing scarcity of taxonomists. The development of alternative methods to classify phytoplankton (HPLC, molecular analyses, etc.) has reduced time and effort and/or increased identification accuracy, leading to a diminished need to manually classify phytoplankton cells via microscopy. However, emerging imaging technologies reemphasize the importance of taxonomic experts in helping to build the phytoplankton image libraries required to develop automatic classification systems. The question, then, is how to expand taxonomic expertise within the next generation of phytoplankton researchers to meet this demand while continuing to advance their knowledge of new methods and technologies? Addressing this question will require a forward-looking assessment of the required skills and expertise needed for the future phytoplankton workforce to tackle important issues such as warming oceans, coastal eutrophication, ocean acidification, etc. that impact marine ecosystems.

Acknowledgments

Corresponding Author:

Schuyler C. Nardelli
Rutgers University Center for Ocean Observing Leadership,
New Brunswick, NJ
Email: scn50@scarletmail.rutgers.edu

References

Annett, A.L., Carson, D.S., Crosta, X., Clarke, A., & Ganeshram, R.S. 2010. Seasonal progression of diatom assemblages in surface waters of Ryder Bay, Antarctica. Polar Biol. 33:13–29. https://doi.org/10.1007/s00300-009-0681-7.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems 32, NeurIPS 2019. pp. 1–12. Curran Associates Inc.

Cheng, K., Cheng, X., Wang, Y., Bi, H., & Benfield, M.C. 2019. Enhanced convolutional neural network for plankton identification and enumeration. PLoS ONE. 14(7):e0219570. https://doi.org/10.1371/journal.pone.0219570.

Culhane, E., Haentjens, N., Chase, A.P., Gaube, P., & Morrill, J. 2020. Leveraging unsupervised methods to train image classification networks with fewer labelled inputs: Application to species classification of phytoplankton imagery from an Imaging Flow Cytometer. In: Ocean Sciences Meeting 2020.

Ducklow, H.W., Fraser, W.R., Meredith, M.P., Stammerjohn, S.E., Doney, S.C., Martinson, D.G., … Amsler, C.D. 2013. West Antarctic Peninsula: An ice-dependent coastal marine ecosystem in transition. Oceanography. 26(3):190–203. https://doi.org/10.5670/oceanog.2013.62.

Garibotti, I.A., Vernet, M., & Ferrario, M.E. 2005. Annually recurrent phytoplanktonic assemblages during summer in the seasonal ice zone west of the Antarctic Peninsula (Southern Ocean). Deep Sea Res Part 1 Oceanogr Res Pap. 52(10):1823–41. https://doi.org/10.1016/j.dsr.2005.05.003.

Geng, C., Huang, S.-J., & Chen, S. 2021. Recent advances in open set recognition: A survey. IEEE Trans Pattern Anal Mach Intell. 43(10):3614–31. https://doi.org/10.1109/TPAMI.2020.2981604.

Hasle, G.R., Syvertsen, E.E., Steidinger, K.A., Tangen, K., Throndsen, J., & Heimdal, B.R. 1997. Identifying Marine Phytoplankton. (C.R. Tomas, Ed.). Academic Press. 858 pp.

Hernando, M., Schloss, I.R., Malanga, G., Almandoz, G.O., Ferreyra, G.A., Aguiar, M.B., & Puntarulo, S. 2015. Effects of salinity changes on coastal Antarctic phytoplankton physiology and assemblage composition. J Exp Mar Bio Ecol. 466:110–9. https://doi.org/10.1016/j.jembe.2015.02.012.

Hutchinson, G.E. 1961. The paradox of the plankton. Am Nat. 95(882):137–45. https://doi.org/10.1086/282171.

Kerr, T., Clark, J.R., Fileman, E.S., Widdicombe, C.E., & Pugeault, N. 2020. Collaborative deep learning models to handle class imbalance in FlowCam plankton imagery. IEEE Access. 8:170013–32. https://doi.org/10.1109/ACCESS.2020.3022242.

LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. Nature. 521:436–44. https://doi.org/10.1038/nature14539.

Lee, H., Park, M., & Kim, J. 2016. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In: 2016 IEEE International Conference on Image Processing (ICIP).

pp. 3713–7. IEEE. https://doi.org/10.1109/ICIP.2016.7533053.

Li, W.K.W., McLaughlin, F.A., Lovejoy, C., & Carmack, E.C. 2009. Smallest algae thrive as the Arctic Ocean freshens. Science. 326(5952):539. https://doi.org/10.1126/science.1179798.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S.X. 2019. Large-scale long-tailed recognition in an open world. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2532–41. IEEE. http://doi.org/10.1109/CVPR.2019.00264.

Mendes, C.R.B., Tavano, V.M., Leal, M.C., de Souza, M.S., Brotas, V., & Garcia, C.A.E. 2013. Shifts in the dominance between diatoms and cryptophytes during three late summers in the Bransfield Strait (Antarctic Peninsula). Polar Biol. 36:537–47. https://doi.org/10.1007/s00300-012-1282-4.

Meredith, M.P., Stammerjohn, S.E., Ducklow, H.W., Leng, M.J., Arrowsmith, C., Brearley, J.A., … Waite, N. 2021. Local- and large-scale drivers of variability in the coastal freshwater budget of the Western Antarctic Peninsula. J Geophys Res Ocean. 126:e2021JC017172. https://doi.org/10.1029/2021jc017172.

Miao, Z., Liu, Z., Gaynor, K.M., Palmer, M.S., Yu, S.X., & Getz, W.M. 2021. Iterative human and automated identification of wildlife images. Nat Mach Intell. 3:885–95. https://doi.org/10.1038/s42256-021-00393-0.

Mitchell, B.G., & Holm-Hansen, O. 1991. Bio-optical properties of Antarctic Peninsula waters: Differentiation from temperate ocean models. Deep Sea Res Part A Oceanogr Res Pap. 38(8–9):1009–28. https://doi.org/10.1016/0198-0149(91)90094-V.

Moline, M.A., Claustre, H., Frazer, T.K., Schofield, O., & Vernet, M. 2004. Alteration of the food web along the Antarctic Peninsula in response to a regional warming trend. Glob Chang Biol. 10(12):1973–80. https://doi.org/10.1111/j.1365-2486.2004.00825.x.

Montes-Hugo, M., Doney, S.C., Ducklow, H.H., Fraser, W., Martinson, D., Stammerjohn, S.E., & Schofield, O. 2009. Recent changes in phytoplankton communities associated with rapid regional climate change along the Western Antarctic Peninsula. Science. 323(5920): 1470–3. https://doi.org/10.1126/science.1164533.

Nguyen, A., Yosinski, J., & Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 427–36. IEEE. https://doi.org/10.1109/CVPR.2015.7298640.

Olson, R.J., & Sosik, H.M. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. Limnol Oceanogr Methods. 5:195–203. https://doi.org/10.4319/lom.2007.5.195.

Orenstein, E.C., & Beijbom, O. 2017. Transfer learning and deep feature extraction for planktonic image data sets. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1082–8. IEEE. https://doi.org/10.1109/WACV.2017.125.

Pastore, V.P., Zimmerman, T.G., Biswas, S.K., & Bianco, S. 2020. Annotation-free learning of plankton for classification and anomaly detection. Sci Rep. 10:12142. https://doi.org/10.1038/s41598-020-68662-3.

Picheral, M., Colin, S., & Irisson, J.-O. 2022. EcoTaxa, a tool for the taxonomic classification of images. http://ecotaxa.obs-vlfr.fr.

Prézelin, B.B., Hofmann, E.E., Mengelt, C., & Klinck, J.M. 2000. The linkage between Upper Circumpolar Deep Water (UCDW) and phytoplankton assemblages on the west Antarctic Peninsula continental shelf. J Mar Res. 58(2):165–202. https://doi.org/10.1357/002224000321511133.

