Improving Rare-Class Recognition of Marine Plankton with Hard Negative Mining

Joseph L. Walker
Scripps Institution of Oceanography
jlwalker@ucsd.edu

Eric C. Orenstein
Monterey Bay Aquarium Research Institute
eorenstein@mbari.org

Abstract

Biological oceanographers are increasingly adopting machine learning techniques to conduct quantitative assessments of marine plankton. Most supervised plankton classifiers are trained on labeled image datasets annotated by domain experts under the closed world assumption: all object classes and their priors are the same during both training and deployment. This assumption, however, is hard to satisfy in the actual ocean where data is subject to dataset shift due to shifting populations and from the introduction of object categories not seen during training. Here we present an alternative approach for training and evaluating plankton classifiers under the more realistic open world scenario. We specifically address the problems of out-of-distribution detection and dataset shift under the class imbalance setting where downsampling is needed to reliably detect and classify relatively rare target classes. We apply a hard negative mining approach called Background Resampling to perform downsampling and compare it to other strategies. We show that Background Resampling improves detection of novel particle classes while simultaneously providing competitive classification performance under dataset shift.

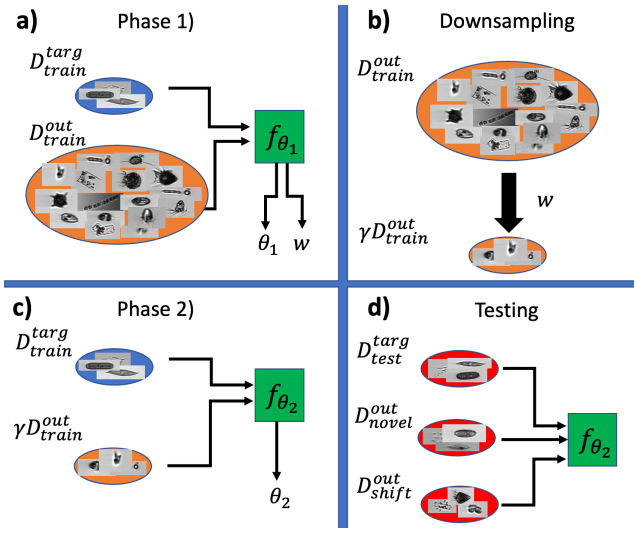

Figure 1. **a)** Using target and background training datasets, denoted as D^{targ}_{train} and D^{out}_{train} respectively, the parameters of a classifier, f_{θ_1}, and background image weights, w, are jointly learned using an alternative optimization approach. **b)** the background dataset is downsampled by interpreting w as resampling probabilities. **c)** the original target dataset and downsampled background dataset is used to train a new classifier, f_{θ_2}. **d)** testing is then performed using target, novel, and dataset shift datasets with f_{θ_2}.

1. Introduction

Marine plankton are a critical component of the biogeochemical processes that are responsible for regulating the climate, supporting the aquatic food web, and producing oxygen [21, 1]. The innumerable ecological roles of plankton make it imperative to monitor their populations as a function of natural and anthropogenic environmental change. Quantifying the fluctuations of individual taxa and the diversity of planktonic communities in response to perturbations is fundamental to understanding planktonic ecosystem dynamics. However, technological limitations constrain our ability to obtain highly temporally resolved time series of individual taxa.

Plankton ecologists are increasingly using *in situ* imaging and deep learning to make population estimates of plankton. Numerous imaging systems have been developed and deployed to study plankton in their natural environment [4, 11, 9, 41]. One of the most widely used plankton imagers is the Imaging FlowCytobot (IFCB), which was developed at Woods Hole Oceanographic Institution (WHOI) to study microorganisms within the 10-100 µm range [36]. The WHOI-Plankton annotated dataset is one of the largest, best maintained labeled plankton image sets available [39].

Together with *in situ* imaging, advances in deep learning have enabled oceanographers to sample ocean populations with higher spatiotemporal resolution and provide the opportunity to produce long, highly resolved time se-

ries of individual taxa. Convolutional Neural Networks (CNNs), a family of deep neural network architectures, have been shown to improve classification accuracy on marine plankton imagery versus ensemble or margin-based methods [38, 16, 39]. CNNs obviate the need for defining handcrafted features by learning the feature extraction and classification process end-to-end. This training paradigm enables the learning of feature representations with more discriminative power [26, 25]. CNNs could therefore alleviate the human cost of manually examining the collected data in order to extract ecologically relevant information.

In many cases, biological oceanographers are only interested in identifying organisms that belong to a small set of classes, referred to here as *target classes*. Some projects are specifically formulated to reduce the number of target classes: Harmful Algal Bloom (HAB) monitoring and parasite tracking to name a few [37, 14, 24, 35, 5, 10, 6]. The annotation process requires a trained taxonomist to search through a large set of images obtained from an experiment or deployment and sort them into ecologically meaningful classes. Either by design or due to circumstance, the available data for classifier training will consist of labeled images associated with the target classes and a large pool of unlabeled data often simply called "other". Classifier training is thus often formulated as an *N+1* classification problem, where there are N target classes and all other object types are mapped to an additional *background* class. The term *background* therefore refers to data that the classifier has been trained to distinguish from target examples, whereas *out-of-distribution* (OOD) refers to data from novel classes that the classifier has not been trained on and is only exposed to during the testing stage or deployment.

The combined abundance of objects from the target classes is often much smaller than the prevalence of all other objects that form the background, the so-called class *imbalance* problem [27, 7, 6, 17]. The issue is exacerbated by the size structure of particles in the ocean and design constraints of imaging systems. There are orders of magnitude more small objects near the lower resolution limit of any optical imaging system. As a result, *in situ* optical imaging systems will image more small indistinguishable objects than large easy-to-identify particles [41, 36, 17]. Therefore, the background class will often be populated by many examples of these small undifferentiated particles.

