ICES Journal of Marine Science, 2026, Vol. 83, Issue 1, fsaf212
https://doi.org/10.1093/icesjms/fsaf212
Received: 24 July 2025; **revised:** 7 November 2025; **accepted:** 10 November 2025
Original Article

Practical guidance on automated sorting of underwater images in plankton ecology research

Kevin Sorochan [1,*], Ankita Ravi Vaswani [2], Ann Howard [1], Saskia Rühl [3], Emily O'Grady [1], Klas Ove Möller [2], Catherine L. Johnson [1]

[1]Fisheries and Oceans Canada, Bedford Institute of Oceanography, Dartmouth, NS B2Y 4A2, Canada
[2]Institute of Carbon Cycles, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany
[3]Plymouth Marine Laboratory, Prospect Place, Devon PL1 3DH Plymouth, United Kingdom

*Corresponding author. Fisheries and Oceans Canada, Bedford Institute of Oceanography, P.O. Box 1006, Dartmouth, NS, Canada B2Y 4A2. E-mail: Kevin.Sorochan@dfo-mpo.gc.ca

Abstract

In situ plankton imaging complements classical sampling approaches by obtaining observations of plankton composition and traits at finer spatial and temporal resolutions. These imaging techniques can provide valuable ecological insight and are also notorious for generating a massive volume of images that require classification to generate quantitative data. Automating image segmentation and classification can accelerate data extraction; however, the high diversity and uneven distribution of plankton taxa, variation in image characteristics obtained from different imagers, and limited availability of human classification expertise present challenges to development of user-friendly and universally accepted image processing and classification tools. Differences in desired taxonomic or trait resolution, classifier performance, and spatial variability in community composition often necessitate the development of tailored automated classifiers for specific cases. This customization typically requires expertise in computer vision and machine learning that many ecologists do not acquire through traditional training. In this paper, we review the plankton imaging and classification workflow and present two case studies for a plankton ecology audience. We emphasize Convolutional Neural Network (CNN) classifiers and demonstrate strategies to address common challenges in image classification using semiautomated classification and unsupervised learning approaches. The overarching aim is to provide practical guidance for ecologists and encourage broader adoption of *in situ* plankton imaging in ecological research.

Keywords: image classification; plankton; computer vision; deep learning

Introduction

In situ plankton imaging and traditional sampling methods, such as nets or water samples, have different advantages and limitations that complement one another in terms of spatiotemporal coverage and characterization of plankton community composition. Imagers can sample at a higher spatiotemporal resolution, provide observations of particle orientation and morphology, and capture particles that are missed by nets or destroyed during sample collection or preservation (Culverhouse et al. 2006, Ohman 2019, Orenstein et al. 2022). On the other hand, imagers tend to sample a much smaller volume and typically produce two-dimensional images that may limit taxonomic resolution and estimation of plankton biovolume (e.g. Culverhouse et al. 2006). The advantages of plankton imaging have long been acknowledged (Davis et al. 1996); however, widespread use of this sampling method was hindered for decades by limited availability of image acquisition systems and accessibility to computer vision solutions that automate image sorting (Benfield et al. 2007). Commercialization of imagers and development of publicly available, powerful, and relatively user friendly image classification software have facilitated greater use of plankton imaging for research and monitoring purposes (Irisson et al. 2022). In turn, plankton imaging dataset availability and publications have increased substantially over the last two decades (Irisson et al. 2022, Ciranni et al. 2024).

In situ plankton imaging presents the challenge of obtaining ecologically relevant data products (e.g. taxa- or trait-specific counts) from large quantities of raw image data (Benfield et al. 2007). Automated methods for extracting Regions of Interest (ROIs, i.e. cropped areas, each containing at least one imaged object) from full-frames and subsequent image classification reduce human effort in the process (Irisson et al. 2022). By automating ROI acquisition and sorting, biases in image sorting (Culverhouse 2007) can also be reduced. Supervised machine learning, including feature engineering and deep learning approaches, are commonly used to speed up image classification by building a model from a training set to predict the class of unlabeled images (González et al. 2017).

The feature engineering approach is dependent on a human selecting image features that are appropriate to the classification task (Gorsky et al. 2010). This approach can be efficient and elegant but requires expertise in feature selection (Chollet et al. 2022). In contrast, the deep learning approach utilizes Convolutional Neural Networks (CNNs) to automatically learn important image features and make predictions based on them (LeCun et al. 2015). Advances in performance of deep CNNs were made in the 2010s, which facilitated their

broader adoption in computer vision (LeCun et al. 2015) and classification of plankton images (reviewed by Irisson et al. 2022, Rubbens et al. 2023, Ciranni et al. 2024). The benefit of automatically obtaining learned features comes with a cost of increased human and computational effort because deep CNNs require large quantities of labeled images when trained from scratch (e.g. Luo et al. 2018, Ellen and Ohman 2024). This problem can be addressed with use of "transfer learning", in which a CNN trained on a different set of images is adapted for a new classification task. Transfer learning can, therefore, leverage the power of CNNs for applications with limited training data (Rubbens et al. 2023, Ciranni et al. 2024) and is commonly used in marine science and plankton research (e.g. Campbell et al. 2020, Le et al. 2022). In addition to transfer learning, the significant human labeling effort associated with building training sets can be reduced by using clustering methods to identify groups of images with similar features (Pastore et al. 2020, Schröder et al. 2020) or semisupervised learning approaches that utilize unlabeled and labeled images during training (e.g. Schanz et al. 2023).

As pointed out by Irisson et al. (2022), ecologists typically do not have formal training in image processing and machine learning techniques, and this may limit implementation of automated image classification methods in the field of plankton ecology. Many publications that address classification of plankton images are either broad in their scope with less emphasis on practical guidance for image classification methodology (e.g. Lombard et al. 2019, Goodwin et al. 2022, Irisson et al. 2022, Malde et al. 2020, Rubbens et al. 2023) or include a high degree of technical detail and are published in computer vision or machine learning journals (e.g. Ciranni et al. 2024, Eerola et al. 2024). The goal of this paper is to provide an overview of the general image classification framework and examples of its implementation to a plankton ecology audience.

Automated classification of plankton images is an active field of research with a diverse suite of methods that vary in their level of sophistication (e.g. Ciranni et al. 2024). While this precludes a detailed set of guidelines for best practice, there are overarching concepts and common challenges relevant to classification of plankton images regardless of the exact methodology (Eerola et al. 2024). In this paper, these concepts and challenges are presented in the context of the generalized plankton image analysis workflow with a focus on application of CNNs for image classification. The practical value and flexibility of this framework is also demonstrated in two case studies, highlighting the potential of plankton imaging methods to accelerate discovery and deepen ecological insight in diverse research contexts.

Overview of the supervised image classification workflow

Generating data from images of plankton via machine classification (e.g. "the plankton quantitative imaging process," Irisson et al. 2022) involves several steps, which are described in the following subsections: (1) image acquisition, segmentation, and preprocessing; (2) image annotation; (3) automated classification; (4) performance evaluation; and (5) (if necessary) improving classification performance to achieve research objectives (Fig. 1).

In practice, this process is typically iterative rather than strictly sequential (Fig. 1). For example, classifier training in the machine learning step may reveal issues with image quality or consistency of human labeling, necessitating adjustments to image preprocessing or annotation. Evaluation of classifier performance often identifies misclassifications or low-confidence predictions, requiring additional refinements to the annotated image set or model parameters. This cycle of refinement continues until performance reaches an acceptable level for the research objective, ensuring accurate and reliable quantitative data products.

There are alternative approaches to image classification that do not fit within the framework described herein and are beyond the scope of this paper. For example, You Only Look Once (YOLO, Redmon et al. 2016) and Single Shot Multibox Detector (SSD, Liu et al. 2016) carry out simultaneous object detection and classification from input images that can contain multiple objects (see Ciranni et al. 2024). In addition, "zero-shot learning" approaches utilize semantic information, such as human-defined attributes (e.g. Li et al. 2023) or natural language descriptions (e.g. Huo et al. 2025) to predict classes that are absent in the training set.

Acquisition, segmentation, and preprocessing

Plankton imagers vary in their optical method (e.g. dark or bright field illumination, shadowgraph), sampling rate, sampling procedure (pumped or open path imaging), and deployment mode (fixed, profiling, towed) (Lombard et al. 2019). The optical method and camera specifications determine fundamental characteristics of the images, including the field of view, resolution, pixel pitch, contrast, and color (Lombard et al. 2019, Eerola et al. 2024, Ellen and Ohman 2024). Different instruments can, therefore, produce images of the same individual plankter but with different features, which poses a challenge for development of classifiers that can effectively utilize plankton images acquired from different instruments (Ciranni et al. 2024, Eerola et al. 2024).

After image acquisition, ROIs are extracted from the larger image frame containing the camera's full field of view. Image segmentation is typically automated (e.g. Bi et al. 2015, Panaïotis et al. 2022), although manual segmentation combined with annotation is also used in some cases (e.g. Richards et al. 2019). The choice of segmentation method and parameterization directly influence the number and quality of ROIs obtained, affecting attributes such as intensity and sharpness. Ideally, segmentation should be optimized to capture ROIs with desirable attributes while minimizing the inclusion of non-target regions, thereby minimizing images belonging to nuisance classes, such as blurry images or artifacts. This can be achieved by refining the segmentation algorithm (Luo et al. 2018, Panaïotis et al. 2022). Images belonging to nuisance classes increase data storage requirements, computational load during training, annotation effort, and class imbalance, potentially degrading classification performance for target classes.

It may also be necessary to preprocess full frame images to minimize the number of images in nuisance classes (Panaïotis et al. 2022) or to preprocess ROIs to improve automated classification performance (e.g. noise reduction, Bi et al. 2015). When using CNN classifiers, ROIs are usually preprocessed prior to automated classification to ensure compatibility with the classifier or preserve aspect ratio (Campbell et al. 2020, Li and Du 2022, Ellen and Ohman 2024). CNN architectures require image inputs with specific dimensions; therefore, images

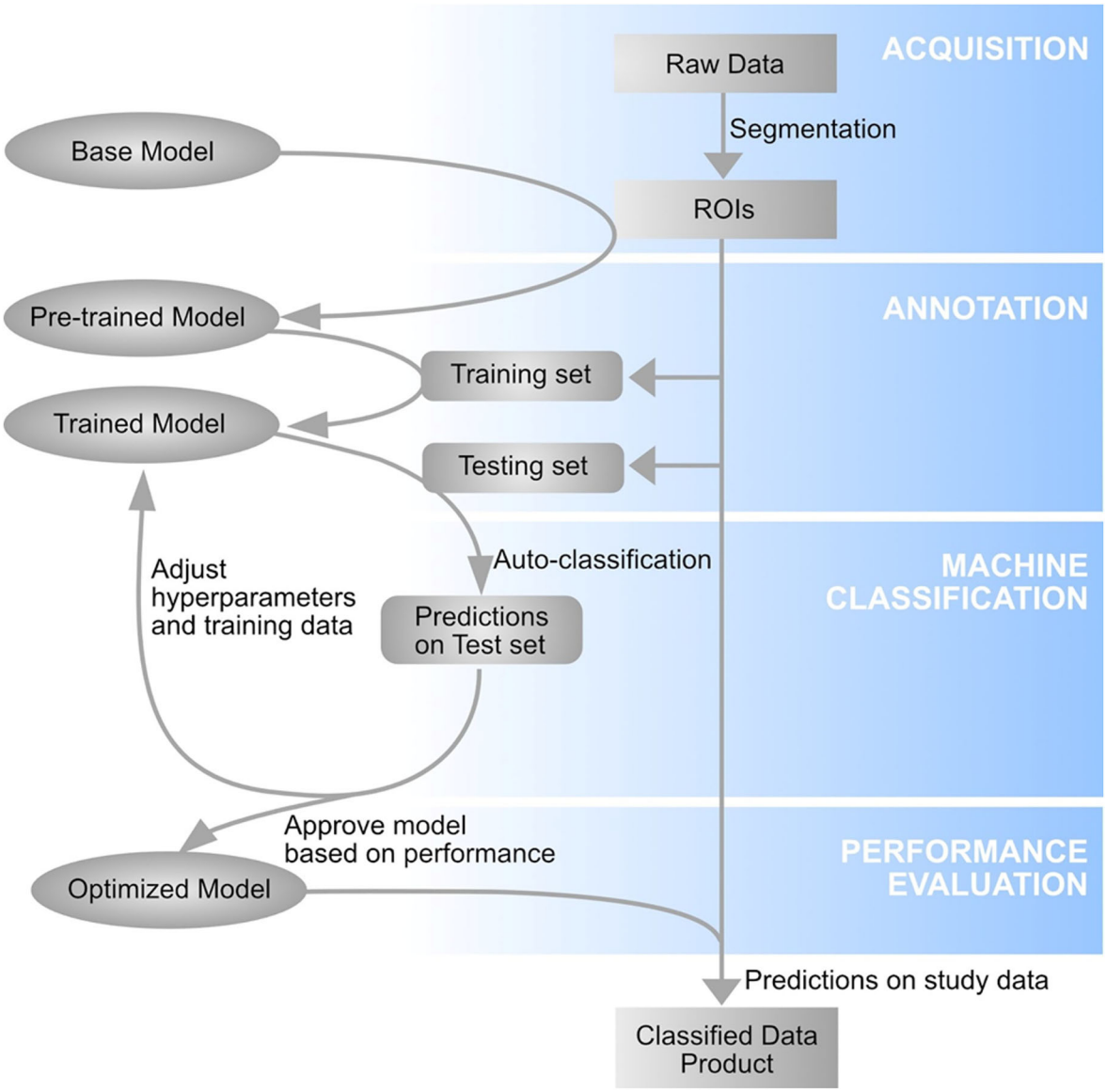

Figure 1. Schematic of supervised learning process for classification of plankton images.

are resized before they are introduced to the classifier, potentially resulting in information loss and distortion. To preserve aspect ratio, the image objects can be made square prior to resizing by "padding" with additional pixels consistent with the images' background (e.g. Plonus et al. 2021, Ellen and Ohman 2024). Image padding is not required and may not always improve classification performance (Lumini and Nanni 2019, Kerr et al. 2020, Schanz et al. 2023).

