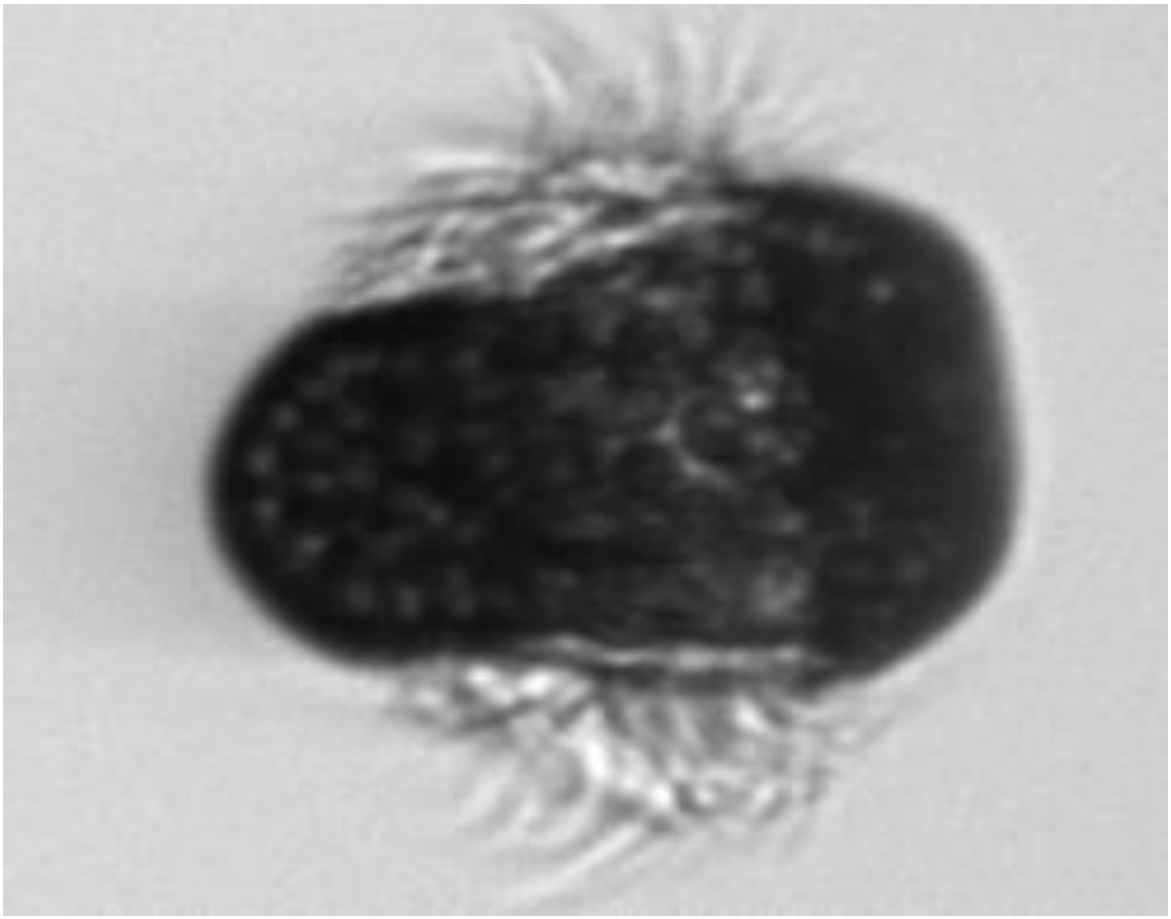
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Biotechnology

Second cycle, 30 credits

Machine learning-assisted image analysis and metabarcoding for monitoring of plankton in the Baltic Sea

KARIN GAREFELT



Stockholm, Sweden 2023

Contents

Abstract

In environmental monitoring of the seas around Sweden, manual counting with microscope is used to monitor the plankton communities and algal blooms. New techniques are currently being evaluated, including imaging flow cytometry and DNA metabarcoding, but it is not known how results from the different techniques relate to one another. Previous work has not compared imaging flow cytometry with metabarcoding, although both methods have been compared to traditional microscopy. In this project, samples for DNA metabarcoding and imaging flow cytometry with the Imaging FlowCytobot (IFCB) have been collected in parallel in the Baltic Proper, Öresund, Kattegat, and Skagerrak. To be able to process the large amount of images from the cytometry, an image classification algorithm based on convolutional neural networks and transfer learning was developed, which was used to classify the images collected and compare the classification results with 18S rRNA metabarcoding of the protist community. This new approach of comparing imaging flow cytometry with metabarcoding resulted in strong (R>0.8) correlation for some diatom taxa, but discrepancies between the technologies were also observed. The discrepancies can be further studied to identify weaknesses in both techniques and refine them further.

Keywords: IFCB, Imaging FlowCytobot, Baltic Sea, imaging-in-flow cytometry, phytoplankton imaging, convolutional neural network, metabarcoding

Sammanfattning

I miljöövervakningen av haven runt Sverige har manuell mikroskopi av plankton länge varit den huvudsakliga tekniken för att övervaka växtplanktonbestånden och algblomningar. Nya tekniker utvärderas, men det är inte känt hur resultaten från de nyare teknikerna relaterar till varandra. Två tekniker som utvärderas av SMHI, flödesmikroskopi och DNA-streckkodning, har inte tidigare jämförts i litteraturen. Båda teknikerna har dock jämförts med traditionell mikroskopi. I det här projektet har provserier för DNA-streckkoding och automatiserad mikroskopi med Imaging FlowCytobot (IFCB) samlats in parallellt under en expedition i Egentliga Östersjön, Öresund, Kattegatt och Skagerrak. En bildklassificerare konstruerades med ett konvolutionellt neuronnät, som användes för att klassificera bilderna som tagits med IFCB:n och jämföra resultaten med DNA-streckkodning av 18S rRNA-genen. Jämförelsen visade stark korrelation mellan klassificeringen av bilder och DNA-streckkodning för vissa kiselalger (R>0.8), men teknikernas resultat skilde sig också åt i många fall. Skillnaderna kan studeras för att hitta svagheter i de båda teknikerna och utveckla dem vidare.

Acronyms

ASV amplicon sequence variant

CNN Convolutional Neural Network

IFCB Imaging FlowCytobot

PR^2 Protist Ribosomal Reference

ReLU Rectified Linear Unit

ResNet Residual Network

SMHI The Swedish Meteorological and Hydrological Institute

SYKE The Finnish Environment Institute

WoRMS World Register of Marine Species

1 Introduction

1.1 Phytoplankton Monitoring

In Sweden, phytoplankton communities are regularly monitored within the Swedish National Marine Monitoring Programme. The data is used for marine planning, identification of algal blooms, and tracking progress towards environmental goals like Environmental quality objectives. Today, the phytoplankton abundance (cells/liter), species composition, chlorophyll-a concentration (µg/L), biovolume of phytoplankton (mm^3/liter), and biomass of phytoplankton (mg C/L) are measured, and satellite images are used to estimate algal bloom intensities during summertime. [1] [2] Taxonomic analysis by microscopy is an important method to monitor species composition. One of the ships used to monitor the state of the seas surrounding Sweden is the research vessel Svea.