Training on imbalanced data will encourage a machine learning based classifier to minimize its loss by accurately and reliably classifying majority class examples at the price of diminishing recall of minority class examples [33]. A widely adopted strategy for addressing the imbalance problem is to upsample the minority classes via data augmentation and downsample the larger classes via random downsampling [40, 7]. However, random downsampling is likely to lose crucial information regarding the distribution of possible features that are associated with objects belonging to the background class.

Developing effective machine classifiers for plankton imagery is further complicated by the diversity and constant flux of novel taxa present in the sampling environment [20, 44]. *N+1* classifier training implicitly assumes the classifier's learned representations are robust enough that any and all future objects that do not belong to the set of target classes will be mapped to the background class. But this kind of generalization is not explicitly enforced or encouraged when the classifier is trained and evaluated on datasets that share the same set of labels; a common practice in plankton ecology studies [35, 38, 16, 17, 9, 12].

When a classifier is tasked with labeling unlabeled data, another assumption is made: that the class priors and distribution of features characterizing the classes are unchanging. Changes in these distributions are broadly referred to as dataset shift, and have been shown to impact classifier performance. This problem has received a significant amount of attention in both the plankton ecology [40, 19, 2] and machine learning [34, 54, 18] communities.

Plankton recognition in the open ocean is a particularly challenging endeavor because incoming data is almost guaranteed to be imbalanced, composed of novel classes, and subject to dataset shift. In this work, we present an effective solution to this integrated recognition task for the case where the goal is to identify images belonging to relatively uncommon plankton groups. We examine how the construction of the background class training set via downsampling can impact out-of-distribution detection and dataset shift classification performance. We use a hard negative mining approach called Background Resampling to optimize the downsampling procedure to preserve information regarding the set of features associated with the background class. Our study makes the following three particular contributions:

1. We present a new framework for training and evaluating plankton classifiers that addresses the challenges that are encountered in an open ocean deployment, primarily OOD detection and dataset shift.
2. We show that downsampling via hard negative mining can endow models with greater generalization abilities across a range of challenging test scenarios where other approaches are inconsistent.
3. We benchmark a contemporary OOD detection technique on a fine-grained OOD detection problem.

2. Related Work

2.1. Out-of-distribution detection

Out-of-distribution (OOD) detection methods seek to train a classifier to successfully recognize data that does not belong to the set of target classes. In the case of marine

plankton classification, OOD data would present as novel object classes. Outlier Exposure (OE), a popular new approach for OOD detection, leverages the fact that deep networks produce an estimate of the posterior class distribution. OE measures the entropy of the posterior class distribution to estimate the likelihood that a given data point is OOD [23]. This is implemented with a softmax network layer, which models the probability of an input x being recognized as class i as

$$P(i|x) = \frac{exp\left(w_i^T g\left(x;\theta\right) + b_i\right)}{\sum_{j=1}^{N} exp\left(w_j^T g\left(x;\theta\right) + b_j\right)} \quad (1)$$

where $i \in \{1, 2, ..., N\}$ indexes one of the N target classes. $g(x;\theta)$ denotes the embedding of example x in feature space as a function of network parameters θ. w_j and b_j denote the weight vector and bias terms for class j respectively. The classifier is trained to output a high entropy (i.e., uniform) distribution $P(i|x)$ for background examples, and a confident low entropy distribution for examples from the target classes. For OE, classification is performed by thresholding the softmax scores, where the threshold T is determined empirically from a validation set and calibrated to provide a desired recall on the target classes. If $T < max_i\, P(i|x)$ then the classification is upheld, otherwise, the example is classified as non-target.

OE has been shown to generalize well to OOD examples that come from an entirely different domain [23, 28, 15]. This inspired a wave of OOD detection models which build from the OE concept. [15] introduced Objectosphere loss which aims to minimize the magnitude of $g(x;\theta)$ for background data, which naturally results in low confidence softmax outputs. [30] incorporated scaling and input preprocessing to further increase the softmax output disparity between target and background data. However, these methods are typically tested using OOD and target data from completely separate domains. This is unlike many real-world applications, where OOD data is from the same domain as, and looks very similar to, target class examples. In the case of marine particle classification, particle classes can be visually very similar, which makes OOD detection a challenging problem [12, 48, 32, 55].

2.2. Hard negative mining

Hard negative mining (HNM) approaches seek to identify a set of negative (or background) examples that are likely to generate a false positive [45, 13, 53]. Focusing classifier training on these hard examples has been shown to improve classification performance relative to other downsampling methods [13, 29]. While similar techniques have been applied to datasets consisting of hand-crafted features to predict phytoplankton blooms, to our knowledge, they have not been applied to plankton image classification [49].

3. Dataset

We use the WHOI-Plankton dataset[1] for all experiments [52]. This fully annotated dataset is comprised of 103 classes totaling over 3.5 million grayscale IFCB images, ranging from millions to as few as four examples per class. The bulk of the images belong to the *mix* category which corresponds to small unidentifiable particles. This dataset was amassed over 9 years (2006-2014) from nearly continuous sampling at the Martha's Vineyard Coastal Observatory. An expert taxonomist labeled all images collected in two randomly selected, non-consecutive, single hour time points from each two-week period. Each hour thus represents a complete, independent sample of the plankton population at that point in time. The image data is sorted into subfolders reflecting the image acquisition year.