Annotation

Supervised machine learning requires a set of images that has been annotated with class labels. The labeled images are used to train the classifier and evaluate its performance. Images may be labeled based on taxonomy or traits. Trait groupings may be used to address trait-based research questions or improve classification performance by reducing intra-class variation (e.g. "Appendicularia" and "Appedicularia_with_house" classes, Plonus et al. 2021). Obtaining a set of images that has accurate labels and is representative of the population of unlabeled images is crucial for accurate automated classification (e.g. González et al. 2017).

A large number of images may need to be labeled to train the classifier (e.g. Luo et al. 2018, Ellen and Ohman 2024). At the same time, minimizing error in manual labeling requires human expertise in plankton taxonomy and consensus among experts (Culverhouse 2007). These logistical constraints limit the size of the annotated image set, posing a major challenge in automated image classification in marine ecology (e.g. Richards et al. 2019, Kenitz et al. 2023, Eerola et al. 2024, Ciranni et al. 2024). Freely available tools facilitate human interaction with the images by providing a platform for browsing and labeling images, collaboration among multiple classification experts, and export for classifier training (e.g. Langenkämper et al. 2017, Picheral et al. 2017).

Ideally, the training set should be constructed in an unbiased way so that the frequency of images among classes and the features associated with each class are consistent between labeled training and unlabeled image domains (González et al. 2017, Orenstein et al. 2020, Eerola et al. 2024, Ellen and

Ohman 2024). This can be accomplished by randomly sampling and labeling ROIs from the overall dataset. The resulting frequency of images among classes is often unbalanced, and images belonging to rare classes may be missed entirely. This can be problematic because features from dominant classes contribute more to the optimization procedure in the process of training the classifier. Consequently, classifier performance is typically lower for classes with less representation in the training set (e.g. Schanz et al. 2023).

More images belonging to non-dominant classes may need to be labeled to deal with class imbalance. Since it is difficult to visually locate instances of rare classes in a large set of unlabeled ROIs, methods have been developed that can help to "rebalance" an annotated image set. For example, anomaly detectors can be used to identify images belonging to rare-classes (e.g. Pastore et al. 2020), and a larger training set can be built out from an initially limited set using an iterative "active learning" method (Bochinski et al. 2018). In this method, (1) a classifier is trained and used to obtain predictions; (2) images are added to the training set only if they were auto-classified with low confidence; and (3) the updated training set is then used to re-train the classifier (Bochinski et al. 2018). Both anomaly detection and active learning are featured in Case Study 2.

In standard CNN algorithms, the classifier outputs a probability score for each class for each image input (e.g. LeCun et al. 2015, Faillettaz et al. 2016). The probability scores can be used to evaluate classifier confidence, and the class with the highest probability score is typically assigned as the categorical prediction. Rebalancing the annotated image set comes at the cost of potentially introducing a domain shift in class frequency between labeled training and unlabeled target images, resulting in classifier bias (e.g. González et al. 2017, Orenstein et al. 2020). Schanz et al. (2023) adjusted probability score outputs to address bias introduced from rebalancing of the training set using a Bayesian correction.

Automated classification

During training, CNN weights (i.e. adjustable parameters) are "learned" such that classification performance is optimized by minimizing loss, a measure of prediction error as defined by the loss function (also referred to as an objective function, Lecun et al. 2015). In supervised learning for image classification, loss is derived from the model predictions and labels in annotated image sets reserved for training and validation. Both loss and accuracy are typically computed and reported during training iterations. High training accuracy and low validation accuracy indicate model overfitting. Several techniques can be used to avoid overfitting and improve generalization, including the use of image augmentation and implementation of regularization techniques such as weight regularization and dropout (Chollet et al. 2022).

The typical supervised machine learning process includes designing or selecting and modifying an existing CNN architecture, setting hyperparameters (i.e. model parameters that are set initially and not updated during training), and training. The process is iterative and may involve updates to the annotated dataset, CNN architecture, and tuning of hyperparameters (Wojciuk et al. 2024) to increase model performance and efficiency (Fig. 1). A key hyperparameter is the learning rate, which influences the extent to which weights are adjusted in each iteration of the optimization process. Increasing the learning rate can decrease the time to convergence, but this can also introduce instability and prevent convergence altogether (Bartz et al. 2023, Ellen and Ohman 2024).

CNN architectures and associated weights obtained from pretraining on large training sets including ImageNet (Deng et al. 2009) are freely available (e.g. Lumini and Nanni 2019, Eerola et al. 2024). Image features relevant to differences among specific plankton classes can be subsequently obtained from further training with plankton images. The use of established CNN architectures and transfer learning is a practical and powerful way to address problems associated with a limited number of annotated images and limited expertise in CNN architecture design. Strengths of different CNN architectures can be combined in an analysis using an ensemble of CNN architectures (Ellen et al. 2019).

Pretraining with unsupervised, contrastive learning has been demonstrated to improve classification performance when labeled images are scarce (Chen et al. 2020a, Chen et al. 2020b, Schanz et al. 2023). In contrastive learning, image augmentations (i.e. transformations such as cropping, flipping, color jitter, and rescaling) are introduced, one for each image (Chen et al. 2020b). In this method, loss is minimized when similarity between the original image and its augmented pair is higher than between the original image and an unrelated pairing. The basic premise of contrastive learning is that the network's output should differ for different images but should remain the same between augmented variations of an image. Hence, the network learns features of an image which are invariant under the introduced augmentations, and the choice of augmentations thus impacts which features are learned as "essential" to an image. Ultimately, the CNN is initialized with weights obtained from contrastive learning and then further trained with labeled plankton images for classification. The use of these unsupervised and supervised methods together is referred to as semisupervised machine learning and is utilized in Case Study 2.

Performance evaluation

The performance of a classifier is best evaluated by examining its predictions on labeled images in a test set that the classifier has not been exposed to during training (Fig. 1). Key metrics that are used to evaluate the performance of an automated classifier include recall, precision, F1 score, and accuracy (Goodwin et al. 2022, Orenstein et al. 2022, Oldenburg et al. 2023, Ciranni et al. 2024). These metrics, as well as the overall performance of a classifier, can be visualized by plotting a confusion matrix (Fig. 2). The confusion matrix facilitates visualization of the complete record of model performance by mapping the true positives (TP), false positives (FP), and false negatives (FN) for each class.

Recall and precision are calculated from TP and FP or FN as follows:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

Recall measures the ability of a classifier to predict the class that matches the true label; high recall is a result of low FN (Eq. 1). On the other hand, precision measures the ability of the classifier to exclude instances in which the predicted class does not match the true label; high precision is a result of low FP (Eq. 2). Ideally, a model would be characterized by high

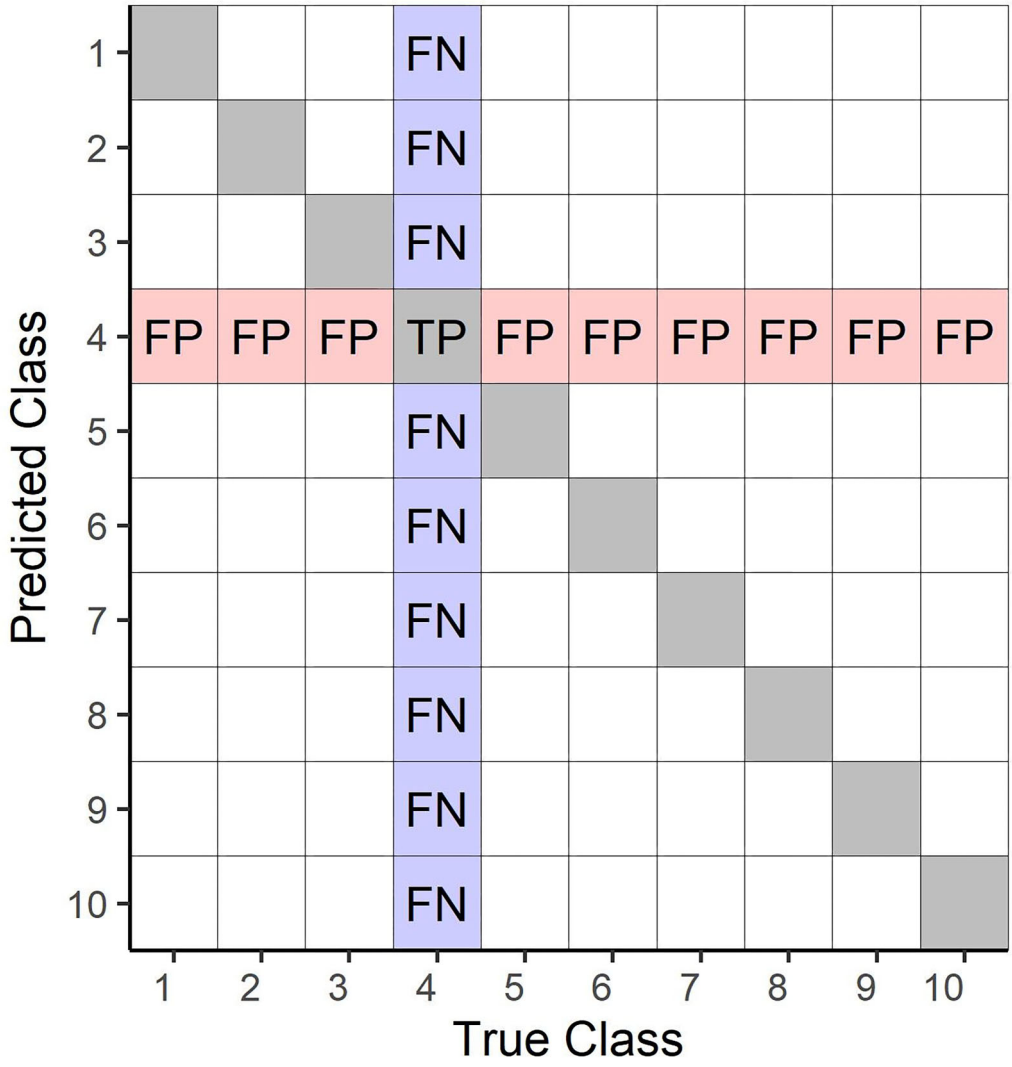

Figure 2. Confusion matrix indicating position of TP, FP, and FN regarding class 4; all unlabeled cells are true negatives (TN) with respect to class 4. The matrix diagonal, which indicates TP for each class, is shaded grey.

recall and precision; however, reducing FN often comes at the cost of increasing FP (e.g. Goodwin et al. 2022). When evaluating classification performance, the relative importance of recall and precision may vary according to the research objectives. Precision may be emphasized in species distribution modeling or habitat suitability studies because FPs result in incorrect characterization of class presence. Recall may be emphasized in studies where it is important that the detection of a class is not missed, such as the detection of invasive species or rare but important ecological indicators; however, human effort would likely be required to confirm TPs. Studies that aim to accurately quantify plankton concentration would require both high precision and recall; however, recall may be prioritized for dominant classes if contamination by less abundant classes occurs within an acceptable level.

The F1 score is a common metric of classifier performance, defined as the harmonic mean of recall and precision:

$$F1 = \left(2 * \text{Precision} * \text{Recall}\right) / \left(\text{Precision} + \text{Recall}\right) \quad (3)$$

Another common metric is accuracy. Overall accuracy is the ratio of TP to total number of observations, N:

$$\text{Accuracy} = \sum \text{TP}/\text{N} \quad (4)$$

In binary classification (including "one versus all," class-specific scenarios), accuracy is computed as follows:

$$\text{Accuracy} = \left(\text{TP} + \text{TN}\right)/\text{N} \quad (5)$$

In both cases, accuracy represents the sum of the diagonal of the confusion matrix divided by the number of observations. Overall accuracy (Eq. 4) can be misleading if there is strong imbalance in the number of image objects among categories, because this metric will not reflect misclassification of rare categories. Plotting confusion matrices and evaluating class-specific recall, precision, F1, and accuracy (Eq. 5) provides a comprehensive understanding of classifier efficacy.

Improving classification performance

If the classifier performance is not satisfactory for the research objective, additional measures are required to reduce the number of classification errors. Automated classification could be

used solely to increase the efficiency of manual validation of the entire unlabeled image set (e.g. Gorsky et al. 2010) at the potential cost of substantial human effort. Alternatively, predictions could be visually verified and FP instances removed from the analysis (Bi et al. 2015) or instances with probability score outputs below a threshold could be removed from the analysis (Faillettaz et al. 2016, Luo et al. 2018, Campbell et al. 2020). Both of these methods aim to improve precision at the cost of losing information. Excluding images with relatively low probability scores reduces recall and may result in exclusion of rare classes (Plonus et al. 2021, Oldenburg et al. 2023). If rare classes are of particular interest, it may be advantageous to accept and manually verify instances where the class of interest appears within the top-k predicted classes, rather than relying solely on the top-1 prediction (i.e. $k = 1$, Plonus et al. 2021). Additionally, classification probability scores can be used to prioritize manual verification efforts, focusing verification on images associated with lower confidence predictions (see Case Study 1). While this approach can enhance both recall and precision, it may also lead to a significant increase in manual verification workload.

The use of probability score predictions for decision making in image classification pipelines (Luo et al. 2018, Plonus et al. 2021, Bochinski et al. 2018, Conradt et al. 2022, Oldenburg et al. 2023) implies that probability scores are effective indicators of classifier confidence. A classifier is "well-calibrated" if it correctly quantifies the confidence associated with its predictions (Silva Filho et al. 2023). For example, in binary classification, an image assigned a probability score of 0.7 for the class "diatom" and 0.3 for the class "not diatom" should actually be a diatom 70% of the time. Calibration can be checked by comparing predicted and empirical probabilities in a sample of images (Silva Filho et al. 2023), such as a test set. Since classifier calibration influences the frequency distribution of predicted probability scores, it may be useful to examine calibration when applying probability score thresholds (as described above). Note that classifiers that exhibit high classification accuracy are not necessarily well-calibrated (Guo et al. 2017).