1.2 Imaging FlowCytobot

Phytoplankton abundance and species composition are normally determined with microscopy, but a method with higher throughput and reproducibility is currently being evaluated by SMHI, which carries out plankton monitoring in Sweden. It is based on using Imaging FlowCytobot (IFCB), an automated microscope. [1]

The Imaging FlowCytobot is an automated flow cytometer that captures images of phytoplankton and microzooplankton in the size range of 10 to 100 µm. It was introduced by Heidi M. Sosik and Robert J. Olsen in 2007. [3] Each Imaging FlowCytobot includes a quartz flow cell, where cells are flowing in a single file through a red (635 nm) laser beam. When phytoplankton and other chlorophyll-containing cells pass through the laser beam, they emit red fluorescence (680 nm), which triggers a flash and a camera. For each cell passing in the file, an image is produced and metadata is saved. With this capture method, the Imaging FlowCytobot can generate more than 10 000 plankton images per hour. This data volume makes manual classification impractical. Therefore, Sosik and Olsen also presented a machine learning-based technique for quantifying the abundance of phytoplankton

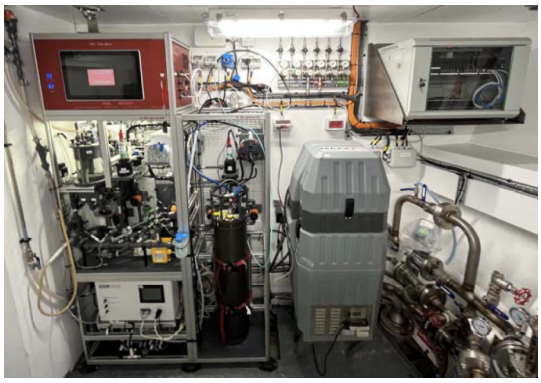


Figure 1: The Imaging FlowCytobot onboard Svea

taxa by classification of individual plankton images. [3]

Since the introduction of the Imaging FlowCytobot, it has been used for different applications all over the world. In the Gulf of Mexico, a system automatically detects harmful algal blooms of *Karenia brevis*, and alerts state agency representatives in Texas. [4] In the Amundsen Sea in Antarctica, it has been used to research the effect of ice melting on phytoplankton communities. [5] SMHI plans to include IFCB as a regular tool on the R/V Svea cruises, and continuously sample at 4 meters depth with the Ferry Box system. [6] This will enable monitoring at a higher spatiotemporal resolution than the traditional microscopy method. The Imaging FlowCytobot on board R/V Svea is the black, cylinder-shaped tool in Figure 1.

The first image classifier for IFCB data was presented in 2007. It was developed with a dataset consisting of 6600 images from an IFCB, all manually classified into 22 categories. Image features that reflect the size, shape, and other features of the captured object were extracted in MATLAB using image processing functions. With this data set consisting of features of IFCB images, a support vector machine was developed. [3] Previously, most classifiers for plankton image data were based on hand-crafted features extracted from images just like the first classifier. However, it has been shown that deep neural networks can learn features that are superior to hand-crafted features for image classification. [7]

1.3 Convolutional Neural Networks

1.3.1 CNN Architecture

Convolutional Neural Networks (CNN:s or ConvNets) are a class of artificial neural networks commonly used in image classification tasks. They consist of two sequential sections: feature learning and classification, as depicted in Figure 2. Typically, there are four types of layers in a CNN: convolutional layers, activation (ReLU) layers, pooling layers, and fully connected layers. [8]

The convolutional layers include filters or kernels, which are passed over subsections of the image. For each image section, the scalar product between the filter and the section is computed to the output of the layer. This is illustrated in Figure 3. The convolutional layers detect features of an image, and the outputs of convolutional layers are called feature maps or convolutional maps. Each filter produces a feature map, and the combined maps from all filters can therefore be larger than the input to the convolutional layer. This can be observed in the first step in Figure 2, and in the four maps produced in Figure 3. Not all inputs to a convolutional layer are connected to all outputs, which means these layers are not fully connected.

The convolutional layers are followed by activation layers, often in the form of Rectified Linear Unit layers (ReLU layers). ReLU replaces negative numbers in

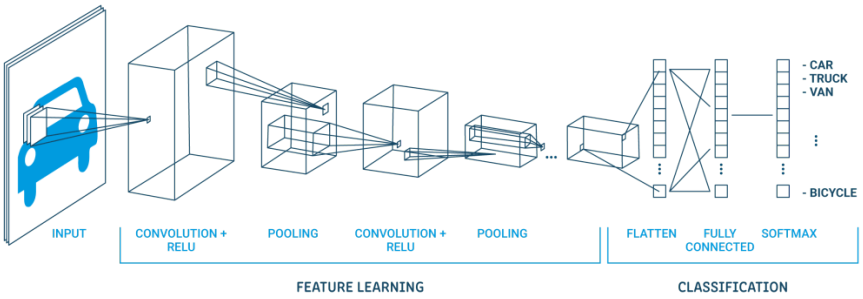


Figure 2: The structure of a small Convolutional Neural Network (CNN). Credit: [9]

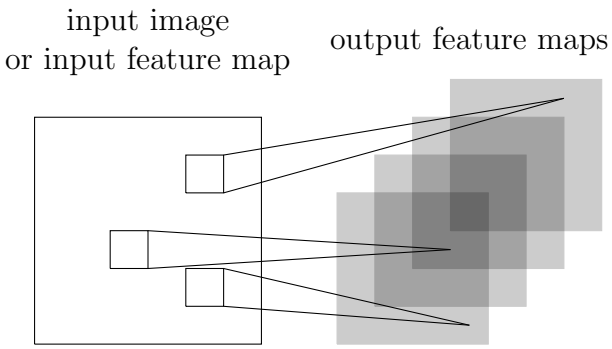


Figure 3: Illustration of a single convolutional layer. Credit: [10]

the feature map with zeros, which creates non-linearity in the net. Following the activation layer, there are pooling layers to reduce the size of the image. In Figure 2, this is illustrated as the volume decreasing in the pooling step. The main types of pooling are max pooling (retaining the maximum pixel value from an area) and average pooling (retaining the average pixel value from an area). As seen in Figure 2, the feature learning section of the CNN consists of multiple iterations of convolutional layers, activation layers, and pooling layers. [8]