4. Methods

4.1. Background resampling

Background Resampling (BR) is a HNM approach which aims to ameliorate the class imbalance problem while improving OOD detection [29]. BR assigns each background training image a weight that is proportional to the confidence with which the image is classified as one of the target classes. Then a subset of the background images is sampled according to the image weights which are interpreted as resampling probabilities. This downsampled set is then used to train a new classifier. BR is from the family of OE methods for OOD detection and therefore requires both background and target training datasets, denoted as D_{train}^{out} and D_{train}^{targ} respectively. D_{train}^{out} and D_{train}^{targ} are used to train the parameters θ_1 of a classifier, denoted as f_{θ_1}, to output high and low entropy distributions over the softmax outputs (eq. 1) respectively. The BR procedure can be broken into two distinct phases:

Phase (1): Using D_{train}^{out} and D_{train}^{targ}, learn the background image weights w and θ_1 with the alternative optimization

$$\theta_1^{(t)} = \underset{\theta_1}{argmin} \left[L_{targ}\left(\theta_1\right) + \alpha L_{out}(\theta_1; w^{(t-1)}) \right] \quad (2)$$

$$w^{(t)} = \underset{w}{argmax} \left[L_{targ}(\theta_1^{(t)}) + \alpha L_{out}(\theta_1^{(t)}; w) \right] \quad (3)$$

where L_{targ} is the cross-entropy classification loss term used to penalize incorrect classifications for target class data. L_{out} is the loss term that penalizes overly confident predictions on background examples and is defined as the Kullback-Leibler divergence between the uniform distribution and the softmax outputs. The solution to this system of equations is approximated using a differential relaxation (stochastic gradient descent) and batches of images from

[1] doi:10.1575/1912/7341

both D^{out}_{train} and D^{targ}_{train}. $w^{(t)}$ is defined as the set of image weights that *maximize* the associated loss at time step t, where t denotes the batch number. This ensures that the reweighting algorithm will assign high weight values to images from D^{out}_{train} that are difficult to classify, guaranteeing that the resampling process selects challenging background images that are visually very similar to the target class examples. The adversarial nature of the iterative process – classification vs selection of difficult examples for the classifier – is critical to accurately learning the boundary between target and background classes. The hyperparameter α controls the trade-off between learning to output confident and low-confident predictions for target class and background examples respectively. For all experiments, we set $\alpha = 0.5$ in accordance with the standard OE default [23, 29].

Phase (2): A resampling percentage γ is empirically set and will typically reflect the degree of imbalance between the background and target classes. Using the learned background image weights, the background class is downsampled to γ percent of its original size. This is done by selecting each background image x_i, associated with weight w_i, independently with probability $p_i = min\left(1, \frac{\gamma D^{out}_{train}}{\sum_{j=1} w_j} w_i\right)$. Once the background image weights are obtained, f_{θ_1} is discarded. The downsampled background dataset and full target training dataset are then used to train another classifier, denoted as f_{θ_2}. Testing is then performed with f_{θ_2}. For all experiments, we use $\gamma = 0.05$. A schematic diagram of the entire process is shown in Fig. 1.

4.2. Experimental setup

Our experiments were designed to simulate a scenario where a biological oceanographer is interested in tracking the prevalence of a few relatively rare plankton groups. The abundance of these groups can fluctuate over a very large background class whose images are not of interest. We construct subsets of the WHOI-Plankton dataset to perform our experiments:

Target Data. The classes to be detected, or target classes, are *Ceratium*, *Dinobryon*, *Pleurosigma* and *Ephemera*. All the available data for these four classes is denoted as D^{targ}. Both the *Dinobryon* and *Ceratium* genus are associated with algal blooms. The *Pleurosigma* genus is of interest in biomedical applications because they are believed to produce rare but important organic compounds [3, 56]. The *Ephemera* class is taxonomically ambiguous, but previous studies have identified it as difficult to classify because of its visual similarity to other organisms in the WHOI-Plankton dataset [55]. Images from these four classes were drawn from each year of the WHOI-Plankton dataset but capped at 900 examples per class. This was to prevent significant class imbalance within the set of target classes and to provide a realistic amount of data that could be obtained relatively easily from a low-budget data annotation campaign.

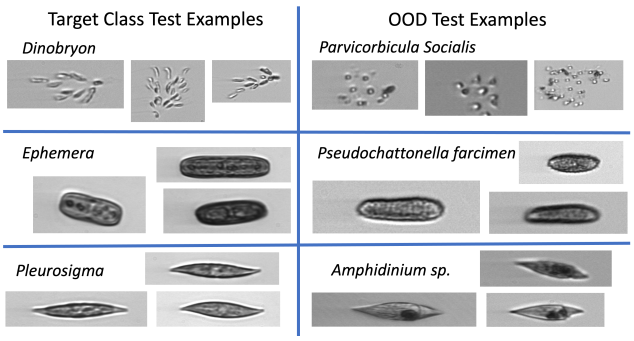

Figure 2. Three examples, selected by a human annotator, from classes in D^{targ}_{test} (left column) and D^{out}_{novel} (right column) showing the morphological similarity between specimen of these classes

Training Data. 55% of data from D^{targ} (1783 images) was randomly selected (stratified by class) to serve as a training set for the target classes and is referred to as D^{targ}_{train}. The background training dataset, denoted as D^{out}_{train}, consists of all images from the year 2006 (totaling 134,293 images) that do not belong to the target classes.

Validation data. 22% of data from $D^{targ} \backslash D^{targ}_{train}$ (311 images), referred to as D^{targ}_{val}, was randomly selected to be used to learn the OOD detection decision threshold T.

Testing Data. We construct three different testing datasets to assess classification performance on target examples, novel object classes, as well as a dataset shift scenario where the prior probabilities of the classes in D^{out}_{train} are subject to change. The remaining 78% of data from $D^{targ} \backslash D^{targ}_{train}$ (1125 images) was selected for testing target class classification and is referred to as D^{targ}_{test}.

There are 13 classes in the WHOI-Plankton dataset that are not present in $D^{targ}_{train} \cup D^{out}_{train}$. These classes were used to form a hold-out set of novel classes to test OOD detection performance. This hold-out set is referred to as D^{out}_{novel} (totaling 1112 images). Many of the classes in D^{out}_{novel} look remarkably similar to the classes in D^{targ}_{test}, underscoring the difficulty of OOD detection in plankton imagery (Fig. 2). For OOD detection testing, we utilize datasets D^{targ}_{test} and D^{out}_{novel}. Since they are approximately the same size, testing on the combination of these sets implicitly assumes that the number of target class and OOD examples is similar. This may be realistic if the novel classes are relatively rare, but in practice the target examples are often rare compared to non-target examples. Therefore, we test classifier performance using several ratios of target to OOD examples using subsamples from D^{targ}_{test} and the full D^{out}_{novel} set.