Case studies

In this section, two case studies highlight different plankton imaging applications and associated image classification methods in the context of the framework described above. While the above framework provides an overview of the general image acquisition and automated classification procedure, these case studies provide examples of decisions that are made in practice based on the acquisition method, research question, and available resources.

Case study 1: classification of mesozooplankton images to characterize the prey field of the North Atlantic right whale, *Eubalaena glacialis*

Background

The objective of this case study was to demonstrate a semiautomated image classification scheme for sorting *in situ* images of mesozooplankton with the research goal of characterizing the prey field of the endangered North Atlantic right whale, *E. glacialis*. Specifically, the primary research objective was to quantify the concentration and depth of aggregations of late-stage *Calanus* spp. copepods (e.g. Sorochan et al. 2023, Johnson et al. 2024). *Calanus* spp. are abundant in the western North Atlantic (Pepin et al. 2015), have a relatively large body size (several millimeters in length) and high capacity for lipid storage (Lee et al. 2006), and are an important food source for higher trophic levels including *E. glacialis* (e.g. Pershing and Stamieszkin, 2020). A secondary objective was to examine potential for other abundant zooplankton taxa to supplement the diet of *E. glacialis*. It was, therefore, necessary to sort images of mesozooplankton into multiple classes.

Acquisition

In situ images were captured with a SeaScan Inc Digital Auto Video Plankton Recorder (DAVPR) from 27 deployments in the southern Gulf of St. Lawrence, Canada, in June 2022 (Johnson et al. 2024). The DAVPR captured full frame images at a frequency of 20 Hz and logged images and associated Conductivity-Temperature-Depth data internally. The DAVPR camera (SVS-Vistek Charge-Coupled Device camera, model ECO285CVGE) provided color images with an optical field of view of 41.1 × 30.7 mm ("S3" optical setting) and resolution of 1392 × 1040 pixels. Illumination was provided by a strobe positioned opposite the camera. In each deployment, the DAVPR was repeatedly lowered and raised to within ∼5 m of the seafloor and 3 m of the surface at a wire speed of 0.5 m s^{-1} while the vessel was stationary.

ROIs were extracted from the full frames with image segmentation software developed for the DAVPR data processing stream (AutoDeck, SeaScan Inc). ROIs were identified and extracted using thresholds for intensity, focus, and SD of high frequency content. The values of these thresholds were selected to optimize the number of manually identifiable ROIs of large copepods (i.e. clearly defined individuals with high pixel intensity) while reducing the number of manually unidentifiable out-of-focus ROIs (i.e. blurry images). This was done by extracting ROIs from the same subset of deployments but with different threshold settings. Ultimately, the thresholds determine the number of ROIs in the image data set, their distribution among classes, and the associated imaged volume. The settings used corresponded to an imaged volume of 0.27 l $frame^{-1}$. The number of ROIs obtained across all deployments was 106392.

Annotation

Approximately 10% of the total number of ROIs ($n = 9646$) were sampled from the full image set and labeled by manually sorting into class-specific folders. To obtain ROIs for annotation, an equal number of ROIs was randomly selected from each DAVPR deployment to ensure an unbiased representation of ROIs balanced among deployments. For each classification category, 70% of the image objects were allocated to classifier training and validation (of these, 80% training and 20% validation) and 30% to testing classifier performance (Fig. 3).

In the annotation step, ROIs were sorted into 16 classes with a minimum of 40 individuals per class (Fig. 3). Other classes with < 40 images were assigned to the "plankton_other" class. It was expected that the classifier would output relatively low probability scores for the plankton_other class, given that this class was characterized by inconsistency in visual attributes among instances (e.g. Campbell et al. 2020). The plankton_other class was, therefore, intended to be used to evaluate contamination of the 15 identified classes by rare taxa.

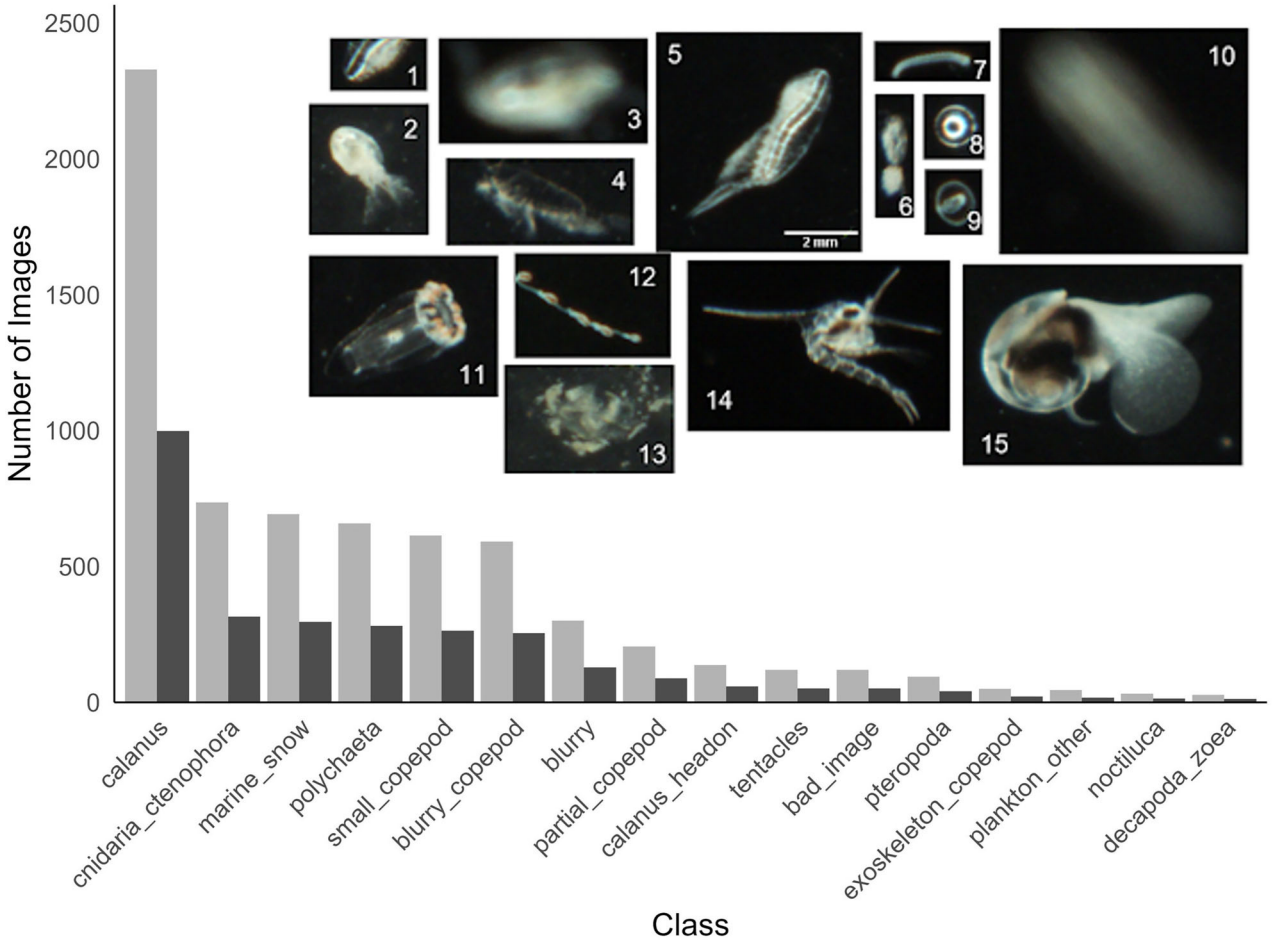

Figure 3. Annotated images used for training (grey) and testing (black) for each class. An example image for each class is shown in the inset. The number associated with each image indicates its class. 1, partial_copepod; 2, headon_copepod; 3, blurry_copepod; 4, exoskeleton_copepod; 5, *Calanus*; 6, small_copepod; 7, polychaeta; 8, bubble; 9, polychaeta_2 (potentially polygordiid trochophore); 10, blurry; 11, cnidaria_ctenophora; 12, tentacle; 13, marine_snow; 14, decapoda_zoea; 15, pteropoda. The scale bar in panel 5 of the inset corresponds to all images.

With the exception of plankton_other, the 15 classes were defined based on taxonomy, visual attributes, or to distinguish nuisance classes (i.e. blurry, bubble). In some cases, images belonging to the same taxonomic group were separated into multiple classes with the goal of reducing intra-class variation of visual attributes (e.g. polychaeta and polychaeta_2, cnidaria_ctenophora, and tentacle) or to account for human uncertainty in classification. For example, "*Calanus* adjacent" classes represented large copepods that were likely *Calanus* spp., as the *Calanus* spp. taxon dominated large copepods in associated plankton net samples (not shown), but could not be classified with the same level of confidence. Individuals that were (1) not oriented along their major axis of length were labeled headon_copepod; (2) out of focus were labeled blurry_copepod; and (3) partial objects not fully captured in the field of view were labeled partial_copepod (Fig. 3).

For each class, a subset of the image objects was verified by at least two humans that included a zooplankton ecologist and taxonomist. The three most abundant classes were *Calanus* spp. copepods (*Calanus*, 35%), cnidarian or ctenophore jellies (cnidaria_ctenophore, 11%), and detritus (marine_snow, 10%). The three least abundant classes included polychaeta_2, plankton_other, and decapoda_zoea (all < 1%) (Fig. 3).

Automated classification

Automated image classification was done on a Dell Precision 5820 machine equipped with a NVIDIA RXT4000 GPU, 8 GB of GPU RAM, and CUDA software (v. 10.1). The Keras deep learning Application Programming Interface was used for automated classification in R (R Core Team 2020, Chollet et al. 2022). Keras is built on top of the Tensorflow computing platform (Chollet et al. 2022), and Keras (v. 2.4.0) and Tensorflow (v. 2.3.0) were run in a Python (version 3.8.5) session in R. Interfaces to Keras and Tensorflow in R utilized the reticulate R package, which facilitates interoperability with Python (Ushey et al. 2025).

The Inception v3 CNN architecture (Iv3, Szegedy et al. 2016) was used as the classifier. Transfer learning was employed by parameterizing the Iv3 CNN with weights obtained from training on ImageNet (Deng et al. 2009). The final fully connected layer was modified such that there were 16 outputs. A softmax activation function computed a classification probability for each class and image, and automated classifications were determined by selecting the class for each image with the maximum probability score assigned by the classifier. Training was done over a total of eight epochs using the standard categorical cross-entropy loss function (e.g. Schanz et al. 2023). During training, only weights associated with the final

Table 1. Hyperparameter values over which the CNN was evaluated. All image augmentations (brightness, zoom, flip) were turned on and off together

Hyperparameter	Value
Image augmentation	TRUE, FALSE
Learning rate	0.00025, 0.001, 0.004
Dropout	0.25, 0.5
ReLU	TRUE, FALSE

dense layer were updated, while the remainder of the weights in the base model were frozen. Image objects were introduced in batches of 32 such that there were 170 steps per epoch. Prior to being introduced to the CNN, image objects were resized to 299 × 299 pixels without cropping or padding, and pixels were scaled to values between −1 and 1.

Hyperparameters of the CNN were tuned by training with different options for learning rate, image augmentation, dropout, and inclusion of a Rectified Linear Unit (ReLU) activation function (Table 1). Both dropout and ReLU contributed to sparce representation by zeroing elements in the feature vector received by the final dense layer. The dropout function zeroed values randomly to prevent overfitting, whereas ReLU zeroed only negative values, introducing non-linearity and allowing the network to learn complex patterns efficiently (Chollet et al. 2022). All 24 CNN configurations were evaluated and ranked by validation accuracy in the final epoch of training using the R package tfruns (Allaire 2024). Training and validation loss recorded over the training history were then examined, and models that exhibited signs of overfitting (e.g. Fig. 4) were disqualified.

The hyperparameters and their range of values were chosen carefully. Learning rate was initially evaluated over values of 10^{-4}, 10^{-3}, and 10^{-2}. However, all top ranking models were characterized by learning rates of 10^{-3} in preliminary tuning exercises; therefore, the range was subsequently narrowed to 2.5×10^{-4}, 10^{-3}, and 4×10^{-3}. A dropout rate of 0.5 was chosen (Srivastava et al. 2014), and a lower rate of 0.25 was also allowed to account for potential cases in which a rate of 0.5 was too restrictive (Table 1). Augmentations of training images (i.e. brightness, zoom, and flip) were visually examined, and the parameter ranges were selected to be consistent with variations observed among images. In the tuning exercise, all image augmentations were either turned on or off together. When turned on, brightness and zoom were augmented by randomly selecting adjustments from an even probability distribution with ranges of ± 0.25 and ± 0.2, respectively, and images were also randomly flipped horizontally, vertically, horizontally and vertically, or not flipped.