The classification section consists of fully connected layers (FC). The feature maps from the last step in the feature extraction are flattened, and forwarded through one or more fully connected layers. This section can also include pooling layers and activation layers. The last step in the classification section is the softmax function which assigns a probability to the image belonging to the different image classes. [8]

1.3.2 Residual Networks

Network depth (the number of stacked layers) has been identified as an important factor for success in classification tasks. However, simply adding more layers to a standard CNN will not necessarily improve performance as network depth introduces new problems including vanishing gradient descent, and in some cases can lead to a decrease in accuracy. To enable greater network depth without these problems, deep residual learning was introduced in 2015. The Residual Networks (ResNets) are a type of Convolutional Neural Network with added features. The defining feature of ResNets is the shortcut connections, which forward the output of one layer into the output of a later layer, thereby skipping one or more layers. ResNets have been out-performing other classification methods, and winning multiple image classification and detection competitions. [11]

1.3.3 Transfer learning

The training of CNN:s is computationally expensive and training time is usually counted in weeks. However, CNN:s already trained with one set of images can be used as a starting point for the training towards a new task. This is the concept of transfer learning. The network is then adapted to the new task by adding a new set of fully connected layers and/or fine-tuning the weights in the feature extraction section. [7]

1.3.4 Measuring Classification Performance

Multiple metrics can be used to measure classifier performance. In this report, the metrics precision, recall, and F1 score will be used. These are the definitions of precision and recall, where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1 score is the harmonic mean of the precision and recall, and therefore the value of the F1 score falls between the value of the precision and the recall. [12]

$$\text{F1} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

1.4 Metabarcoding

Metabarcoding is another technique for assessing abundance of phytoplankton taxa. Like monitoring with the Imaging FlowCytobot, metabarcoding can be used

to create datasets of high spatiotemporal resolution. Metabarcoding uses a genetic marker-based approach. Marker genes like 16S and 18S ribosomal RNA (rRNA) genes are amplified with PCR, and then sequenced. [13] Metabarcoding is superior to traditional microscopy and IFCB in identifying small plankton (< 10 µm). [14]

To evaluate if metabarcoding is suitable for plankton monitoring, multiple studies have been published comparing the method with traditional taxonomic analysis by microscopy. Currently, there is a joint project between Umeå University, KTH/SciLifeLab, and SMHI aiming to optimize metabarcoding for phytoplankton, compare with microscopy and harmonize the barcode databases with traditional taxonomic analysis. [15]

1.5 Objectives

To my knowledge, there are no publications comparing metabarcoding to imaging flow cytometry for assessing phytoplankton community structures. The working hypothesis is that there is a correlation between the biodiversity assessments with metabarcoding and IFCB. This project aims to compare the performance of the two techniques. The focus of the project will be more on the IFCB and classifier development, as it is a less established technique.

2 Materials and Methods

2.1 Classifier Development

2.1.1 Image Data Sources

The datasets for training and evaluation of classifier performance were obtained from two organizations: The Swedish Meteorological and Hydrological Institute (SMHI) and The Finnish Environment Institute (SYKE).

The Swedish dataset was collected in Tångesund (Skagerrak) by SMHI in the JERICO-NEXT project (`https://www.jerico-ri.eu`). An observatory with an Imaging FlowCytobot was installed in the area of Tångesund in 2016 in proximity to a mussel farm. Example images from the SMHI dataset are displayed in Figure 4. Between August and October 2016, the IFCB was used to study harmful algal blooms. Water was collected autonomously at six depths and analyzed in the IFCB. [16] A subset of the IFCB images is organized in folders by taxonomic class (annotated by microbiologists). The images organized in folders by class are in .png-format, and folders containing no images were excluded from the dataset. The SMHI dataset was obtained in personal communication with Bengt Karlson, a scientist at Research and Development Oceanography at SMHI.

Two separate annotated datasets were acquired from SYKE: SYKE-plankton_IFCB_2022 and SYKE-plankton_IFCB_2021. Example images from SYKE-plankton_IFCB_2022 are displayed in Figure 5. The dataset SYKE-plankton_IFCB_2022 is from water samples collected at multiple locations in the Baltic. Samples were collected on cruise ships M/S Finnmaid and Silja Serenade, and on the research station of Utö (Finland). On the Marine Research Station on Utö, an IFCB has been in use continually since early 2020, sampling water at a 5-meter depth 250 m offshore. The other dataset from SYKE, SYKE-plankton_IFCB_2021, consists of images collected with the IFCB on Utö during 2021. 59 samples were selected to form the SYKE-plankton_IFCB_2021 dataset, covering all weeks of the year. The selected samples were labeled by expert taxonomists into the same classes as SYKE-plankton_IFCB_2022, and an Unclassified group. This dataset covers seasonal variation, and is more similar to natural samples as all images from the selected samples are included. The datasets from SYKE are publicly available at `https://b2share.eudat.eu/records/abf913e5a6ad47e6baa273ae0ed6617a` and `https://b2share.eudat.eu/records/7c273b6f409c47e98a868d6517be3ae3`.

All three datasets have a strong imbalance between classes. This is illustrated in Figure 6. The image dimensions and sizes differ, as a reflection of the natural morphological diversity of phytoplankton. This is seen in both Figure 5 and 4.

The image classes have different taxonomic levels. For example, *Mesodinium rubrum* is a species, while Oscillatoriales is an order. [17] These are two examples from the SMHI dataset. SYKE explains the difference in taxonomic level as follows: "Due to differences in the features of the organisms visible in the images, which form the basis of the identification, some classes have been determined to the species level while others have been determined at a higher taxonomic level." [18]

2.1.2 Dataset Organization

The SMHI and SYKE-plankton_IFCB_2022 datasets were merged and split into training, validation, and test sets (60%, 20%, and 20%, respectively). Two classes were merged since they refer to the same taxon: *Mesodinium rubrum* (SYKE) with *Mesodinium rubrum* (SMHI) and *Dinophysis acuminata* with *Dinophysis acuminata* (SYKE). Also, a copy of each of the validation and test sets was complemented with an equal number of unclassifiable images from SYKE-plankton_IFCB_2021. These expanded sets were later used to determine probability thresholds for classification (validation set) and evaluate model performance (test set). These sets are more representative of natural phytoplankton communities, as they include difficult-to-classify images. The training set was modified by oversampling images until > 100 images per class were reached. This is an effort to reduce class imbalance in the training dataset.