Saba, G.K., Fraser, W.R., Saba, V.S., Iannuzzi, R.A., Coleman, K.E., Doney, S.C., … Schofield, O.M. 2014. Winter and spring controls on the summer food web of the coastal West Antarctic Peninsula. Nat Commun. 5:4318. https://doi.org/10.1038/ncomms5318.

Sailley, S.F., Ducklow, H.W., Moeller, H.V., Fraser, W.R., Schofield, O.M., Steinberg, D.K., … Doney, S.C. 2013. Carbon fluxes and pelagic ecosystem dynamics near two western Antarctic Peninsula Adélie penguin colonies: An inverse model approach. Mar Ecol Prog Ser. 492:253–72. https://doi.org/10.3354/meps10534.

Schofield, O., Saba, G., Coleman, K., Carvalho, F., Couto, N., Ducklow, H., … Waite, N. 2017. Decadal variability in coastal phytoplankton community composition in a changing West Antarctic Peninsula. Deep Sea Res Part I Oceanogr Res Pap. 124:42–54. https://doi.org/10.1016/j.dsr.2017.04.014.

Scott, F.J., & Marchant, H.J. (Eds). 2005. Antarctic marine protists. Australian Biological Resources Study. 563 pp.

Sosik, H.M., & Olson, R.J. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol Oceanogr Methods. 5(6):204–16. https://doi.org/10.4319/lom.2007.5.204.

Stammerjohn, S.E., Martinson, D.G., Smith, R.C., & Iannuzzi, R.A. 2008. Sea ice in the western Antarctic Peninsula region: Spatio-temporal variability from ecological and climate change perspectives. Deep Sea Res Part II Top Stud Oceanogr. 55(18–19): 2041–58. https://doi.org/10.1016/j.dsr2.2008.04.026.

Steinberg, D.K., Ruck, K.E., Gleiber, M.R., Garzio, L.M., Cope, J.S., Bernard, K.S., … Ross, R.M. 2015. Long-term (1993-2013) changes in macrozooplankton off the Western Antarctic Peninsula. Deep Sea Res Part I Oceanogr Res Pap. 101:54–70. https://doi.org/10.1016/j.dsr.2015.02.009.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., & Yan, J. 2020. Equalization loss for long-tailed object recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11659–68. IEEE. https://doi.org/10.1109/CVPR42600.2020.01168.

Trefault, N., De la Iglesia, R., Moreno-Pino, M., Lopes dos Santos, A., Gérikas Ribeiro, C., Parada-Pozo, G., … Vaulot, D. 2021. Annual phytoplankton dynamics in coastal waters from Fildes Bay, Western Antarctic Peninsula. Sci Rep. 11:1368. https://doi.org/10.1038/s41598-020-80568-8.

Van Horn, G., & Perona, P. 2017. The devil is in the tails: Fine-grained classification in the wild. arXiv:1709.01450 https://doi.org/10.48550/arXiv.1709.01450.

Venables, H.J., Clarke, A., & Meredith, M.P. 2013. Wintertime controls on summer stratification and productivity at the western Antarctic Peninsula. Limnol Ocean. 58(3): 1035–47. https://doi.org/10.4319/lo.2013.58.3.1035.

Vernet, M., Martinson, D., Iannuzzi, R., Stammerjohn, S., Kozlowski, W., Sines, K., … Garibotti, I. 2008. Primary production within the sea-ice zone west of the Antarctic Peninsula: I- Sea ice, summer mixed layer, and irradiance. Deep Sea Res Part II Top Stud Oceanogr. 55(18–19):2068–85. https://doi.org/10.1016/j.dsr2.2008.05.021.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C.C. 2022. Domain generalization: A survey. IEEE Trans Pattern Anal Mach Intell. 1–20. https://doi.org/10.1109/TPAMI.2022.3195549.