Performance evaluation

Of the 24 CNN configurations (Table 1), 5 were selected for further analysis based on rank of validation accuracy and training history (Fig. 4). The final CNN configuration (i.e. CNN configuration 2, Table 2) was selected after examination of class-specific performance metrics (recall, precision, F1, see Table 3) and confusion matrices derived from the test image set. High recall of the *Calanus* class was prioritized at the cost of lower precision and F1 (Table 3) to avoid loss

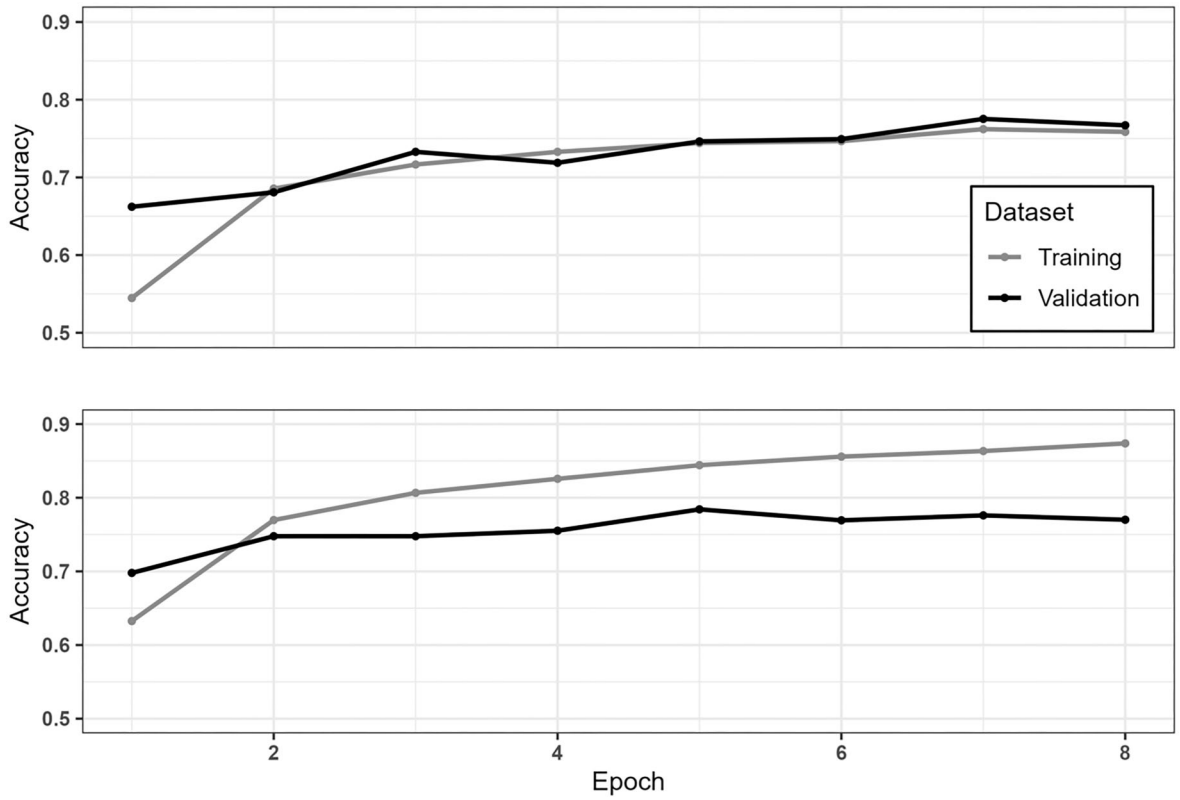

Figure 4. Training history indicating changes in accuracy of predictions derived from training and validation sets over training iterations (i.e. epochs) in which the classifier has cycled through all training images. The training history is shown for the final model configuration selected for image classification (upper panel) and the same configuration but without dropout and image augmentation (lower panel), which resulted in overfitting.

Table 2. Hyperparameter values for the top five CNN configurations.

CNN	ReLU	Learning rate	Dropout	Image augmentations
1	True	1×10^{-3}	0.50	True
2	**False**	$\mathbf{1 \times 10^{-3}}$	**0.50**	**True**
3	True	2.5×10^{-4}	0.25	True
4	False	2.5×10^{-4}	0.25	True
5	False	2.5×10^{-4}	0.50	False

The CNN selected for image classification is indicated in bold.

Table 3. Performance of the top five CNN configurations on the test set.

CNN	Average precision	Average recall	Average F1 score	Overall accuracy	*Calanus* precision	*Calanus* recall	*Calanus* F1
1	0.778	0.572	0.692	0.772	0.820	0.910	0.863
2	**0.659**	**0.577**	**0.681**	**0.755**	**0.735**	**0.960**	**0.832**
3	0.763	0.534	0.694	0.752	0.834	0.876	0.854
4	0.755	0.529	0.636	0.750	0.790	0.911	0.846
5	0.745	0.536	0.690	0.752	0.800	0.906	0.849

The average of precision, recall, F1 score and overall accuracy across all 16 classes are provided in addition to precision, recall, and F1 of the *Calanus* class. The CNN configuration selected for image classification is indicated in bold.

of *Calanus* images to other classes. Images in the *Calanus* class were far more abundant in the annotated set than most other classes; therefore, loss of relatively small proportions of *Calanus* could have large impacts on the precision of less abundant classes. The final CNN configuration that was selected had an overall accuracy (Eq. 4) of ~75%, and *Calanus* recall and precision of 0.96 and 0.74, respectively (Table 3).

Improving classification performance

A semiautomated image classification approach was used to achieve a performance standard (e.g. recall and precision > 0.9) higher than that of the top ranked automated classifiers (e.g. Table 3) to address project requirements for depth-stratified and class-specific mesozooplankton concentrations. This was done by manually verifying images with maximum probability score outputs below a threshold value.

For each image object in the test image set, the chosen CNN configuration (Table 2) was used to generate probability score predictions for all classes, and the class with the maximum probability score and the value of that probability score were retained. A reliability plot derived from these predictions (Fig. 5) resembled an "overconfident classifier" (Silva Filho et al. 2023). For example, predicted maximum probability scores between 0.2 and 0.4 were lower than the frequency of TP in the test set and those between 0.5 and 0.9 were higher than the frequency of TP in the test set. The distribution of maximum probability scores was strongly skewed, with more than half of the scores > 0.9 (Fig. 5). Despite not being perfectly calibrated, the reliability plot indicated a positive relationship between maximum probability score and proportion of TPs. Thus, manual verification of images based on probability score thresholds would indeed focus human verification effort where it is most likely to correct automated classification errors.

The relationship between probability score threshold and semiautomated classification performance (recall, precision, and F1) on the test set increased in a linear or sigmoidal relationship between thresholds of 0.25 and 1 (Fig. 6), and precision of the *Calanus* class increased linearly with the proba-

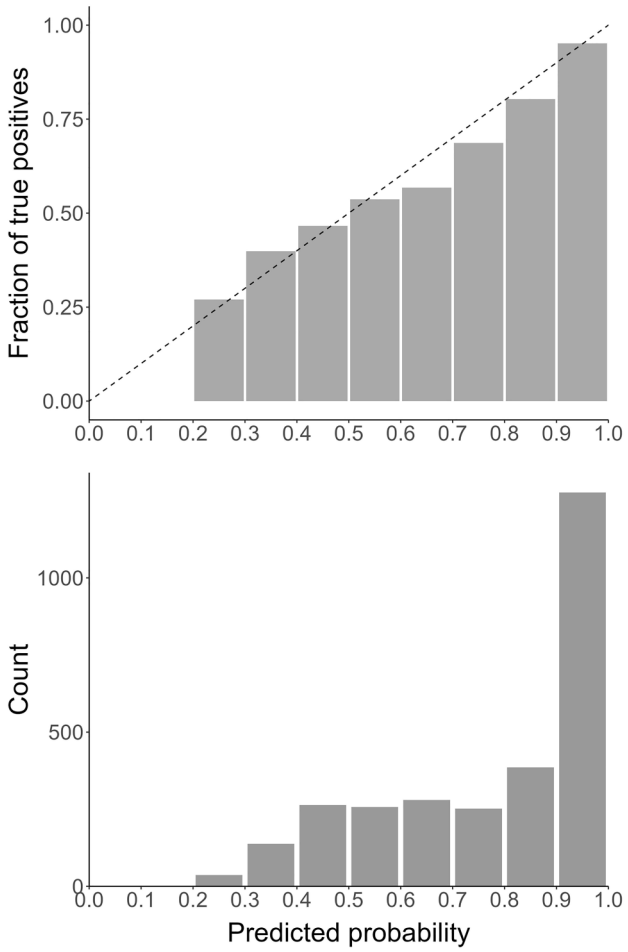

Figure 5. Reliability plot derived from predictions on test image set (top panel) and histogram indicating the frequency distribution of probability scores (bottom panel). In the reliability plot, the stippled line indicates the 1:1 relationship.

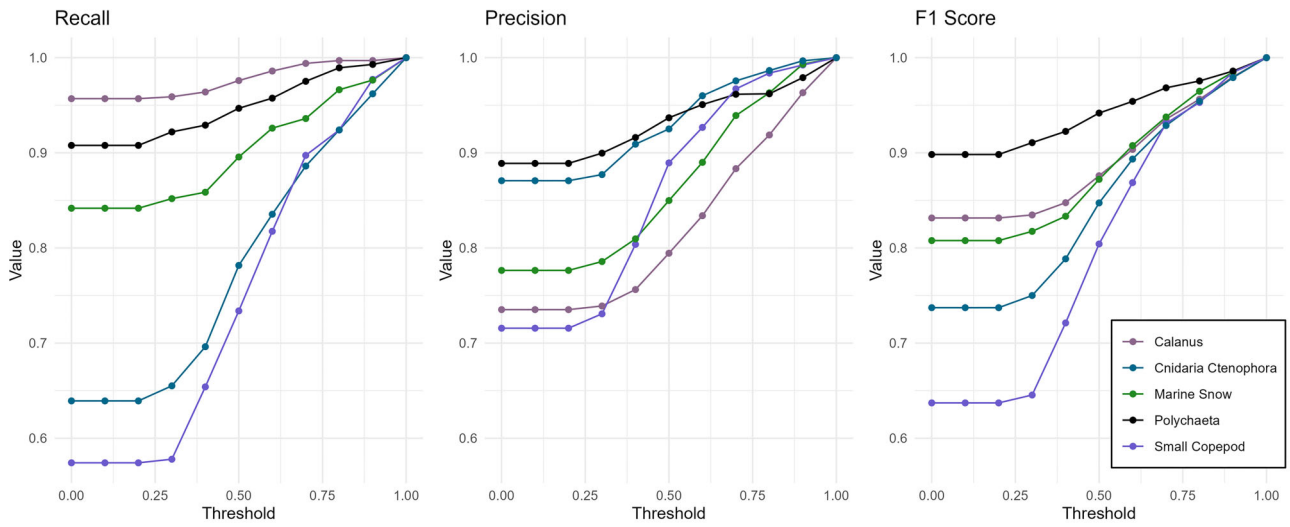

Figure 6. Relationships between class-specific evaluation metrics (recall, precision, and F1) and probability threshold for dominant classes in the test image set. Automated classifications of images with maximum probability scores below the threshold were manually verified. For example, a threshold of zero indicates all automated classifications were accepted without manual verification, whereas a threshold of 1 indicates no automated classifications were accepted and all classifications were manually verified.

bility score threshold from a minimum of ~0.75. The appropriate value of the probability score threshold depends on the performance standard and trade off with human effort. For example, the percentage of images that required manual verification in the full population of images objects (i.e. 106 392) increased linearly from probability score thresholds between 0.4 and 0.9, after which the rate increased substantially (Fig. 7). At a probability score threshold of 0.8, both recall and precision of the main classification categories were > 0.9 (Fig. 6), and 43% of the image population would need to be manually verified (Fig. 7). At a probability threshold of 0.9, precision and recall were > 0.95, and 56% of the image population would need to be manually verified. This level of manual verification (~50%) is intensive but not insurmountable, and it has strong potential to produce high quality data for the different mesozooplankton (and marine snow) classes.

Summary—Case Study 1

A semiautomated classification method was demonstrated that aims to achieve high precision and recall by accepting automated classification of only those images classified with confidence by the CNN. Transfer learning facilitated an effective use of a relatively small training set to obtain rapid automated predictions from the CNN. To optimize CNN hyperparameters, a tuning algorithm was implemented that ranked 24 CNN configurations based on validation accuracy. Of these configurations, the top five that did not show signs of overfitting were evaluated by a suite of performance metrics and examination of confusion matrices. While several candidate CNN configurations exhibited similar classification performance, the final classifier was chosen to prioritize recall within the *Calanus* class. Manual verification substantially improved precision of the *Calanus* class, but also improved precision and recall of less abundant classes with poor automated classification performance, notably larval decapods (decapoda_zoea) and copepod exuviae (exoskeleton_copepod) (Fig. 8).

This workflow contributed to the quantification of vertical distributions of late-stage *Calanus* spp. copepods in the southern Gulf of St. Lawrence, producing information to identify and characterize foraging habitat for *E. glacialis* in Canadian waters (Johnson et al. 2024). Since the research objective required accurate estimates of *Calanus* spp. copepod concentrations, there were particularly high standards of precision and recall in the image classification step. For example, in Johnson et al. (2024), a 0.9 probability threshold was used resulting in manual verification of slightly more than half of the full set of images. Practical use of the semiautomated classification method is dependent on the extent of manual effort required. This will depend on the number of ROIs in the full image set and the proportion of these images that require manual verification, which varies with classifier performance, calibration, and the performance standard of the user.

Case study 2: Classification of plankton images to investigate community response to harmful algal blooms in the North Sea

Background

The research objective of this case study was to assess the impact of a *Noctiluca scintillans* (henceforth *Noctiluca*) bloom on plankton community composition. *Noctiluca* is a non-toxic, heterotrophic dinoflagellate whose massive blooms disrupt ecosystems by reducing oxygen and increasing ammonium and negatively impact fishing and aquaculture yields (Rameshkumar et al. 2023). *Noctiluca* cells break easily in nets, which poses a challenge for quantification of *Noctiluca* dynamics using traditional sampling methods (Kordubel et al. 2024a). *In situ* imaging can overcome this problem while also quantifying spatiotemporal variation in morphological traits exhibited among *Noctiluca* individuals (Kordubel et al. 2024a).