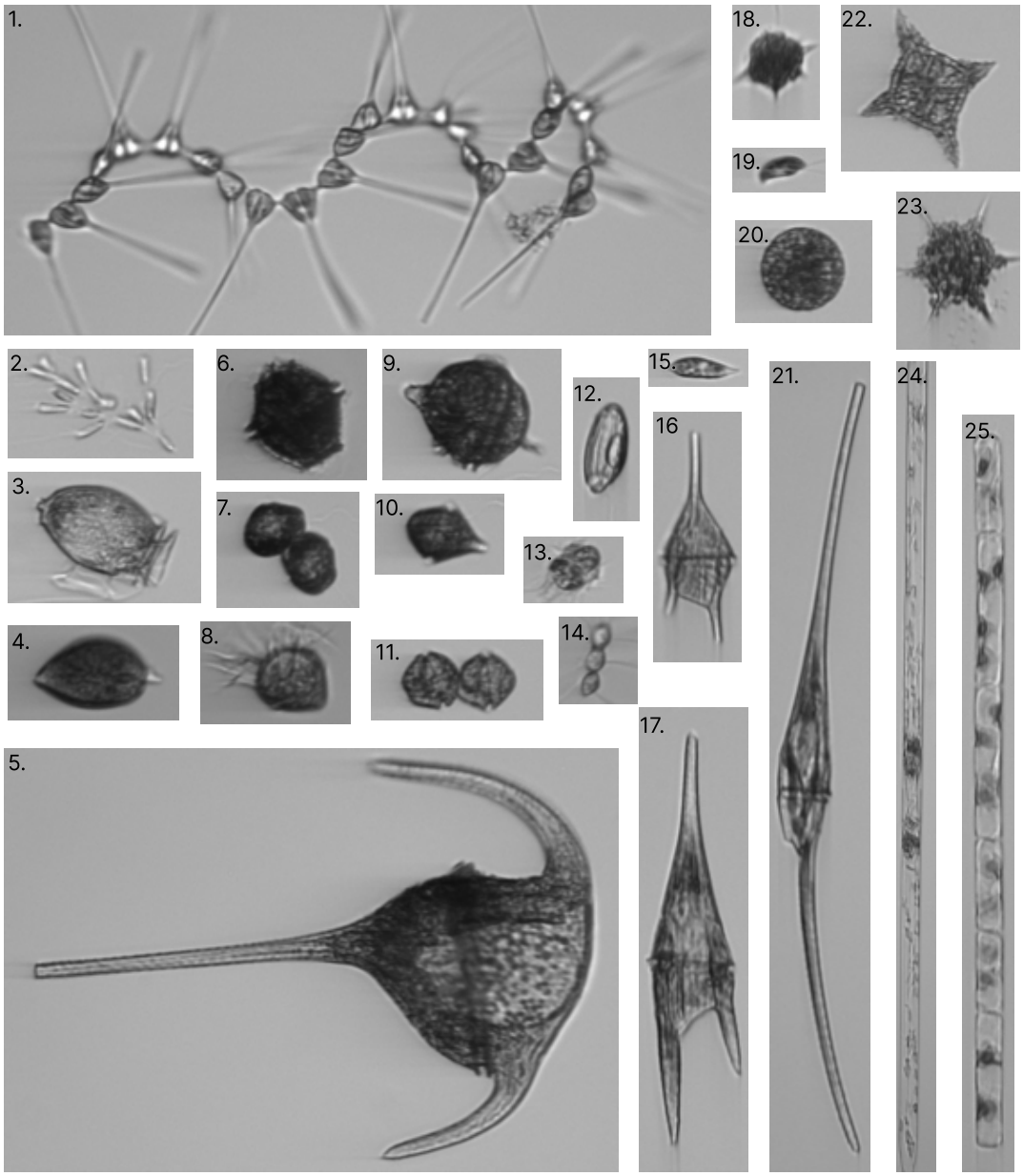


Figure 4: Example images from the SMHI dataset. 1) *Asterionellopsis glacialis*, 2) *Dinobryon* sp., 3) *Dinophysis acuminata*, 4) *Prorocentrum micans*, 5) *Ceratium tripos*, 6) *Lingulodinium polyedrum*, 7) *Scrippsiella* pair, 8) *Strombidium* sp., 9) *Gonyaulax spinifera*, 10) *Scrippsiella CPX*, 11) Double cells, 12) *Torodinium robustum*, 13) *Mesodinium rubrum*, 14) *Chaetoceros*, 15) *Prorocentrum triestinum*, 16) *Ceratium lineatum*, 17) *Ceratium furca*, 18) *Alexandrium pseudogonyaulax*, 19) *Cryptomonads*, 20) *Dictyocha* naked, 21) *Ceratium fusus*, 22) *Dictyocha fibula*, 23) *Dictyocha speculum*, 24) *RhizoPseudosolenia*, 25) *Guinardia delicatula*.