While we wish to develop classifiers that reliably detect novel examples, it is important that improved OOD detection does not diminish classifier performance on other important aspects of plankton recognition, such as recognition under dataset shift. To simulate a real world deployment,

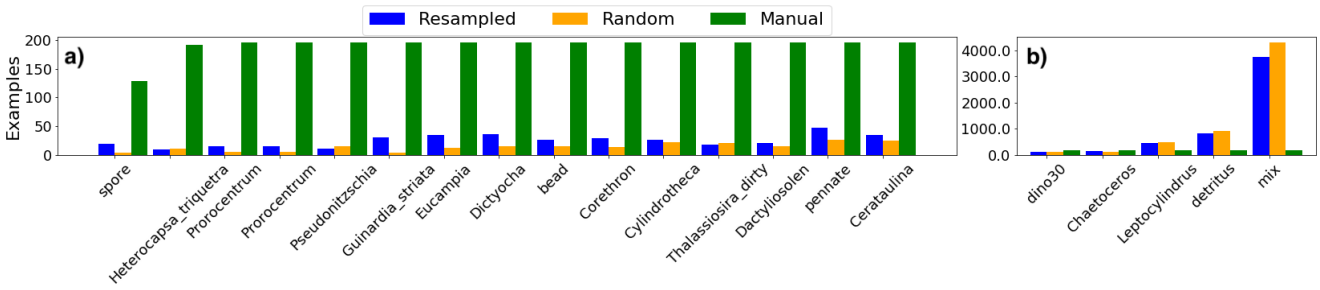

Figure 3. Class distribution of the downsampled background sets generated from each of the downsampling strategies. **a)** class distribution for the 15 least abundant classes that have at least 100 examples in the original D^{out}_{train} set. **b)** class distribution for the five most abundant classes in D^{out}_{train}. Note the change in vertical axis scale between **a)** and **b)**.

each day's worth of data from the WHOI-Plankton 2014 image directory is used to test each classifier under a dataset shift scenario, where the classes are the same as D^{out}_{train}, but the prior probabilities and appearance of the background classes are subject to change. This testing dataset is denoted as D^{out}_{shift} (totaling 329,835 images). For dataset shift testing, each day's worth of data from D^{out}_{shift} is combined with a random 75% sample of D^{targ}_{test}, ensuring that the background and target classes are subject to dataset shift.

4.3. Models and training

The dataset size ratio D^{targ}_{train}:D^{out}_{train} is approximately 1:75. For the class with the fewest examples in D^{targ}_{train}, denoted as $D^{targ\_min}_{train}$, the ratio $D^{targ\_min}_{train}$:D^{out}_{train} is approximately 1:400. Training on data with this degree of imbalance typically results in poor detection for minority class examples [7, 49, 50]. Instead, it is common to downsample the majority classes and upsample the minority classes to improve results [51, 31, 47]. Using our downsampling percentage γ = 0.05, the ratio $D^{targ\_min}_{train}$:γD^{out}_{train} is approximately 1:20. To fully balance the classes, we upsample the four target classes using random image rotations.

It is possible that all images of a rare background object class are lost when using random downsampling. This increases the risk that the classifier trained on the randomly downsampled data will assign examples of that class to one of the target classes. In the case where the images in D^{out}_{train} are assigned their true class label, D^{out}_{train} can be downsampled by taking an even number of examples from each class within D^{out}_{train}, referred to as class-balanced downsampling. This guarantees that every class is represented in the downsampled dataset, therefore maximizing the feature diversity in this new downsampled set. For this reason, we consider the scenario where a fine-grained labeled image set is available and class-balanced downsampling is possible. In this setting, images associated with the background meta-class are assigned their true class label, but still trained as one class.

All classifiers are fine-tuned ResNet-18 models [22], pre-trained on ImageNet [46]. Three downsampling methods are used to train the classifiers and compared. Each classifier uses the same training procedure, using D^{targ}_{train} but a different subset of D^{out}_{train}:

1. **Resampled**: Trained on a subset of images from D^{out}_{train} of approximate size γD^{out}_{train} that was selected according to the resampling probabilities described in Sec. 4.1.
2. **Random**: Trained on a subset of D^{out}_{train} of approximate size γD^{out}_{train} drawn randomly. This represents the standard downsampling approach and therefore serves as a baseline for comparison.
3. **Manual (class-balanced)**: Each subclass within the background meta-class is downsampled by capping the number of examples at 196. This upper limit was determined empirically to yield a downsampled background meta-class of approximate size γD^{out}_{train}. Note that this mode of downsampling is only possible if labeled data is available for background examples. Using this classifier as a baseline, we seek to determine whether BR is beneficial when labeled data is available for background examples.

For each phase (defined in Sec. 4.1), we used image batch sizes of 64 from both D^{out}_{train} and D^{targ}_{train}. The weight learning optimization is performed until the loss associated with the background image weights (eq. 3) fails to decrease for 10 epochs. For phase 2 training, each classifier was trained on its respective subset of D^{out}_{train} for 50 epochs, using an initial learning rate of 0.0003 which was reduced by a factor of 0.5 after every 10 epochs.

Predetermining the number of epochs is common for studies involving OE [23, 28, 29, 42] since the validation set is used to learn the decision threshold T rather than to perform early stopping. The values for all other hyperparameters used during training are those of [29]. After training, a decision threshold is calculated for each classifier as the largest threshold that allows for 95% recall of examples from D^{targ}_{val}.