In the summer of 2022, high-resolution *in situ* sampling of a *Noctiluca* bloom was carried out off the German Coast of Helgoland, an island 60 km off the German Coast of the North Sea (Fig. 9) and a known hotspot for *Noctiluca* blooms (Kordubel et al. 2024b). This case study describes the approach to classify plankton images from this dataset with emphasis on

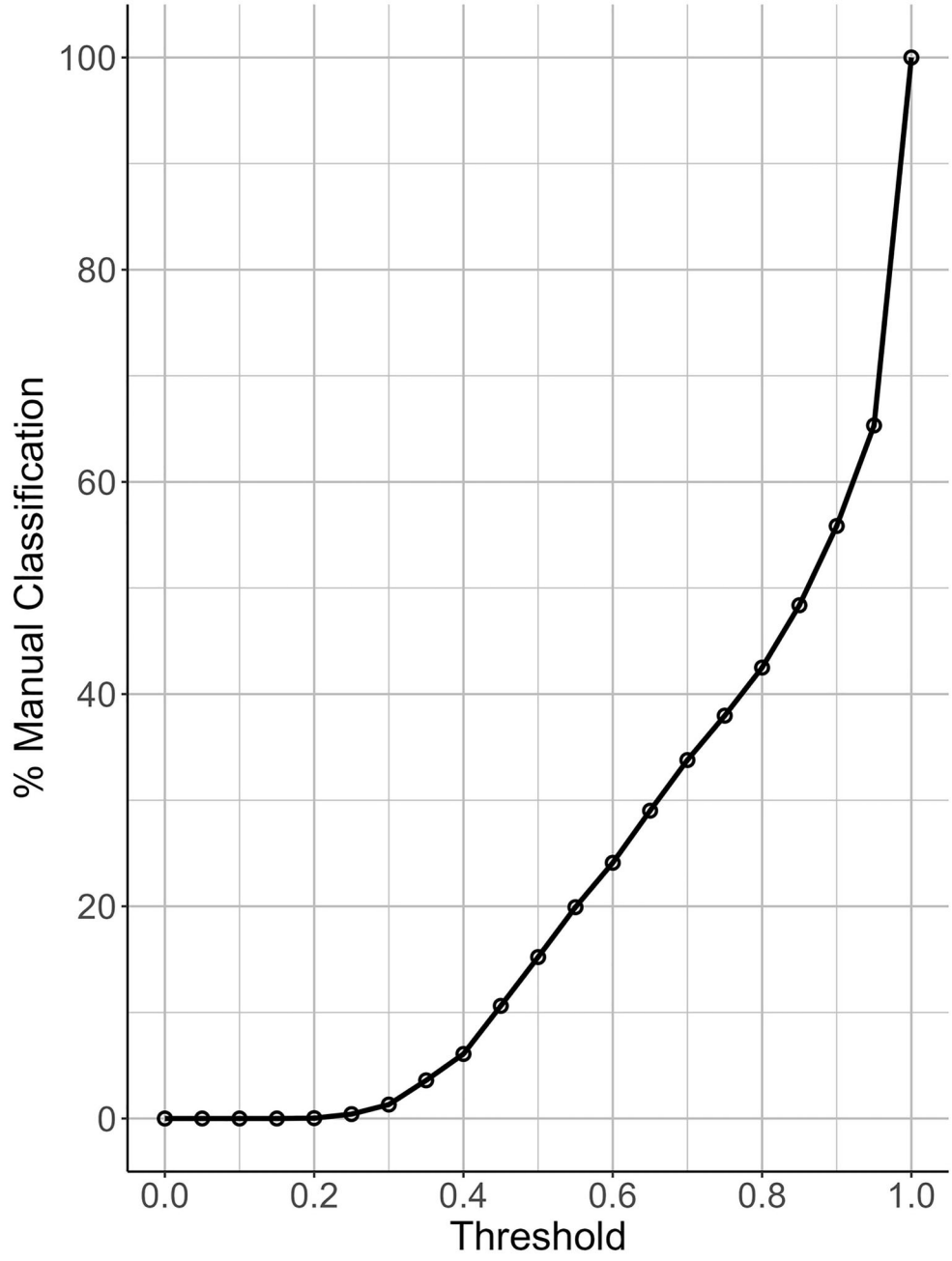

Figure 7. The relationship between the percentage of manual classification required for each probability threshold for the entire population of image objects ($n = 106\,392$).

identifying *Noctiluca* cells and large diatom chains. There was also interest in extracting and assessing morphological traits of *Noctiluca* through bloom progression. Image classification methods were developed in the context of large unannotated datasets acquired in the region (JERICO-RI COSNYA 2024), where manual labeling of images is unfeasible.

Acquisition

The CPICS (CoastalOceanVision, Inc.) was used to capture images of plankton sized > 200 μm. The CPICS was deployed in early June, late June, and early August of 2022 on a rosette frame that profiled from 0 to 50 m depth to sample successive stages of a *Noctiluca* bloom event (Fig. 9a). The dataset consisted of 26 CPICS deployments.

The CPICS is an automated *in situ* imaging microscope with an open flow design, providing dark field images of fragile organisms in their natural environment (Gallager 2019). Images were captured with a 6 MegaPixel Point Grey Grasshopper3 camera with a bi-telecentric lens and 0.9× magnification. A real-time frame rate recorded by the CPICS was used to calculate the total volume sampled. With these settings, the CPICS sampled a volume of 0.83 ml per full frame at an average of 13 frames per second, hence sampling at an average rate of ~10.75 mL s^{-1}. Each frame (2736 × 2192 pixels) was analysed in real-time to detect and save ROIs exceeding thresholds for area of consecutive pixels and pixel intensity. A total of 632 529 ROIs were used in this study. Examples of ROIs in predominant classes are shown in Fig. 9.

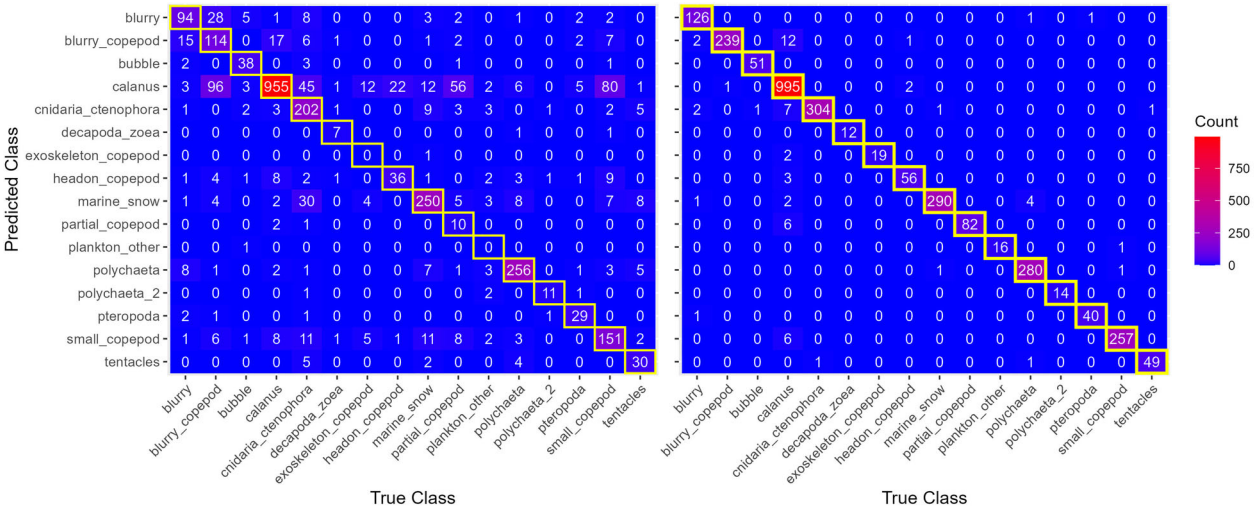

Figure 8. Confusion matrices derived from predictions on the test image set for probability thresholds of 0 (0% manual classification, left panel) and 0.9 (56% manual classification, right panel).

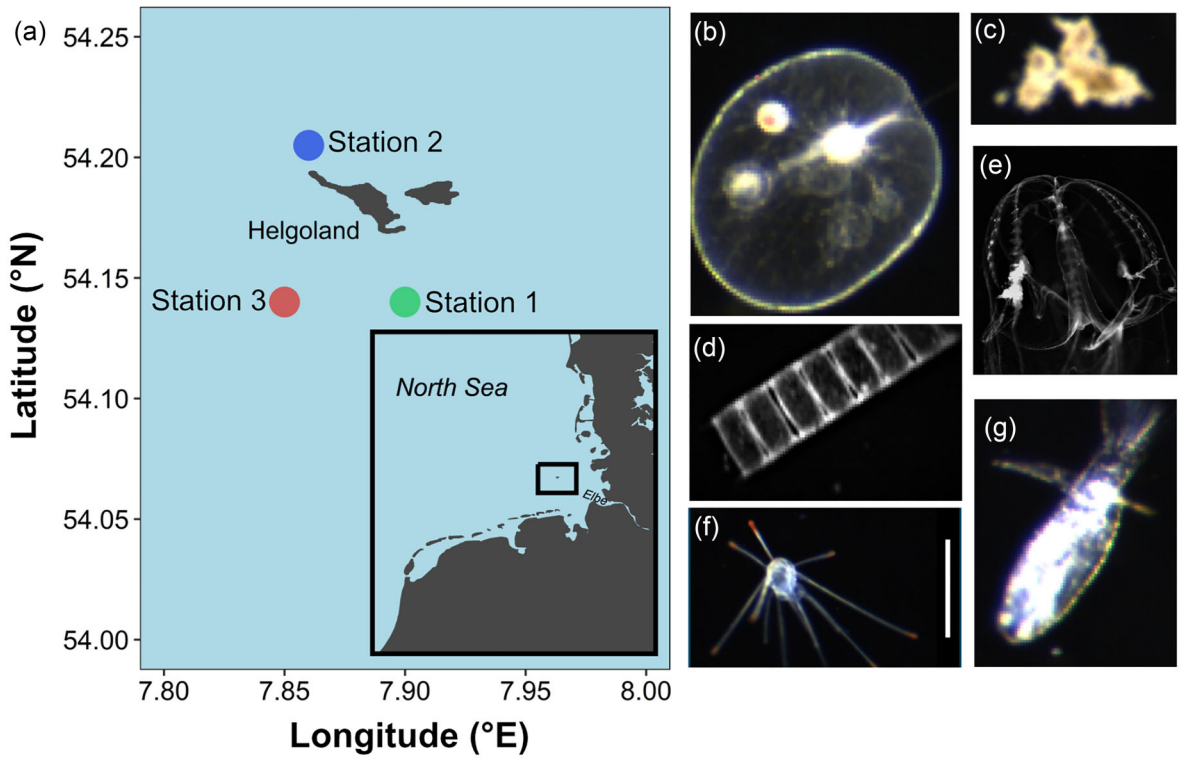

Figure 9: (a) Map of sampled stations. (b–g) Examples of images from Continuous Particle Image and Classification System (CPICS) belonging to predominant classes: (b) Noctiluca, (c) Marine snow, (d) Diatom chain, (e) Ctenophore, (f) Pluteus larva, and (g) Copepod.

Annotation

An iterative active learning process was employed to generate a labeled image set (Bochinski et al. 2018) for subsequent automated classification using the ResNet101 CNN architecture (see Automated classification subsection of this case study). This facilitated construction of a final set of images with class labels that included adequate representation of most plankton classes from a population of images with extreme class imbalance (Fig. 10). In the image annotation step, a high-performance computing (HPC) system and an open-source labeling platform (LabelStudio, Tkachenko et al. 2025) allowed multiple plankton identification experts (three experts were involved in the case study) to browse, filter, and label images concurrently online.

In the active learning process, a subset of 6617 ROIs was randomly selected from the larger image set and manually labeled. Twelve classes were identified from this subset, which was used for CNN training (3022 images), validation (622 images), and testing (2973 images). In the initial labeled image set, the classes Marine snow, Blurry, and *Noctiluca* occurred

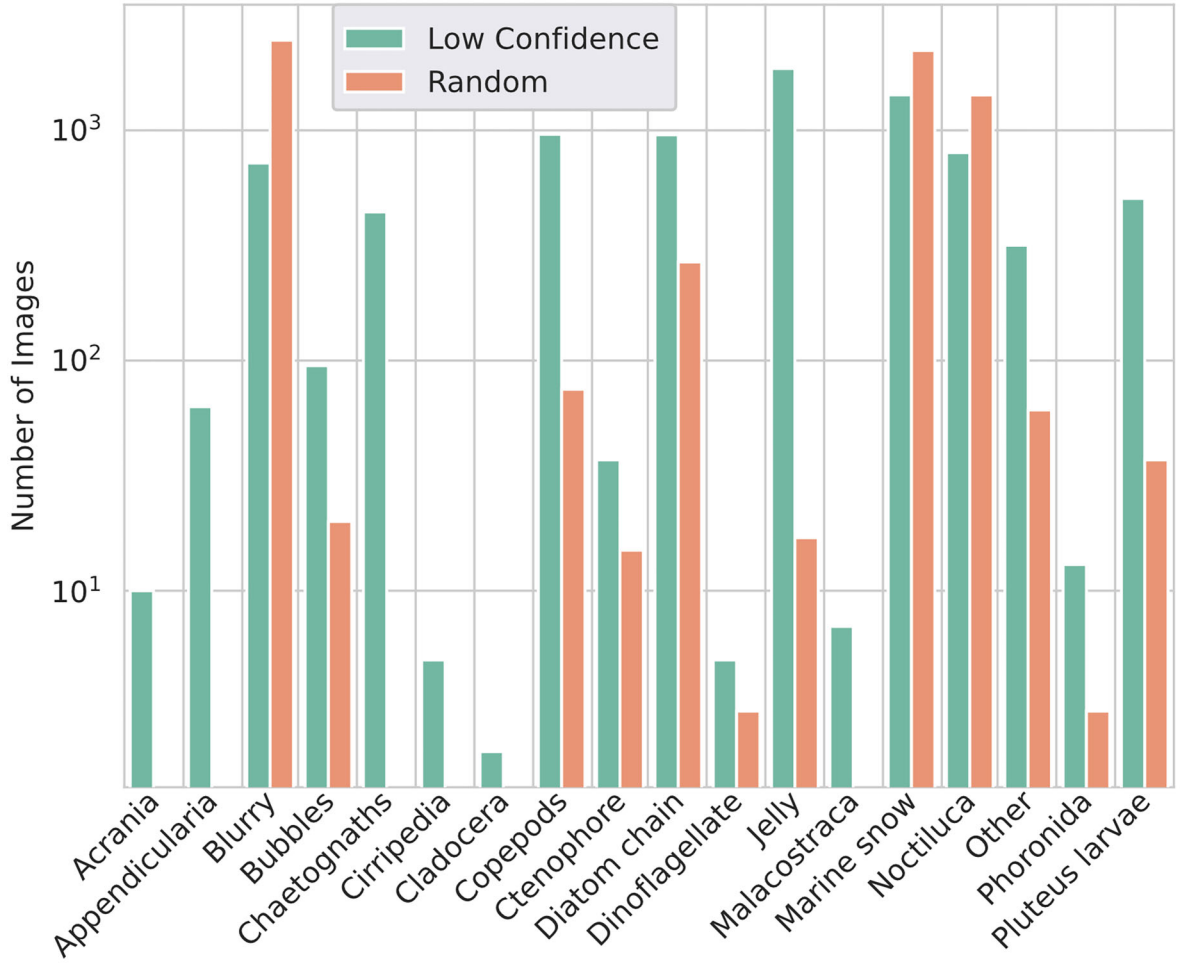

Figure 10. Comparison of the class distribution in the low confidence labels (green, probability score < 0.4, ~8000 images) and random subset labels (orange, ~6000). Class distribution in the low confidence subset is much more balanced and more classes are represented, while the random subset consisted of images predominantly assigned to Blurry, Marine Snow and *Noctiluca* classes.

far more frequently than other classes (Fig. 10). The CNN was trained and used to generate predictions for the full image set (i.e. ~6.3×10^5 ROIs), and images with a maximum probability score < 0.4 (8223 ROIs) were manually labeled. A probability score threshold of 0.4 was chosen to keep the size of the "Low Confidence" labeled image set comparable to the initial randomly selected subset. This resulted in a substantial increase in the number of labeled images belonging to non-dominant classes (Fig. 10). The updated and final annotated set was composed of 12 423 images, and 30% of these images were withheld for performance evaluation.