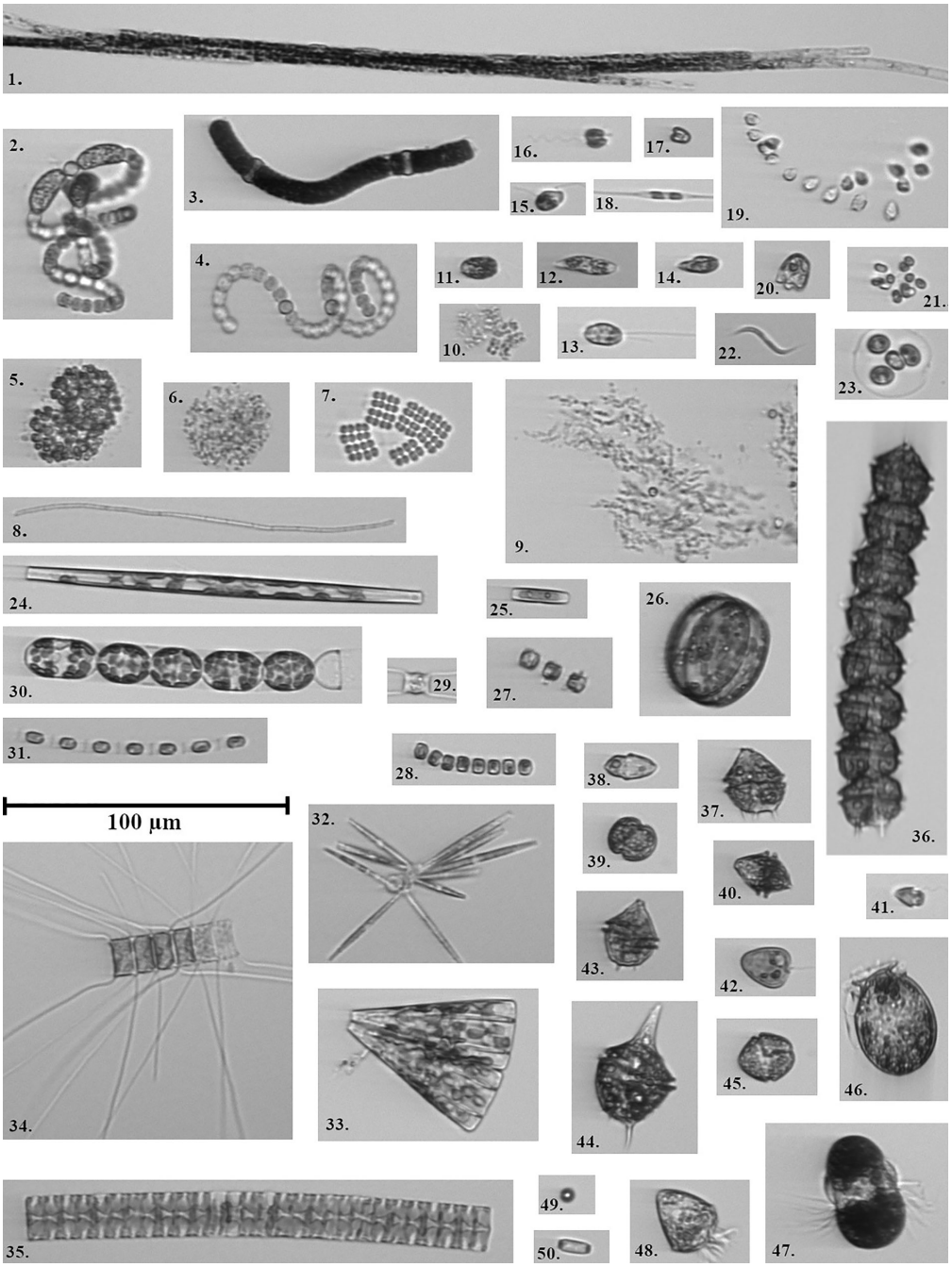


Figure 5: Example images from SYKE-plankton_IFCB_2022. 1) *Aphanizomenon flosaquae*, 2) *Dolichospermum sp./Anabaenopsis* sp. coiled, 3) *Nodularia spumigena*, 4) *Dolichospermum* sp./*Anabaenopsis* sp., 5) *Snowella* sp./*Woronichinia* sp., 6) Chroococcales, 7) *Merismopedia* sp., 8) Oscillatoriales, 9) *Aphanothece paralleliformis*, 10) *Chroococcus* sp., 11) *Eutreptiella* sp., 12) Euglenophyceae, 13) Cryptomonadales, 14) Cryptophyceae-Teleaulax sp., 15) *Katablepharis remigera*, 16) *Pseudopedinella* sp., 17) *Pyramimonas* sp., 18) *Ceratoneis closterium*, 19) *Uroglenopsis* sp., 20) *Cymbomonas tetramitiformis*, 21) Chlorococcales, 22) *Monoraphidium contortum*, 23) *Oocystis* sp., 24) Pennales thin, 25) Pennales thick, 26) Centrales, 27) *Thalassiosira levanderi*, 28) *Cyclotella choctawhatcheeana*, 29) *Chaetoceros* sp. single, 30) *Melosira arctica*, 31) *Skeletonema marinoi*, 32) *Nitzschia paleacea*, 33) *Licmophora* sp., 34) *Chaetoceros* sp., 35) *Pauliella taeniata*, 36) *Peridiniella catenata* chain, 37) *Peridiniella catenata* single, 38) Gymnodiniales, 39) *Gymnodinium* like cells, 40) *Heterocapsa triquetra*, 41) *Heterocapsa rotundata*, 42) *Prorocentrum cordatum*, 43) *Gonyaulax verior*, 44) *Amylax triacantha*, 45) Dinophyceae, 46) *Dinophysis acuminata*, 47) *Mesodinium rubrum*, 48) Ciliata, 49) Beads, 50) Heterocyte. Credit:[18]

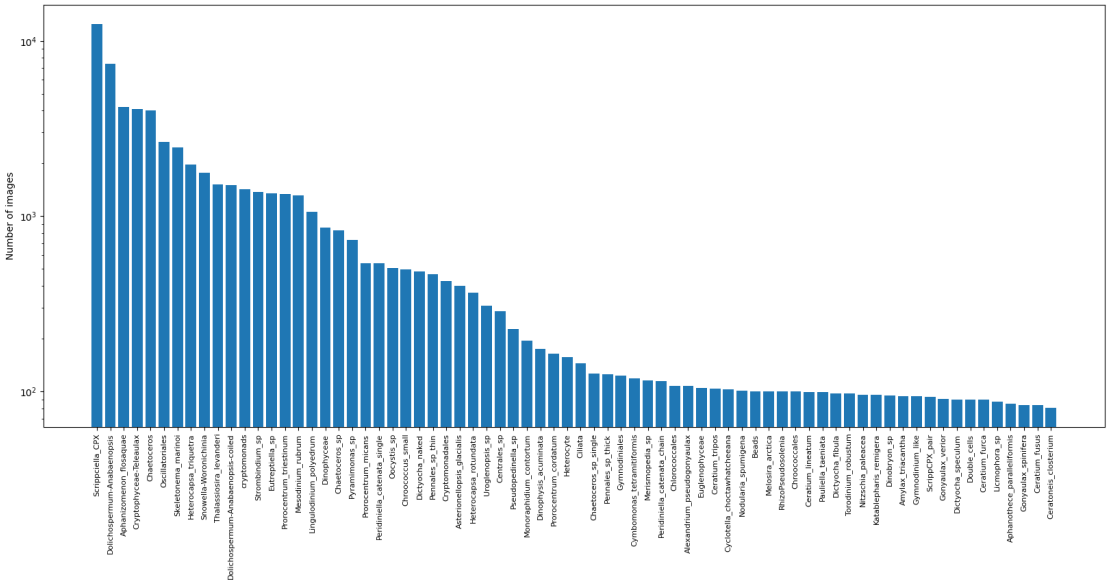


Figure 6: The distribution of images between the image classes in the training set.

The image transform used when importing images to dataloaders includes padding images to a square shape of 180x180 pixels with the mode pixel color of each image (the most frequent pixel value). Three color channels were used to make the IFCB images compatible with the pre-trained neural net. For the training set, the data transform also includes random simple image transforms like random horizontal and vertical flips, translation, zoom, and brightness changes. This introduces variation in the oversampled images and decreases overfitting to the training set.

2.1.3 Construction of Classifier

Model Construction The chosen classifier algorithm is a replication of a classifier described by Kaisa Kraft et al. at SYKE in the paper [18]. In the paper from SYKE, a pre-trained network is adapted to the task of plankton classification using the SYKE-plankton_IFCB_2022 dataset. This project uses a larger dataset, as the SYKE-plankton_IFCB_2022 dataset was merged with the SMHI dataset. All computations using the classifier were performed on Berzelius at National Supercomputer Center, an NVIDIA® SuperPOD consisting of 60 NVIDIA® DGX-A100 compute nodes. [19] A pre-trained ResNet-18 was obtained from TorchVision, which is a part of PyTorch [20]. The pre-training of ResNet-18 is performed with 1000 classes of RGB images from ImageNet. The ImageNet classes are not plankton species, but instead subjects like "fire truck" and "Golden Retriever".