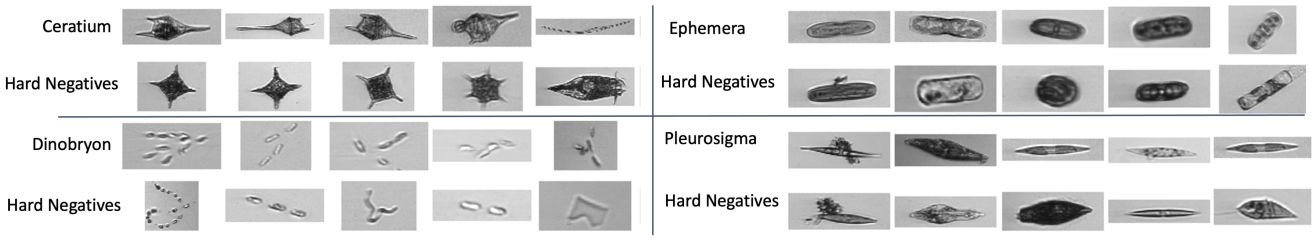

Figure 4. Samples from each target class and their respective hard negatives. Ten background classes are represented by the hard negative examples.

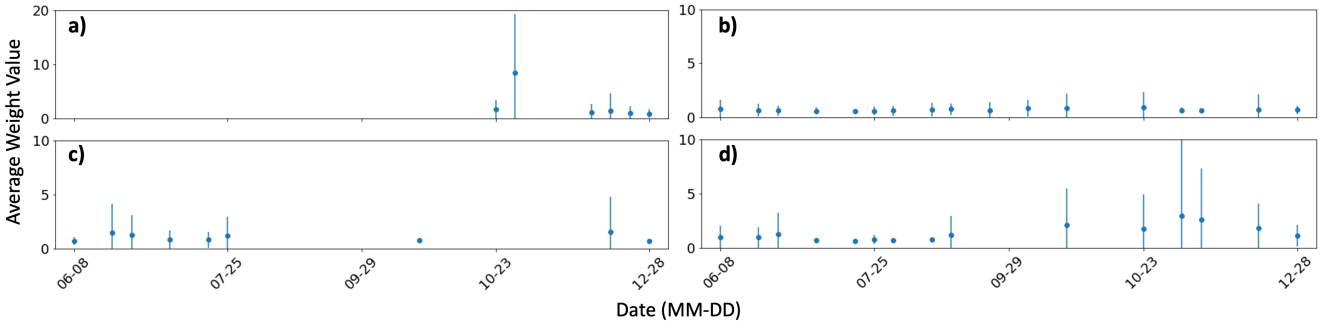

Figure 5. Daily average background image weight value associated with each of the hard background classes presented in the order: **a)** *Dictyocha*; **b)** *detritus*; **c)** *Skeletonema*; **d)** *pennate*. The bars represent the standard deviation of the weight values. Note that the vertical axis scale in **a)** is greater than that of **b)-d)**.

4.4. Performance metrics

We use two metrics to assess model performance:

1. **F1 score** is a metric for binary classification, defined as the harmonic mean of the precision and recall for a given class. The F1 score is calculated for each class, by treating all other classes as a single class. The F1 scores are then uniformly averaged over each class.
2. **Accuracy (overall precision)** is the fraction of correctly classified images from the four target classes and background class.

Using these two metrics, we benchmark each classifier on an OOD detection task (Sec. 5.2) and dataset shift scenario (Sec. 5.3).

4.5. Alternative target classes

To test the generalization of BR, we repeated all experiments for five different sets of four target classes. These classes were randomly selected but were restricted to classes with 600-10,000 examples. This restriction was added to preserve the D_{train}^{targ}:D_{train}^{out} ratio across all sets of target classes. For all target classes considered, the number of examples per class was capped at 900.

5. Results

5.1. Downsampling analysis

Based on the class frequency distribution, BR draws disproportionately more examples from the minority classes compared to random downsampling (Fig. 3). While manual downsampling also samples disproportionately from the minority classes, it creates a class distribution that is radically different than the natural population distribution.

To visualize difficult OOD samples, we drew examples from the background class that were classified into one of the target labels with high confidence (Fig. 4). These "hard negatives" reveal that the classifier confused background examples from more than just a select few classes. Four background classes are represented in these hard negatives more than others: *Dictyocha*, *detritus*, *Skeletonema*, and *pennate*. We refer to these as "hard background classes".

The daily average image weight values associated with the hard background classes vary substantially between classes and over time (Fig. 5). This information allows us to determine whether examples within the hard classes were consistently ascribed higher weight values or if hard negatives are outlier examples for those classes.

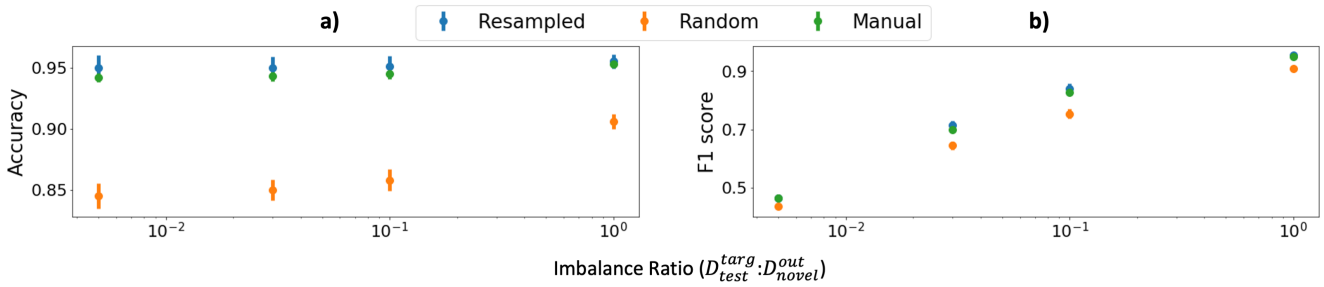

Figure 6. OOD testing results. Left-to-right: Accuracy and F1 score obtained from an average of model runs using D_{test}^{targ}:D_{novel}^{out} ratios of $\{1{:}\,400, 1{:}\,50, 1{:}\,10, 1{:}\,1\}$. Error bars reflect standard deviation.