Automated classification

The machine learning framework deployed in this study utilized a semisupervised learning approach (Schanz et al. 2023), in which computationally expensive and unsupervised pretraining was followed by supervised fine-tuning. Three automated image classification methods were evaluated in this case study: (1) supervised classification only (hereafter referred to as "supervised"); (2) supervised classification initialized with CNN weights obtained from unsupervised pretraining on the entire set of ~6.3×10^5 images (hereafter referred to as "semisupervised"); and (3) supervised classification with initial CNN weights obtained from unsupervised pretraining conducted in a previous study (Schanz et al. 2023, hereafter referred to as "transfer learning"). In the transfer learning method, the weights were derived from a smaller set of CPICS images (on the order of 10^5) acquired close to the site of the *Noctiluca* bloom in this case study (Schanz et al. 2023). In all three methods, optimization and use of fixed CNN hyperparameters were carried out as detailed in Schanz et al. (2023). Machine classification was done using the ResNet101 CNN architecture (He et al. 2016) and implemented with PyTorch (Paszke et al. 2019), PyTorch Lightning (Falcon 2019), and Hydra (Yadan 2019) and run on 2 nodes with NVIDIA Tesla V100 GPUs.

In the supervised method, the ResNet101 CNN architecture was used but without a pretraining step. In the semisupervised and transfer learning methods, unsupervised pretraining was carried out using SimCLR contrastive learning (Chen et al. 2020a) on a ResNet101 backbone. During unsupervised pretraining, the image augmentations were identical to those reported in Schanz et al. (2023); however, decoupled contrastive learning (Yeh et al. 2022) was used to increase learning efficiency by reducing batch sizes relative to typical contrastive learning (Chen et al. 2020a). During pretraining, a batch size of 88 was used and the number of epochs was limited to a maximum of 60. The learning rate

was increased from 0 to 1 in the first 10 epochs and then reduced back to 0 over the remaining epochs using cosine decay.

After unsupervised pretraining, the SimCLR head ("feature extractor" head) was replaced with a single, fully connected softmax layer (referred to as "classifier head"), which allowed for supervised fine-tuning and classification tasks to be carried out. In the classifier head, a 2048-dimensional feature vector was taken as input and a probability score for each class in the labeled test set was output. During supervised fine-tuning, a standard cross-entropy loss (e.g. Schanz et al. 2023) was minimized on labeled images, all CNN weights were updated, and the number of epochs was limited to 1000 (Fig. 11).

Performance evaluation

Classification performance was evaluated by measuring the overall accuracy (Eq. 4) on the withheld validation image set during training (Fig. 11a,b) and conditional accuracy for each class (Eq. 5) in the test image set (Fig. 11c). Semisupervised and transfer learning methods exhibited similar model performances (Fig. 11), indicating that the transfer learning method can reduce computation and time demands of the pretraining step. In comparison to the supervised method, the semisupervised and transfer learning approaches reduced the number of training epochs required to achieve peak performance in agreement with previous work (Schanz et al. 2023). In addition, the semisupervised and transfer learning methods boosted overall accuracy (Fig. 11b) and were most effective in increasing conditional accuracy of classes with ∼100–1000 labels (e.g. Diatom chain, Copepods, Pluteus larvae, Chaetognaths, Jelly) (Fig. 11c). For classes with > 1000 labels (e.g. Marine snow, *Noctiluca* and Blurry), semisupervised and transfer learning methods exhibited more modest gains in conditional accuracy, but still outperformed the supervised method. The supervised method may be a viable option for classes with abundant training data if it meets the performance standards of the associated research question. In all three methods, conditional accuracies were very low for classes with < 100 labels (e.g. Acrania, Appendicularia, Cirripedia, Cladocera, Dinoflagellates, and Malacostraca), and more training data would be required to improve classification performance on these rare classes.

Anomaly detection: exploring labeled images for trait identification and novel class detection

Anomaly detection can be a powerful tool for locating images with features associated with morphological traits or classes that are uncommon in the overall image set, ultimately improving efficiency of the image annotation step (Schröder et al. 2020, Pastore et al. 2020, Ciranni et al. 2024). In this case study, feature vectors obtained from the CNN were used to find clusters of images with similar morphological traits. Specifically, feature vectors obtained from the pretraining feature extractor and classifier networks were analysed to identify "outlier" images, find clusters of images with similar characteristics, and explore the utility for expanding libraries of labeled images. An Isolation Forest algorithm was used to derive anomaly scores from the feature vectors (Liu et al. 2012). A strong positive anomaly score indicates the central position of the image in a cluster in multidimensional feature space, while a strong negative anomaly score indicates a peripheral position. Analysis of the feature vectors obtained from unsupervised pretraining resulted in ∼40 000 images (6% of the full image set) with negative anomaly scores (Fig. 12a).

Separate analyses were conducted on feature vectors obtained from the unsupervised pretraining feature extractor head (2048 dimensional feature space) and the semisupervised training classifier head (128 dimensional feature space) to explore the two feature spaces for outliers. The relative proportion of images in the periphery or the center of the cluster derived from both feature spaces were similar, but the cluster composition was different. The anomaly scores associated with images from the two feature spaces were correlated (Fig. 12b), but their differences reflected variations in image distribution within clusters, particularly in their peripheries.

Class frequencies from subsets of 10 000 images of the highest and lowest anomaly scores were explored and compared to class frequencies in the overall image set (Fig. 12e,f). The high anomaly score subsets primarily consisted of the most commonly occurring classes including Marine snow and *Noctiluca*, while low anomaly score subsets contained a large number of images in the Diatom chain, Blurry, or "clipped images" (i.e. ROIs containing partial objects). Notably, ∼7500 images of the highest anomaly scores from the classifier head were of *Noctiluca* in the process of phagocytosis (Fig. 12d,f), and ∼5000 images of the lowest anomaly scores from the unsupervised feature extractor were composed of diatom chains, predominantly belonging to the *Odontella* genus (Fig. 12c,e).

Summary—case study 2

In this case study, image annotation and classification steps were implemented with the goal of minimizing manual image labeling effort from an image set with extreme class imbalance. The active learning process increased the number of labeled images in several classes and increased the total number of classes from 11 to 18 in the final training set. The utility of a simple outlier detection algorithm for discovery of image groupings that share morphological traits was also examined. This method can speed up the development of trait-based libraries by identifying images for batch labeling. Groups of images with specific traits (e.g. diatom chains and *Noctiluca* undergoing phagocytosis) were identified, facilitating reliable classification of *Noctiluca* and their diatom prey and thereby contributing to the research objective of characterizing fine-scale spatiotemporal variation in *Noctiluca* concentration and investigating its impact on plankton community composition.

Unsupervised contrastive learning improved overall and conditional accuracies relative to the supervised-only method. For example, by annotating just 2% of the total image set, a 92% conditional accuracy was achieved for *Noctiluca* from the test image set and ∼140 000 images classified as *Noctiluca* were identified from the full image set. Inspection of 50% of the images classified as *Noctiluca* and 15% of images classified as diatoms confirmed that the classification accuracies met the standards in this study. Conditional accuracies of ∼80% were achieved on classes with ≥ 100 labeled images in the test set. Very rare classes with < 50 images were classified with very low conditional accuracies and too low to allow for reliable analysis. In previous work, ∼75% agreement was obtained between multiple experts highlighting an inherent uncertainty in both human- and machine-mediated classification (Culverhouse 2007, Schanz et al. 2023).

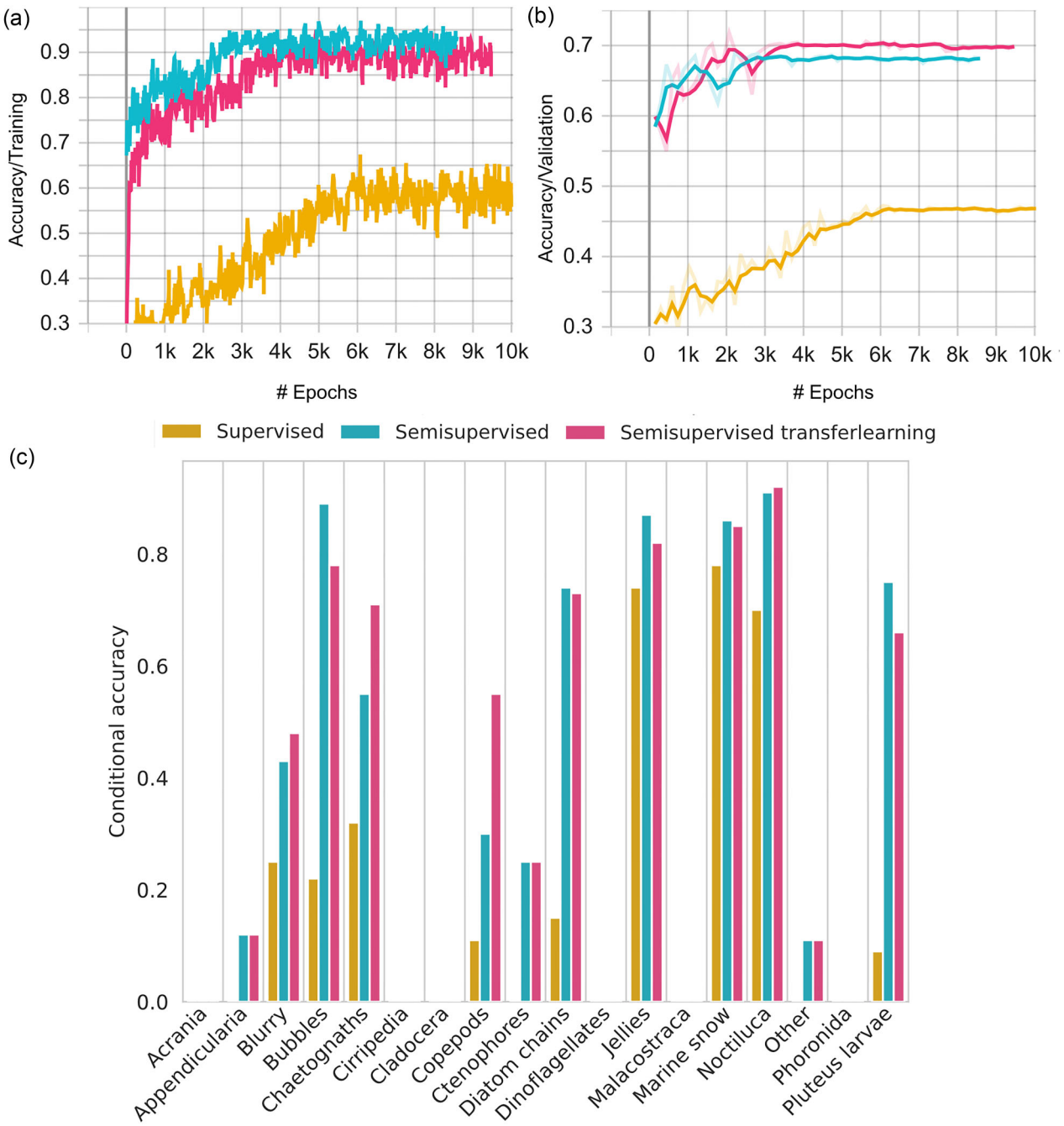

Figure 11. Comparison of supervised only (yellow), semisupervised (blue), and semisupervised with transfer learning (magenta) on training (a) and validation (b) accuracy. (c) Conditional accuracy for the various classes in the dataset using supervised only (yellow), semisupervised (blue) and semisupervised with transfer learning (magenta) methods.

Discussion

The two case studies demonstrate methodological approaches and techniques for overcoming challenges that arise in image classification of plankton to ultimately achieve classification performance deemed appropriate for respective research objectives. While the case studies followed the general image classification workflow reviewed herein, different methods were used in each study to address specific issues that arose in response to the ROI acquisition method, computing resources, and research objectives.