[18]. To adapt the pre-trained model to the plankton classification problem, the fully connected section of the model was replaced with three fully connected linear layers, the last one having as many out-features as the number of plankton image classes. The initial weights in the newly introduced layers are uniformly distributed between $-\sqrt{\mathrm{k}}$ and $\sqrt{\mathrm{k}}$, where $\mathrm{k} = \frac{1}{\mathrm{in}_{\mathrm{features}}}$.

Training During training, a categorical cross-entropy loss (also called softmax loss) was used, and Adam was the optimizer function. A three-step schedule was used to change the learning rates and number of trainable layers during training. In step 1, only the last linear layers were trained and the learning rate 0.01 was used. In step 2, the learning rate for the last linear layers was decreased to 0.005 and training of the last convolutional layer was started with a learning rate of 0.001. In step 3 the learning rate of the last linear layers was further decreased to 0.0025, the learning rate for the last convolutional layer was decreased to 0.001 and the rest of the base layers were trained with a learning rate of 0.0001. The steps lasted for 5, 10, and 15 epochs, respectively. During training the metrics loss, accuracy, weighted F1-score, weighted precision, and weighted recall were monitored using the training and validation dataset. These metrics were weighted by class size as the weighted value is more accurate to the classifier working with natural samples, where some species are more plentiful than others. Training takes less than 30 minutes on one NVIDIA® A100 Tensor Core GPU.

Implementation of Probability Thresholds Not all images can be classified due to lack of information and/or resolution. In the SYKE-plankton_IFCB_2021 dataset, a collection of these images are collected in the class Unclassified. The Unclassified data also includes images that do not belong in any of the image classes currently used by SYKE, so there could be an overlap between the unclassifiable images in the SYKE dataset and the image classes at SMHI. Some of the unclassifiable images from the dataset SYKE-plankton_IFCB_2021 are displayed in Figure 7. The unclassifiable images are used to adapt the model to not make unsure predictions. This is done by implementing probability thresholds.

The probability thresholds used should be different for each class, as some classes are more similar to the unclassifiable images. The thresholds were determined after training the model. Predictions were made on the validation set with added unclassifiable images from SYKE-plankton_IFCB_2021. Thresholds between 1 and 0 were tested, and the threshold which resulted in the highest F1 score for each class was chosen and used in future predictions. If there were multiple maxima, the highest threshold was chosen. The runtime for the threshold selection is less than 10 minutes on one NVIDIA® A100 Tensor Core GPU.

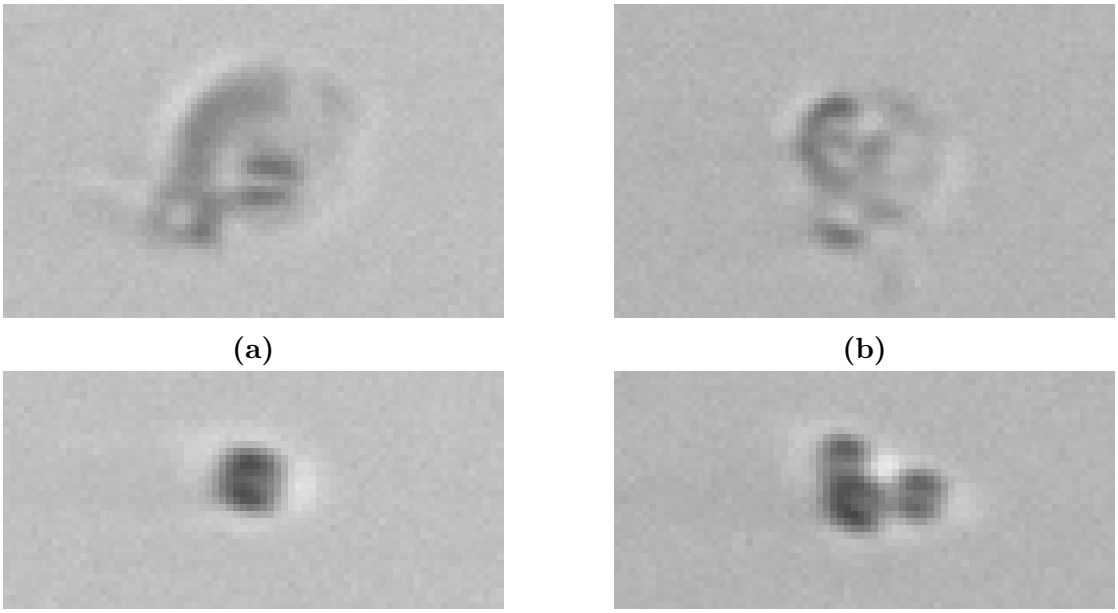


Figure 7: Four examples of images in the Unclassifiable portion of the SYKE-plankton_IFCB_2021 dataset

2.1.4 Evaluation of Classifier Performance

To determine the performance of the classifier, the Test Dataset including unclassifiable images was used. The model was used to predict image classes on all images of the Test Dataset, and the performance metrics precision, recall, and F1 score were calculated class-wise.

To evaluate if the classifier worked equally well independent of the origin of the images, another classifier was developed using a training set only from the SYKE-plankton_IFCB_2022 dataset to be tested on images from SMHI. The classifier was developed according to the method previously described, and thresholds were developed using a validation set from the SYKE-plankton_IFCB_2022 dataset combined with unclassifiable images from the SYKE-plankton_IFCB_2021 dataset. The classifier was applied to all images in the SMHI dataset, and the classification results for images of *Mesodinium rubrum*, *Dinophysis acuminata*, and *Strombidium sp.* were extracted.

2.2 IFCB Data Collection and Application of Classifier

In March 2023, an IFCB was installed on board the research ship R/V Svea and connected to the Ferry Box system. Through the Ferry Box system, water samples