5.2. OOD testing

Testing is done using k-fold cross-validation where the number of folds reflects the desired imbalance ratio. We consider D_{test}^{targ}:D_{novel}^{out} ratios of $\{1{:}\,400, 1{:}\,50, 1{:}\,10, 1{:}\,1\}$ where classifier performance is averaged over each fold (Fig. 6). The ratios considered here are designed to reflect sampling environments that a plankton ecologist may encounter during deployment. The 1:400 ratio represents the most extreme case of population flux, where the deployment environment consists overwhelmingly of novel classes. This is akin to taking a classifier trained on data from the North Atlantic and using it to detect the same four classes in the Tasman Sea. The 1:10 and 1:50 ratios represent a more amenable scenario, where the model is deployed in a similar environment where target class organisms are not as rare. The 1:1 ratio produces a balanced testing set, and assumes an even number of novel and target class examples in the deployment environment.

5.3. Dataset shift testing

When background class population statistics remain roughly constant throughout deployment, the training dataset generated by BR may produce a classifier that is biased against identifying the more common classes. This is because training is disproportionately focused on rare/abnormal examples under the BR procedure (Fig. 3). To assess model performance under a variety of deployment scenarios, we test and average the performance of each classifier over each day's worth of data in D_{shift}^{out} combined with randomly drawn target class examples (Table 1). This assessment measures the classifiers ability to classify under changes in prior distributions and dataset shift. No novel classes were used in this testing scenario.

5.4. Alternative target class testing

For the different sets of target classes, the relative performance among the classifiers was similar to the results shown for the target classes considered in Sec. 4.2.

OOD testing. BR brought the largest performance gains when the target and OOD classes were visually very similar. All OOD detection results are reported for the 1:1 testing ratio. For accuracy, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 5.1% and 0.2% respectively. For F1 score, the *Resampled* classifier on average outperformed *Random* and *Manual* classifiers by 4.7% and 0.0% respectively.

Method	Accuracy	F1 Score
Resampled	**88 ± 1.4**	**94.1 ± .23**
Random	85 ± 1.3	93.9 ± .12
Manual	79.7 ± 2.8	79.2 ± 1.5

Table 1. Dataset shift testing results (in % including ± Std Dev.).

Dataset shift testing. For accuracy, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 0.9% and 9.8% respectively. For F1 score, the *Resampled* classifier on average outperformed the *Random* and *Manual* classifiers by 0.1% and 7.1%.

6. Discussion

We have shown that when downsampling is required, OOD detection performance can be improved by selecting an optimal subset of background training images. In each testing scenario, BR slightly outperformed its nearest competitor. However, BR was the only downsampling method to perform well in both testing scenarios, whereas the performance of the other two downsampling methods varied significantly in each regime. This was observed for the alternative target classes as well. This finding underscores the efficacy of BR since an automated plankton classifier deployed on real-time data is almost guaranteed to experience both novel classes and dataset shift.

For some hard background classes, the distribution of image weights appears to have a seasonal dependence (Fig. 5). The image weights associated with the *detritus* class (Fig. 5b) are comparatively low, suggesting that only occasional instances of *detritus* will possess physical attributes

that pose a challenge to the classifier. This makes sense considering the large range of shapes and textures that objects described as "detritus" can have. This differs from classes like *Dictyocha*, *Skeletonema*, and *pennate* whose images are assigned comparatively larger weight values. However, these three classes show seasonal, and even daily, with-in class variability as indicated by the standard deviation of the image weights (Fig. 5d). This variability suggests that an optimal background training set for OOD detection has to be curated at the example level - randomly sampling from harder background classes may not produce adequate discrimination between the background and target classes. BR optimizes the downsampling strategy by deliberately selecting hard negatives that are very close to the target classes (Fig. 4).

The F1 score from each test set suggests that the 1:10 and 1:50 imbalance ratio produces greater performance differences (Fig. 6a). Despite the relatively higher accuracy obtained by the *Resampled* and *Manual* classifiers, the target classes become so polluted by false positives that absolute and relative performance – as measured by F1 score – degrades significantly with increasing imbalance ratio.

Overall, BR provides competitive or even slightly better OOD detection performance than the class-balanced downsampling used to train the *Manual* classifiers (Fig. 6). The improved performance can likely be attributed to the fact that class-balanced downsampling, while drawing from each class disproportionality, still performs random downsampling within each class. BR, in contrast, selects disproportionately from each class, while simultaneously select challenging examples from within each class. This may be particularly valuable for taxonomic groups where organisms can express different phenotypes as they go through different life stages.

While the *Manual* classifier yields satisfactory detection on novel classes (Fig. 6), this classifier significantly underperforms on natural population changes compared to the other classifiers (Table 1). This is likely because class-balanced downsampling produces a background class distribution that is significantly different from the background class distribution encountered during deployment. The classifiers trained on randomly drawn subsets perform comparatively well in the dataset shift scenario, likely because the background class distribution generally resembles the class priors encountered during testing. BR will typically draw relatively more examples from the minority classes and fewer examples from the most abundant classes as compared to the *Random* training set (Fig. 3). But BR's disproportionate subsampling is not as extreme as the class-balanced downsampling. The fact that BR preserves a significant amount of information regarding the class priors is perhaps why it performs better than class-balanced downsampling for natural population distributions

7. Comments and recommendations

In order to adequately simulate the deployment of a classifier, our testing procedure involved the use of data from training and novel classes as well as shifting prior distributions. We believe this to be the most rigorous form of testing and hope that this study can serve as a framework for future plankton classifier benchmarking. The high visual similarity between the target and OOD examples makes this a challenging detection problem. Most of the methods introduced in the OOD literature, including BR, test using OOD examples that come from an entirely separate domain. This study is one of the first studies to benchmark the performance of a contemporary OOD detection method on in-domain OOD data. It is our hope that these results will facilitate the development of new tools for HAB species monitoring and early detection systems. The reduced false positive rates demonstrated in our experiments make the output of the algorithm more amenable to quality control for verification.