Case study 1 (CS1) highlighted a semiautomated method for practical and accurate quantification of the vertical distribution of copepods for characterization of the preyscape of a marine mammal planktivore (*E. glacialis*). Transfer learning facilitated rapid and effective automated classification with a relatively small training set. Probability score thresholds were used to balance strengths and weaknesses associated with automated prediction and human verification to achieve high precision and recall on the test set. Case study 2 (CS2) highlighted the utility of semisupervised learning, which leverages an unsupervised pretraining step (Chen et al. 2020a), for characterizing plankton community responses to harmful algal blooms. CS2 also highlighted active learning (Bochinski et al. 2018) and anomaly detection (Liu et al. 2012) methods

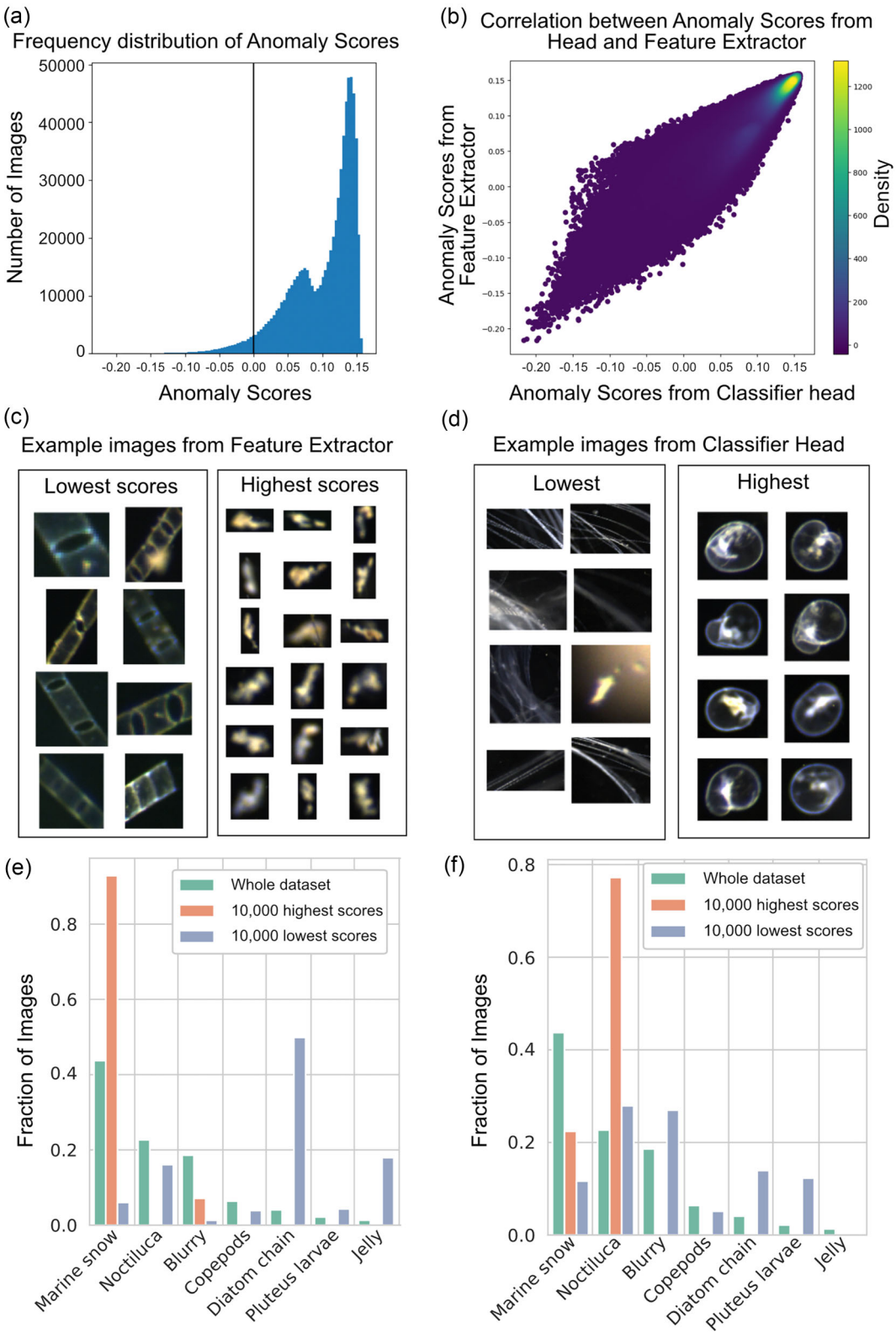

Figure 12: (a) Frequency distribution of anomaly scores from the 2048-dimensional feature space of the pretraining feature extractor over the entire dataset (632 529 images). (b) Correlation between anomaly scores from the feature extractor (y-axis) and classifier head (x-axis). (c) Example images from the lowest and highest anomaly scores from the pretraining feature extractor (d) from the fine-tuning classifier head. (e, f) Relative fraction of classes in the entire dataset (green), high anomaly score subset (orange), and low scoring subset (blue) with scores generated from the pretraining feature extractor (e) and fine-tuning classifier (f).

that can increase efficiency of building annotated image libraries by locating images belonging to different classes, especially underrepresented classes. Ultimately, these methods can be used to improve the ability of CNNs to accurately classify diverse and highly imbalanced datasets.

Both CS1 and CS2 utilized deep CNN architectures (e.g. Iv3 and ResNet) that have previously been shown to be effective tools for classifying plankton images (e.g. Lumini and Nanni 2019, Campbell et al. 2020, Schanz et al. 2023). In CS1, Iv3 was initialized with weights from training on ImageNet, a repository of annotated nonplankton images (Deng et al. 2009), and only weights associated with the final fully connected layer were updated during training on plankton images. This was done to quickly train a robust classifier with a relatively small amount of training data. A higher performance may be achieved by updating weights in additional layers (e.g. Ellen and Ohman 2024) at the potential costs increasing (1) the size of the annotated training set to prevent overfitting; and (2) computation time associated with more parameters being updated. In CS2, all CNN weights were updated during training. Both case studies used ∼10 000 images for training and testing, but the number of epochs to reach convergence in the supervised-only instance (no pretraining) of CS2 was two orders of magnitude higher than CS1. While it is not possible to isolate the cause of this discrepancy, the larger number of training epochs required in CS2 was likely at least in part due to the much larger number of parameters that required optimization. The contrastive semisupervised approach in CS2 reduced the number of epochs required to reach convergence and substantially increased performance, but only after an intensive contrastive learning pretraining step, which utilized hundreds of thousands of images.

Another notable difference between the case studies was that the class with the most images in the annotated set was a class of special interest in CS1 (*Calanus* spp. copepods), whereas that of CS2 was a nuisance class (blurry images). It can be difficult to find a sufficient number of images belonging to less abundant classes for generation of training, validation, and testing sets. This can result in strong class imbalance, which is a major challenge in image classification tasks because underrepresented classes in the training set contribute substantially less to training loss, ultimately resulting in reduced classification performance relative to dominant classes. In CS1, the abundance of *Calanus* images in the samples reflected their abundance in the zooplankton communities of the western North Atlantic (e.g. Pepin et al. 2015) and the ability of the DAVPR to detect them. The number of blurry images in CS1 was intentionally limited in the image segmentation step. In contrast, in CS2, a strong emphasis was placed on over-representing non-dominant classes and reducing the total number of images that required human annotation.

Both case studies highlighted the use of CNN outputs in different aspects of the image classification workflow, including feature vectors for anomaly detection and probability scores for human annotation. The latter contributed to building the training set in CS2 and verification of automated classifications in CS1. Probability scores may be misleading if obtained from models that are not calibrated (Silva Filho et al. 2023); therefore, it may be necessary to evaluate CNN calibration and the frequency distribution of probability scores in certain circumstances, as was done in CS1. While the *Calanus* class was a focus of CS1, manual verification of images with probability scores < 0.9 also corrected poor automated classification performance of rare classes (e.g. decapoda_zoea, polychaeta_2); however, more observations are needed to substantiate this outcome as the sample size for these classes was low in the test set.

It is important to note that classification performance of both case studies cannot necessarily be extrapolated beyond their test sets. In CS1, the derivation of classification performance metrics associated with different levels of human verification (Fig. 6) assumed no human error in verification. In reality, a certain level of human error is unavoidable (Culverhouse 2007), and human errors would need to be quantified to accurately measure the performance of the semiautomated method in practice. Moreover, evaluating only the test set (as was done here for CS1 and CS2) does not account for bias that can be introduced by spatiotemporal drift in class frequency or class-specific features in the target image set (González et al. 2017, Orenstein et al. 2020, Chen et al. 2025). In CS2, rebalancing the training set, as was done with the active learning method, could introduce classification bias in practice through its effect on the frequency of training images among classes (Schanz et al. 2023).

Addressing drift between the training and target domains was beyond the scope of the case studies in this paper; however, verification of the quality of quantitative data derived from image-based sampling beyond the test set is crucial. A qualitative or semiquantitative familiarization with the target image set (pre and postclassification) can help to identify potential for drift, and some human verification may be required in order to quantify and adjust for drift of the target domain from the test domain (e.g. Orenstein et al. 2020, Plonus et al. 2021, Conradt et al. 2022). In addition to adjustments in the classification processes, comparison of image-based data products with parallel data streams from other plankton sampling approaches can be used to validate or calibrate results from image-based analyses (e.g. Benfield et al. 1996, Le et al. 2022, Ollevier et al. 2025).

Conclusion

Effective image classification approaches and techniques are required to facilitate mainstream adoption of plankton imaging and exploit its advantages. Automated image classification is, therefore, not merely a technical innovation but a potentially transformative tool for marine ecology. By addressing current challenges such as class imbalance, human annotation effort, and data shifts, researchers can maximize the value of plankton imaging datasets, ultimately advancing capacity to quantify variability in different aspects of pelagic ecosystems. Approaches that minimize human effort in image classification are crucial for efficiently quantifying variations in plankton populations and communities, monitoring pelagic biodiversity, and detecting plankton responses to environmental change. The latter includes early identification of organisms that pose significant risks to marine ecosystems and human health, such as harmful algal blooms.

This paper underscores the increasing importance of automated image classification in plankton research. In particular, deep learning is a powerful tool to overcome challenges associated with manual image annotation and automated classification performance that contribute to the well-recognized classification bottleneck in quantitative image-based studies. Image classification of plankton is an active field of research (Ciranni et al. 2024), and classification tools are being de-

veloped that leverage millions of plankton images in a transfer learning pipeline (e.g. Ellen and Ohman 2024). However, many plankton ecologists will need to rely on smaller-scale practical methodologies to provide timely deliverables with limited resources in terms of hardware and personnel. Examples of such scenarios were given here in the form of the two case studies that utilized CNNs to build training sets, derive image features, and ultimately automate classification. The use of methods tailored to address specific research questions at each step in the image classification process demonstrates that satisfactory image classification performance can be obtained by users with a variable level of resources including computational power and expertise in computer vision. As image classification tools advance, future research will benefit from a synergistic approach blending traditional ecological expertise with state-of-the-art machine learning techniques.

Acknowledgement

The authors thank David Greenberg and Tobias Schanz for discussions that contributed to this manuscript. Comments from two anonymous reviewers improved the quality of the manuscript. The authors are grateful to the vessels, crews, and field teams that supported collection of data used in this study. Special thanks to Philipp Fischer and team at Helgoland for their support with data collection.

Author contributions

K.S. (Conceptualization [lead], Data curation [lead], Formal Analysis [lead], Investigation [lead], Methodology [lead], Project administration [lead], Supervision [supporting], Visualization [supporting], Writing – original draft [lead], Writing – review & editing [lead]), A.R.V. (Conceptualization [lead], Data curation [lead], Formal Analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing – original draft [lead], Writing – review & editing [lead]), A.H. (Data curation [supporting], Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Validation [lead], Visualization [lead], Writing – review & editing [supporting]), S.R. (Conceptualization [lead], Methodology [supporting], Writing – review & editing [lead]), E.O.G. (Methodology [supporting], Software [lead], Writing – review & editing [supporting]), K.O.M. (Conceptualization [lead], Funding acquisition [lead], Resources [lead], Supervision [lead], Writing – review & editing [lead]), C.L.J. (Conceptualization [lead], Funding acquisition [lead], Resources [lead], Supervision [lead], Writing – review & editing [lead]).

Conflict of interest: The authors have no conflicts of interest to declare.

Funding

Funding for this research was provided by the Fisheries and Oceans Canada Whales Initiative.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

Allaire JJ. *tfruns*: training run tools for TensorFlow. *R package version 1.5.3*. 2024. https://CRAN.R-project.org/package=tfruns(25 November 2025, date last accessed).

Bartz E, Bartz-Beielstein T, Zaefferer M *et al. Hyperparameter tuning for machine and deep learning with R—A practical guide*. Singapore: Springer Nature Singapore Pte Ltd., 2023. https://doi.org/10.1007/978-981-19-5170-1

Benfield MC, Davis CS, Wiebe PH *et al.* Video Plankton Recorder estimates of copepod, pteropod and larvacean distributions from a stratified region of Georges Bank with comparative measurements from a MOCNESS sampler. *Deep Sea Res Part II* 1996;**43**:1925–45. https://doi.org/10.1016/S0967-0645(96)00044-6

Benfield MC, Grosjean P, Culverhouse PF *et al.* RAPID: research on automated plankton identification. *Oceanography* 2007;**20**:172–87. https://doi.org/10.5670/oceanog.2007.63

Bi H, Guo Z, Benfield MC *et al.* A semi-automated image analysis procedure for in situ plankton imaging systems. *PLoS One* 2015;**10**:e0127121. https://doi.org/10.1371/journal.pone.0127121

Bochinski E, Bacha G, Eiselein V *et al.* Deep active learning for in situ plankton classification. In: Z Zhang, D Suter, Y Tian et al.(eds.), *Pattern Recognition and Information Forensics*. ICPR 2018. Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2018, 5–15. https://doi.org/10.1007/978-3-030-05792-3_1

Campbell RW, Roberts PL, Jaffe JS. The Prince William Sound plankton camera: a profiling in situ observatory of plankton and particulates. *ICES J Mar Sci*. 2020;**77**:1440–55. https://doi.org/10.1093/icesjms/fsaa029

Chen C, Kyathanahally S, Reyes M *et al.* Producing plankton classifiers that are robust to dataset shift. *Limnol Oceanogr: Methods* 2025;**23**:39–66. https://doi.org/10.1002/lom3.10659

Chen T, Kornblith S, Norouzi M *et al.* A simple framework for contrastive learning of visual representations. In: H Daumé, A Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*. Proceedings of Machine Learning Research (PMLR), 2020a, 1597–607. https://doi.org/10.48550/arXiv.2002.05709

Chen T, Kornblith S, Swersky K *et al.* Big self-supervised models are strong semi-supervised learners. *Adv Neural In Process Syst* 2020b;**33**:22243–55. https://doi.org/10.48550/arXiv.2006.10029

Chollet F, Kalinowski T, Allaire JJ. *Deep Learning with R*, 2nd edn. Shelter Island NY: Manning Publications, 2022