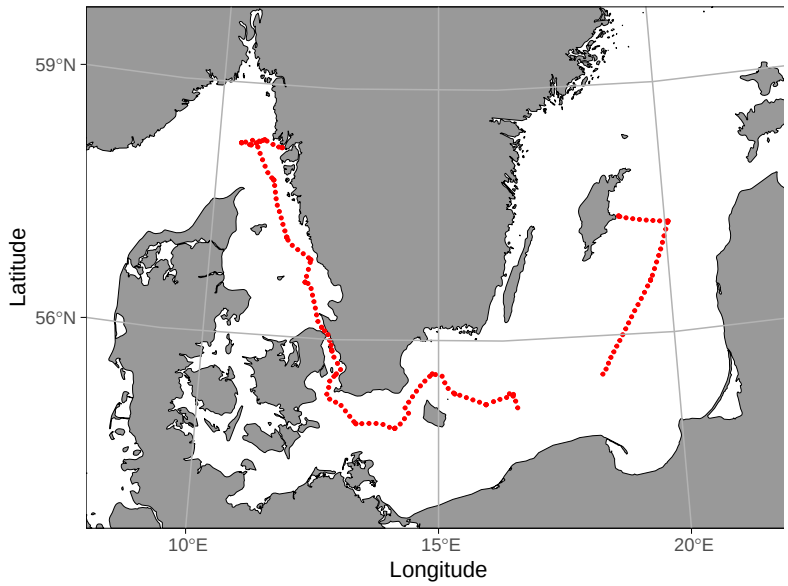


Figure 8: Sampling locations for IFCB. n = 178

(5 ml) were autonomously collected approximately every 27 minutes at a 4-meter depth and analyzed in the IFCB. The samples are analyzed for a set time period, and the remaining volume is ejected if not analyzed within the time period. The research cruise lasted 6 days and the IFCB was active for about 4.5 days due to technical issues and stormy weather. The sample points for the IFCB are presented in Figure 8. The sampling for IFCB starts later than for metabarcoding due to technical issues. The gap in sampling (east of Bornholm) is where the storm was experienced.

The raw data acquired from the IFCB system consisted of .adc, .hdr and .roi-files. With the program IFCB ROI Viewer, the raw outputs were formatted to .png-files which could be uploaded to Berzelius compute cluster. The classifier was applied to the .png images using the probability thresholds.

To compare the IFCB results to the metabarcoding results, the Protist Ribosomal Reference (PR^2) database was used to identify the closest match in PR^2 taxonomy for each image class. For image class names not existing in PR^2 v. 5.0.0, World Register of Marine Species (WoRMS) was used to find synonyms present in PR^2. For all Cyanophyceae, WoRMS was used to annotate them to Cyanophyceae, as PR^2 only covers protists. This matching enables a comparison of the abundance of each image class and its closest taxonomic match.

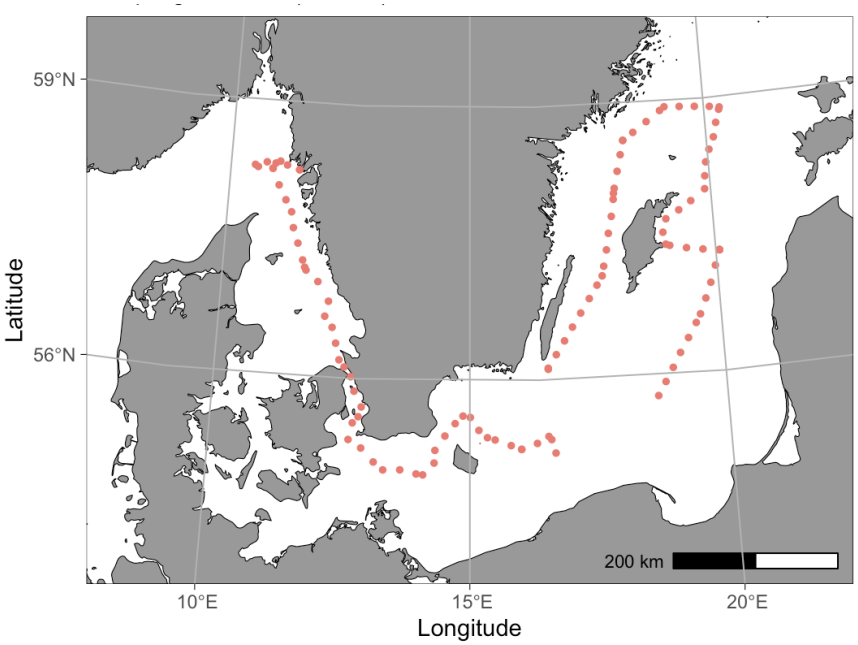


Figure 9: Sampling locations for metabarcoding. n = 107

2.3 Metabarcoding

2.3.1 Sample Collection

On the same cruise with R/V Svea in March 2023, water samples were collected for metabarcoding of planktonic DNA. Water was collected from the Ferry Box system every 60 minutes. These samples are not synchronized with IFCB sampling times but use water from the same source (the Ferry Box system). The sample locations are displayed in Figure 9. The sample volume was around 500 ml. Following the sample acquisition, the water was filtered with MF-Millipore™ Membrane Filters (pore size 0.22 µm). Once dry, the filters were stored at -80 °C in vials.

2.3.2 DNA Extraction

The filters were thawed, and the DNA was extracted according to the protocol "DNA extraction protocol for DNA-metabarcoding of marine phytoplankton using Zymobiomics DNA miniprep kit (Zymo Research; D4300)" [21]. After DNA extraction, DNA was adjusted to equal concentrations (1.25 ng/mL).

2.3.3 18S rRNA Gene Library Preparation and Sequencing

Libraries for 18S rRNA metabarcoding were prepared at the National Genomics Infrastructure (NGI) at SciLifeLab Solna using the primers V4F CCAGCASCYGCGGTAATTCC and V4RB ACTTTCGTTCTTGATYRR, which target the hypervariable V4 region of the 18S rRNA gene. [22] Both the forward and reverse primers were phased to increase the library complexity. The sequences of the phased primers are available in Appendix 6, and they were used in equal proportions. The PCR protocol is described in reference [23]. The primers also have Illumina sequence adapters in their 5' ends (forward: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', reverse: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3') (ordered from IDT DNA (IA, US) at 100 µM in TE buffer). The primers were designed using the Adapterama indexing scheme [24, 25], and include unique forward and reverse indices for all samples sequenced.

The KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems, MA, USA) was used for the PCR reactions. For the first PCR, the reaction mix (25 µL) contained 1x Kapa HiFi HotStart ReadyMix, 5 ng template DNA, and 0.3 µM of each primer. The PCR reaction conditions were 95°C for 3 min, followed by 30 cycles of 98°C for 20 s, 52°C for 15 s and 72°C for 15 s, followed by a final elongation step of 72°C during 2 min. The PCR products were cleaned using magnetic beads with the MagSi-NGS PREP Plus Kit (MDKT00010075, magtivio BV., Nuth, the Netherlands).

The reaction conditions for the second PCR were 95°C for 2 min, 8 cycles of 98°C for 20 s, 55°C for 30 s and 72°C for 30 s, followed by a final elongation step of 72°C for 2 min. Samples were pooled equimolarly and sequenced with NextSeq at SciLifeLab/NGI (Stockholm, Sweden).