The BR procedure can be used to improve classification systems that incorporate an ensemble of "one-versus-all", which are popular within the plankton ecology community and have been used for HAB species monitoring [8, 43]. This could be done by training each one-versus-all classifier using a subset of background images that is optimized to produce the best discrimination for the class that each classifier is trying to detect.

We note that the utility of the background weight learning mechanism is not limited to HNM approaches. It can in and of itself be used to communicate potential failure modes to a human supervisor. For example, by examining the background images that were assigned large weight values, a human user could learn prior to deployment which non-target classes are likely to produce false positives. With this knowledge, they may decide to train the model to detect these tough classes as well.

We have shown that obtaining class labels for background objects for the purpose of class-balanced downsampling does not improve OOD detection performance. Therefore, we conclude that for any future plankton classification campaigns similar to this experimental setup, all human annotation efforts should be focused on the target classes. Instead of random or class-balanced downsampling, automatic procedures such as BR should be used to optimally resample the 'other' category.

Acknowledgements This work was supported by the U.S. Office of Naval Research under Award No. N00014-19-1-2851. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. We would also like to thank Jules S. Jaffe and Nuno Vasconcelos for critical readings of the manuscript and their advisory roles. We thank Yi Li for advising the use of the Background Resampling method.

References

[1] Kevin R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355, Sept. 2005. Number: 7057 Publisher: Nature Publishing Group.

[2] Oscar Beijbom, Judy Hoffman, Evan Yao, Trevor Darrell, Alberto Rodriguez-Ramirez, Manuel Gonzalez-Rivero, and Ove Hoegh Guldberg. Quantification in-the-wild: data-sets and baselines. *arXiv:1510.04811 [cs]*, Nov. 2015. arXiv: 1510.04811.

[3] Simon T. Belt, Guy Allard, Guillaume Massé, Steve Rowland, and Jean-Michel Robert. Important sedimentary sesterterpenoids from the diatom Pleurosigma intermedium. *Chemical Communications*, (6):501–502, 2000.

[4] Mark Benfield, Philippe Grosjean, Phil Culverhouse, Xabier Irigolen, Michael Sieracki, Angel Lopez-Urrutia, Hans Dam, Qiao Hu, Cabell Davis, Allen Hanson, Cynthia Pilskaln, Edward Riseman, Howard Schulz, Paul Utgoff, and Gabriel Gorsky. RAPID: Research on Automated Plankton Identification. *Oceanography*, 20(2):172–187, June 2007.

[5] Tristan Biard, Lars Stemmann, Marc Picheral, Nicolas Mayot, Pieter Vandromme, Helena Hauss, Gabriel Gorsky, Lionel Guidi, Rainer Kiko, and Fabrice Not. In situ imaging reveals the biomass of giant protists in the global ocean. *Nature*, 532(7600):504–507, Apr. 2016. Number: 7600 Publisher: Nature Publishing Group.

[6] Erik Bochinski, Ghassen Bacha, Volker Eiselein, Tim J. W. Walles, Jens C. Nejstgaard, and Thomas Sikora. Deep Active Learning for In Situ Plankton Classification. In Zhaoxiang Zhang, David Suter, Yingli Tian, Alexandra Branzan Albu, Nicolas Sidère, and Hugo Jair Escalante, editors, *Pattern Recognition and Information Forensics*, Lecture Notes in Computer Science, pages 5–15, Cham, 2019. Springer International Publishing.

[7] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, Oct. 2018.

[8] Lisa Campbell, Darren W. Henrichs, Robert J. Olson, and Heidi M. Sosik. Continuous automated imaging-in-flow cytometry for detection and early warning of Karenia brevis blooms in the Gulf of Mexico. *Environmental Science and Pollution Research*, 20(10):6896–6902, Oct. 2013.

[9] R W Campbell, P L Roberts, and J Jaffe. The Prince William Sound Plankton Camera: a profiling in situ observatory of plankton and particulates. *ICES Journal of Marine Science*, 77(4):1440–1455, July 2020.

[10] Svenja Christiansen, Henk-Jan Hoving, Florian Schütte, Helena Hauss, Johannes Karstensen, Arne Körtzinger, Simon-Martin Schröder, Lars Stemmann, Bernd Christiansen, Marc Picheral, Peter Brandt, Bruce Robison, Reinhard Koch, and Rainer Kiko. Particulate matter flux interception in oceanic mesoscale eddies by the polychaete Poeobius sp. *Limnology and Oceanography*, 63(5):2093–2109, 2018.

[11] Robert K. Cowen and Cedric M. Guigand. In situ ichthyoplankton imaging system (ISIIS): system design and preliminary results. *Limnology and Oceanography: Methods*, 6(2):126–132, 2008.

[12] Jialun Dai, Zhibin Yu, Haiyong Zheng, Bing Zheng, and Nan Wang. A Hybrid Convolutional Neural Network for Plankton Classification. In Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision – ACCV 2016 Workshops*, Lecture Notes in Computer Science, pages 102–114, Cham, 2017. Springer International Publishing.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. ISSN: 1063-6919.

[14] Cabell S. Davis, Qiao Hu, Scott M. Gallager, Xiaoou Tang, and Carin J. Ashjian. Real-time observation of taxa-specific plankton distributions: an optical sampling method. *Marine Ecology Progress Series*, 284:77–96, Dec. 2004.

[15] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing Network Agnostophobia. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9157–9168. Curran Associates, Inc., 2018.

[16] Jeffrey S. Ellen, Casey A. Graff, and Mark D. Ohman. Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods*, page lom3.10324, July 2019.

[17] Robin Faillettaz, Marc Picheral, Jessica Y. Luo, Cédric Guigand, Robert K. Cowen, and Jean-Olivier Irisson. Imperfect automatic image classification successfully describes plankton distribution patterns. *Methods in Oceanography*, 15-16:60–77, Apr. 2016.

[18] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, Oct. 2008.