Ciranni M, Murino V, Odone F *et al.* Computer vision and deep learning meet plankton: milestones and future directions. *Image Vision Comput* 2024;**143**:104934. https://doi.org/10.1016/j.imavis.2024.104934

Conradt J, Börner G, López-Urrutia Á *et al.* Automated plankton classification with a dynamic optimization and adaptation cycle. *Front Mar Sci* 2022;**9**:868420. https://doi.org/10.3389/fmars.2022.868420

Culverhouse PF, Williams R, Benfield MC *et al.* Automatic image analysis of plankton: future perspectives. *Mar Ecol Prog Ser* 2006;**312**:297–309. https://doi.org/10.3354/meps312297

Culverhouse PF. Natural object categorization: man versus machine. In: N MacLeod, (ed.), *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, 1st edn. Boca Raton: CRC Press, 2007, 25–46. https://doi.org/10.1201/9781420008074-7

Davis CS, Gallager SM, Stewart WK. Rapid visualization of plankton abundance and taxonomic composition using the Video Plankton Recorder. *Deep Sea Res Part II* 1996;**43**:1947–70. https://doi.org/10.1016/S0967-0645(96)00051-3

Deng J, Dong W, Socher R *et al.* ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE, 2009, 248–55. https://doi.org/10.1109/CVPR.2009.5206848

Eerola T, Batrakhanov D, Vatankhah Barazandeh N *et al.* Survey of automatic plankton image recognition: challenges, existing solutions and future perspectives. *Artif Intell Rev* 2024;**57**:114. https://doi.org/10.1007/s10462-024-10745-y

Ellen JS, Graff CA, Ohman MD *et al.* Improving plankton image classification using context metadata. *Limnol Oceanogr: Methods* 2019;**17**:439–61. https://doi.org/10.1002/lom3.10324

Ellen JS, Ohman MD. Beyond transfer learning: leveraging ancillary images in automated classification of plankton. *Limnol Oceanogr: Methods* 2024;**22**:943–52. https://doi.org/10.1002/lom3.10648

Faillettaz R, Picheral M, Luo JY *et al.* Imperfect automatic image classification successfully describes plankton distribution patterns. *Methods Oceanogr* 2016;**15-16**:60–77. https://doi.org/10.1016/j.mio.2016.04.003

Falcon W, The PyTorch Lightning team. PyTorch Lightning. 2019. https://doi.org/10.5281/zenodo.3828935(25 November 2025, date last accessed).

Gallager SM. Continuous Particle Imaging and Classification System. 2019. Patent No. US 10,222,688 B2.

González P., Álvarez E., Díez J., *et al.*. Validation methods for plankton image classification systems. *Limnol Oceanogr: Methods* 2017;**15**:221–237.

Goodwin M, Halvorsen KT, Jiao L *et al.* Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES J Mar Sci* 2022;**79**:319–36. https://doi.org/10.1093/icesjms/fsab255

Gorsky G, Ohman MD, Picheral M *et al.* Digital zooplankton image analysis using the ZooScan integrated system. *J Plankton Res*. 2010;**32**:285–303. https://doi.org/10.1093/plankt/fbp124

Guo C, Pleiss G, Sun Y *et al.* On calibration of modern neural networks. In: D Precup, WY Teh (eds.), *Proceedings of the 34thInternational Conference on Machine Learning*. Proceedings of Machine Learning Research (PMLR), 2017, 1321–30. https://doi.org/10.48550/arXiv.1706.04599

He K, Zhang X, Ren S *et al.* Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2016,770–8. https://doi.org/10.1109/CVPR.2016.90

Huo Y, Qingxuan L, Dong J. Enhancing phytoplankton recognition through a hybrid dataset and morphological description-driven prompt learning. *J Mar Sci Eng* 2025;**13**:1680. https://doi.org/10.3390/jmse13091680

Irisson J.-O, Ayata S.-D, Lindsay DJ *et al.* Machine learning for the study of plankton and marine snow from images. *Annu Rev Mar Sci* 2022;**14**:277–301. https://doi.org/10.1146/annurev-marine-041921-013023

JERICO-RI COSNYA Underwater Node Helgoland. 2024. Available from: https://www.jerico-ri.eu/infrastructure/underwaternode helgoland/ (27 November 2025, date last accessed)

Johnson CL, Plourde S, Brennan CE *et al.* The southern Gulf of St. Lawrence as foraging habitat for the North Atlantic right whale. *DFO Can Sci Advis Sec Res Doc* 2024. **2024/077**. iv+43.

Kenitz KM, Orenstein EC, Anderson CR *et al.* Convening expert taxonomists to build image libraries for training automated classifiers. *Limnol Oceanogr Bulletin* 2023;**32**:89–97. https://doi.org/10.1002/lob.10584

Kerr T, Clark JR, Fileman ES *et al.* Collaborative deep learning models to handle class imbalance in FlowCam plankton imagery. *IEEE Access* 2020;**8**:170013–32. https://doi.org/10.1109/ACCESS.2020.3022242

Kordubel K, Baschek B, Hieronymi M *et al.* Improving the sampling of red *Noctiluca scintillans* to understand its impact on coastal ecosystem dynamics. *J Plankton Res* 2024;**46**:251–71. https://doi.org/10.1093/plankt/fbae010

Kordubel K, Martínez-Rincón RO, Baschek B *et al.* Long-term changes in spatiotemporal distribution of *Noctiluca scintillans* in the southern North Sea. *Harmful Algae* 2024;**138**:102699. https://doi.org/10.1016/j.hal.2024.102699

Langenkämper D, Zurowietz M, Schoening T *et al.* BIIGLE 2.0–browsing and annotating large marine image collections. *Front Mar Sci* 2017;**4**:83. https://doi.org/10.3389/fmars.2017.00083

Le K, Yuan Z, Syed A *et al.* Benchmarking and automating the image recognition capability of an in situ plankton imaging system. *Front Mar Sci* 2022;**9**:869088. https://doi.org/10.3389/fmars.2022.869088

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

Lee RF, Hagen W, Kattner G. Lipid storage in marine zooplankton. *Mar Ecol Prog Ser* 2006;**307**:273–306. https://doi.org/10.3354/meps307273

Li D, Du L. Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. *Artif Intell Rev* 2022;**55**:4077–116. https://doi.org/10.1007/s10462-021-10102-3

Li Q, Rigall E, Sun X *et al.* Dual autoencoder based zero shot learning in special domain. *Pattern Analysis and Applications* 2023;**26**:797–808. https://doi.org/10.1007/s10044-022-01109-9

Liu FT, Ting KM, Zhou Z-H *et al.* Isolation-based anomaly detection. *ACM Trans Knowl Discovery Data* 2012;**6**:1. https://doi.org/10.1145/2133360.2133363

Liu W, Anguelov D, Erhan D *et al.* SSD: Single Shot MultiBox Detector. In: B Leibe, J Matas, N Sebe et al.(eds.), *Computer Vision – ECCV 2016*. ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Lombard F, Boss E, Waite AM *et al.* Globally consistent quantitative observations of planktonic ecosystems. *Fronti Mar Sci* 2019;**6**:196. https://doi.org/10.3389/fmars.2019.00196

Lumini A, Nanni L. Deep learning and transfer learning features for plankton classification. *Ecol Inf* 2019;**51**:33–43. https://doi.org/10.1016/j.ecoinf.2019.02.007

Luo JY, Irisson J-O, Graham B *et al.* Automated plankton image analysis using convolutional neural networks. *Limnol Oceanogr: Methods* 2018;**16**:814–27. https://doi.org/10.1002/lom3.10285

Malde K, Handegard NO, Eikvil L *et al.* Machine intelligence and the data-driven future of marine science. *ICES J Mar Sci* 2020;**77**:1274–85. https://doi.org/10.1093/icesjms/fsz057

Ohman MD. A sea of tentacles: optically discernible traits resolved from planktonic organisms in situ. *ICES J Mar Sci* 2019;**76**:1959–72. https://doi.org/10.1093/icesjms/fsz184

Oldenburg E, Kronberg RM, Niehoff B *et al.* DeepLOKI—a deep learning based approach to identify zooplankton taxa on high-resolution images from the optical plankton recorder LOKI. *Front Mar Sci* 2023;**10**:1280510. https://doi.org/10.3389/fmars.2023.1280510

Ollevier A, Mortelmans J, Boone W *et al.* Picturing plankton: complementing net-based plankton community assessments with optical imaging across diverse marine environments. *Limnol Oceanogr: Methods*. 2025;**23**:246–60. https://doi.org/10.1002/lom3.10674

Orenstein EC, Kenitz KM, Roberts PLD *et al.* Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. *Limnol Oceanogr: Methods* 2020;**18**:739–53. https://doi.org/10.1002/lom3.10399

Orenstein EC, Saberski E, Briseño-Avena C. Discovery and dynamics of a cryptic marine copepod–parasite interaction. *Mar Ecol Prog Ser* 2022;**691**:29–40. https://doi.org/10.3354/meps14072

Panaïotis T, Caray–Counil L, Woodward B *et al.* Content-aware segmentation of objects spanning a large size range: application to plankton images. *Front Mar Sci* 2022;**9**:870005. https://doi.org/10.3389/fmars.2022.870005

Pastore VP, Zimmerman TG, Biswas SK *et al.* Annotation-free learning of plankton for classification and anomaly detection. *Sci Rep* 2020;**10**:11201. https://doi.org/10.1038/s41598-020-68662-3

Paszke A, Gross S, Massa F *et al.* Pytorch: an imperative style, high-performance deep learning library. In: H Wallach, H Larochelle, A Beygelzimer et al.(eds.), *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates, Inc, 2019. https://doi.org/10.48550/arXiv.1912.01703

Pepin P, Johnson CL, Harvey M *et al.* A multivariate evaluation of environmental effects on zooplankton community structure in the

western North Atlantic. *Prog Oceanogr* 2015;**134**:197–220. https://doi.org/10.1016/j.pocean.2015.01.017

Pershing AJ, Stamieszkin K. The North Atlantic ecosystem, from plankton to whales. *Annu Rev Mar Sci* 2020;**12**:339–59. https://doi.org/10.1146/annurev-marine-010419-010752

Picheral M, Colin S, Irisson J-O. Ecotaxa, a tool for the taxonomic classification of images. 2017. https://ecotaxa.obs-vlfr.fr (25 November 2025, date last accessed).

Plonus R-M, Conradt J, Harmer A *et al.* Automatic plankton image classification—Can capsules and filters help cope with data set shift? *Limnol Oceanogr: Methods* 2021;**19**:176–95. https://doi.org/10.1002/lom3.10413

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. 2020. https://www.R-project.org (26 November 2025, date last accessed).

Rameshkumar P, Thirumalaiselvan S, Raman M *et al.* Monitoring of Harmful Algal Bloom (HAB) of *Noctiluca scintillans* (Macartney) along the Gulf of Mannar, India using in-situ and satellite observations and its impact on wild and maricultured finfishes. *Mar Pollut Bull* 2023;**188**:114611. https://doi.org/10.1016/j.marpolbul.2023.114611

Redmon J, Divvala S, Girshick R *et al.* You only look once : unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE, 2016, 779–88. https://doi.org/10.1109/CVPR.2016.91

Richards BL, Beijbom O, Campbell MD *et al.* . Automated analysis of underwater imagery: Accomplishments, products, and vision. *U.S. Dept. of Commerce, NOAA Technical Memorandum NOAA-TM-NMFS-PIFSC-83. 59* 2019. 50 https://doi.org/10.25923/0cwf-4714.

Rubbens P, Brodie S, Cordier T *et al.* Machine learning in marine ecology: an overview of techniques and applications. *ICES J Mar Sci* 2023;**80**:1829–53. https://doi.org/10.1093/icesjms/fsad100

Schanz T, Möller KO, Rühl S *et al.* Robust detection of marine life with label-free image feature learning and probability calibration. *Mach Learn: Sci Technol* 2023;**4**:035007. https://doi.org/10.1088/2632-2153/ace417

Schröder S-M, Kiko R, Koch R. MorphoCluster: efficient annotation of plankton images by clustering. *Sensors*. 2020;**20**:3060. https://doi.org/10.3390/s20113060

Silva Filho TM, Song H, Perello-Nieto M *et al.* Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach Learn* 2023;**112**:3211–60. https://doi.org/10.1007/s10994-023-06336-7

Sorochan KA, Plourde S, Johnson CL. Near-bottom aggregations of *Calanus* spp. copepods in the southern Gulf of St. Lawrence in summer: significance for North Atlantic right whale foraging. *ICES J Mar Sci* 2023;**80**:787–802. https://doi.org/10.1093/icesjms/fsad003

Srivastava N, Hinton G, Krizhevsky A *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.

Szegedy C, Vanhoucke V, Ioffe S *et al.* Rethinking the Inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition*. New York, NY: IEEE, 2016, 2818–26. https://doi.org/10.1109/CVPR.2016.308

Tkachenko M, Malyuk M, Holmanyuk A *et al.* Label Studio: data labeling software. 2025. https://github.com/HumanSignal/label-studio (25 November 2025, date last accessed).

Ushey K, Allaire J, Tang Y. *reticulate*: interface to Python. *R package version* 1.42.0. 2025. https://github.com/rstudio/reticulate, https://rstudio.github.io/reticulate (25 November 2025, date last accessed).

Wojciuk M, Swiderska-Chadaj Z, Siwek K *et al.* Improving classification accuracy of fine-tuned CNN models: impact of hyperparameter optimization. *Heliyon* 2024;**10**:e26586. https://doi.org/10.1016/j.heliyon.2024.e26586

Yadan O. Hydra—a framework for elegantly configuring complex applications. *GitHub*. 2019. Available from: https://github.com/facebookresearch/hydra (25 November 2025, date last accessed)

Yeh C-H, Hong C-Y, Hsu Y-C *et al.* Decoupled contrastive learning. In: S Avidan, G Brostow, M Cissé et al.(eds.), *Computer Vision – ECCV 2022*. ECCV 2022. Lecture Notes in Computer Science. Cham: Springer, 2022, 668–84. https://doi.org/10.1007/978-3-031-19809-0_38

Handling Editor: Howard Browman