[19] Pablo González, Eva Álvarez, Jorge Díez, Ángel López-Urrutia, and Juan José del Coz. Validation methods for plankton image classification systems. *Limnology and Oceanography: Methods*, 15(3):221–237, 2017.

[20] L. R. Haury, J. A. McGowan, and P. H. Wiebe. Patterns and Processes in the Time-Space Scales of Plankton Distributions. In John H. Steele, editor, *Spatial Pattern in Plankton Communities*, NATO Conference Series, pages 277–327. Springer US, Boston, MA, 1978.

[21] Graeme C. Hays, Anthony J. Richardson, and Carol Robinson. Climate change and marine plankton. *Trends in Ecology & Evolution*, 20(6):337–344, June 2005.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[23] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. *arXiv:1812.04606 [cs, stat]*, Jan. 2019. arXiv: 1812.04606.

[24] Qiao Hu and Cabell S. Davis. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. 2006.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. Conference Name: Proceedings of the IEEE.

[27] Hansang Lee, Minseok Park, and Junmo Kim. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3713–3717, Sept. 2016. ISSN: 2381-8549.

[28] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. *arXiv:1711.09325 [cs, stat]*, Feb. 2018. arXiv: 1711.09325.

[29] Yi Li and Nuno Vasconcelos. Background Data Resampling for Outlier-Aware Classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13215–13224, Seattle, WA, USA, June 2020. IEEE.

[30] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *arXiv:1706.02690 [cs, stat]*, Feb. 2018. arXiv: 1706.02690.

[31] Charles X Ling and Chenghui Li. Data Mining for Direct Marketing: Problems and Solutions. *KDD'98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, page 7, 1998.

[32] George B. McManus and Laura A. Katz. Plankton Identification: Morphology or Molecules or Both? *Limnology and Oceanography Bulletin*, 18(4):86–90, 2009.

[33] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, Jan. 2014.

[34] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, Jan. 2012.

[35] Aleksander Borge Nesse. Classifying Dinoflagellates in Palynological Slides Using Convolutional Neural Networks. 2020. Accepted: 2020-09-28T18:52:08Z Publisher: University of Stavanger, Norway.

[36] Robert J. Olson and Heidi M. Sosik. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods*, 5(6):195–203, 2007.

[37] Eric Coughlin Orenstein. *Automated analysis of oceanographic image data*. PhD thesis, UC San Diego, 2018.

[38] Eric C. Orenstein and Oscar Beijbom. Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1082–1088, Mar. 2017. ISSN: null.

[39] Eric C. Orenstein, Oscar Beijbom, Emily E. Peacock, and Heidi M. Sosik. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. *arXiv:1510.00745 [cs]*, Oct. 2015. arXiv: 1510.00745.

[40] Eric C. Orenstein, Kasia M. Kenitz, Paul L. D. Roberts, Peter J. S. Franks, Jules S. Jaffe, and Andrew D. Barton. Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. *Limnology and Oceanography: Methods*, 18(12):739–753, 2020.

[41] Eric C. Orenstein, Devin Ratelle, Christian Briseño-Avena, Melissa L. Carter, Peter J. S. Franks, Jules S. Jaffe, and Paul L. D. Roberts. The Scripps Plankton Camera system: A framework and platform for in situ microscopy. *Limnology and Oceanography: Methods*, 18(11):681–695, 2020.

[42] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier Exposure with Confidence Control for Out-of-Distribution Detection. *arXiv:1906.03509 [cs, stat]*, June 2020. arXiv: 1906.03509.

[43] Vito P. Pastore, Thomas G. Zimmerman, Sujoy K. Biswas, and Simone Bianco. Annotation-free learning of plankton for classification and anomaly detection. *Scientific Reports*, 10(1):12142, Dec. 2020.

[44] Emily E. Peacock, Robert J. Olson, and Heidi M. Sosik. Parasitic infection of the diatom Guinardia delicatula, a recurrent and ecologically important phenomenon on the New England Shelf. *Marine Ecology Progress Series*, 503:1–10, Apr. 2014.

[45] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Human Face Detection in Visual Scenes. Technical report, Carnegie Mellon University, 1995.

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.

[47] A.H. Schistad Solberg and R. Solberg. A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images. In *IGARSS '96. 1996 International Geoscience and Remote Sensing Symposium*, volume 3, pages 1484–1486 vol.3, May 1996.

[48] Jan Schulz, Kristina Barz, Patricia Ayon, Andree Ludtke, Oliver Zielinski, Dirk Mengedoht, and Hans-Jurgen Hirche. Imaging of plankton specimens with the lightframe on-sight keyspecies investigation (LOKI) system. *Journal of the European Optical Society: Rapid Publications*, 5:10017s, Apr. 2010.

[49] Jihoon Shin, Seonghyeon Yoon, YoungWoo Kim, Taeho Kim, ByeongGeon Go, and YoonKyung Cha. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics*, 61:101202, Mar. 2021.

[50] Akila Somasundaram and U Srinivasulu Reddy. Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data. (978):8, 2016.

[51] Akila Somasundaram and U. Srinivasulu Reddy. *Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data*. Jan. 2016.

[52] Heidi M. Sosik, Emily E. Peacock, and Emily F. Brownlee. WHOI Plankton: Annotated Plankton Images - Data Set for Developing and Evaluating Classification Methods. Tech-

nical report, Woods Hole Oceanographic Institution, 2014. Type: dataset.

[53] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, Jan. 1998. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[54] Dirk Tasche. Fisher consistency for prior probability shift. *The Journal of Machine Learning Research*, 18(1):3338–3369, Jan. 2017.

[55] C. Wang, X. Zheng, C. Guo, Z. Yu, J. Yu, H. Zheng, and B. Zheng. Transferred Parallel Convolutional Neural Network for Large Imbalanced Plankton Database Classification. In *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pages 1–5, May 2018.

[56] Lishu Wang, Bin Yang, Xiu-Ping Lin, Xue-Feng Zhou, and Yonghong Liu. Sesterterpenoids. *Natural Product Reports*, 30(3):455, 2013.