2.3.4 Data Processing of Metabarcoding

The sequences delivered from SciLifeLab/NGI were processed with a pipeline for 18S sequencing data. The pipeline is available in a Github repository, `https://github.com/lfdelzam/ASV_dada2_chunck` The pipeline utilizes DADA2 [26] for denoising the sequences into amplicon sequence variants (ASVs), and for conducting taxonomic annotation of the ASVs using the Protist Ribosomal Reference (PR^2) [17]. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

3 Results

3.1 Classifier Performance

3.1.1 Monitoring During Model Training

Training progress was monitored by calculating the accuracy, recall, precision, and F1 score on the training and validation set. Progress during training on the combined SYKE and SMHI datasets is illustrated in Figure A1 and Figure A2 in the Appendix. The metrics in Figure A1 are weighted by class size, which means that if a class is 10x more abundant than the smaller class, it has 10x more importance to the metrics than thesmaller class. In Figure A2, all classes are weighted equally. The precision, recall, and F1 score weighted by class abundance is higher than the metrics from equal class weighting. In the first epoch, the F1 scores are 0.4 and 0.8 on the validation set, respectively.

Where the metrics are weighted with class abundance, the precision, recall, and F1 score are consistently higher on the Validation set than on the Training set. This is unexpected since the model is fit to the data in the training set. However, it is explained by the oversampling of rare classes in the training set (which was not performed in the validation set). This oversampling increases the abundance of rare classes, and therefore rare classes are given more importance in the calculation of the averages for the training than the validation set in Figure A1. In the graphs where all classes are weighted equally (Figure A2), the performance is better on the training set.

3.1.2 Test of Cross-Dataset Performance

A subset of the results from the classifier developed only with SYKE images on the SMHI images are presented in Tables A1, A2, and A3. *Dinophysis acumniata* and *Mesodinium rubrum* was selected for evaluation as these taxa are image classes in both datasets. A perfect model performance would be to predict all images from the SMHI image class to the corresponding image class. From Table A1, it can be observed that 64% of the *Dinophysis acuminata* images were correctly classified to *Dinophysis acuminata*, and 32% were unclassified. For *Mesodinium rubrum* 24% were correctly classified as *Mesodinium rubrum*, and 47% of the images were unclassified. Additionally, 16% of the images were classified as Ciliata, which is also correct as *Mesodinium rubrum* is in the subdivision Ciliophora (PR^2 database v. 5.0.0). [17] In the case of *Strombidium* sp., it does not have a directly corresponding class in the SYKE-plankton_IFCB_2022 dataset. The largest portion of images (59%) is annotated to Ciliata, which is correct as *Strombidium* sp. is a ciliate, in the subdivision Ciliophora (PR^2 database v. 5.0.0). [17] 33% of *Strombidium* sp. were unclassified.

3.1.3 Performance Metrics on Test Set and Thresholds

In Table 1 the precision, recall, and F1 score is presented for all image classes, calculated on the test dataset including the unclassified images. The thresholds for classification are also presented. In general, the classes have high F1 scores with 70% of the classes having an F1 score over 0.8. For the class *Scrippsiella* pair, the threshold is 1, and the F1 score is 0. This means that all thresholds tested were equal or worse than a threshold of 1.

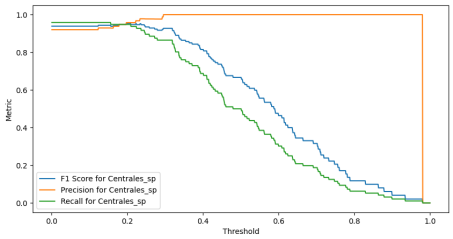
Table 1: Table containing the class-specific evaluation metrics for the image classifier. (N: number of images, Pr: precision, Re: recall, F1: F1 score)

Image class	Training set	Validation set		Test set			
	N	N	Threshold	N	Pr	Re	F1
Alexandrium pseudogonyaulax	54	18	0.30	18	0.62	0.72	0.67
Amylax triacantha	11	3	0.64	5	1.00	0.67	0.80
Aphanizomenon flosaquae	4193	1397	0.10	1399	0.99	0.98	0.98
Aphanothece paralleliformis	17	5	0.37	7	0.80	0.80	0.80
Asterionellopsis glacialis	402	134	0.31	135	1.00	0.69	0.82
Beads	75	25	0.63	25	0.74	0.68	0.71
Centrales sp	288	96	0.21	96	0.96	0.95	0.95
Ceratium furca	9	3	0.62	3	1.00	1.00	1.00
Ceratium fusus	4	1	0.86	2	1.00	1.00	1.00
Ceratium lineatum	73	24	0.43	25	1.00	1.00	1.00
Ceratium tripos	4	1	0.99	3	1.00	1.00	1.00
Ceratoneis closterium	27	9	0.59	9	0.73	0.89	0.80
Chaetoceros	4002	1334	0.24	1334	0.97	0.99	0.98
Chaetoceros sp.	276	0.9	0.52	277	0.89	0.91	0.90
Chaetoceros sp. single	42	0.94	0.30	44	0.97	0.90	0.94
Chlorococcales	57	19	0.27	19	0.50	0.42	0.46
Chroococcales	71	28	0.25	29	0.82	1.00	0.90
Chroococcus small	165	0.97	0.33	166	0.98	0.96	0.97
Ciliata	145	48	0.29	50	0.97	0.79	0.87
Cryptomonadales	427	142	0.36	144	0.72	0.80	0.76
Cryptomonads	1421	473	0.23	475	0.93	0.92	0.92
Cryptophyceae-*Teleaulax*	4098	1366	0.61	1366	0.92	0.85	0.88
Cyclotella choctawhatcheeana	61	20	0.35	21	0.68	0.65	0.67
Cymbomonas tetramitiformis	119	39	0.14	41	0.95	0.90	0.92
Dictyocha fibula	7	2	0.69	3	1.00	1.00	1.00
Dictyocha naked	484	161	0.33	162	1.00	0.97	0.98
Dictyocha speculum	30	10	0.77	10	1.00	1.00	1.00
Dinobryon sp.	44	14	0.66	16	1.00	0.93	0.96
Dinophyceae	859	286	0.23	288	0.92	0.93	0.92
Dinophysis acuminata	175	58	0.41	59	0.90	0.98	0.94
Dolichospermum-Anabaenopsis	7368	2456	0.40	2456	0.98	0.99	0.98
Dolichospermum-Anabaenopsis-coiled	1502	500	0.50	502	0.99	0.92	0.95
Double cells	30	10	0.35	11	0.83	0.50	0.62

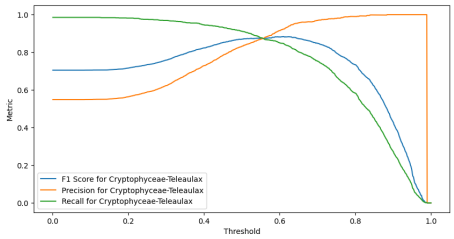
Euglenophyceae	61	20	0.44	21	1.00	0.80	0.89
Eutreptiella sp.	1348	449	0.32	450	0.93	0.91	0.92
Gonyaulax spinifera	6	2	0.42	3	0.50	0.50	0.50
Gonyaulax verior	13	4	0.65	5	1.00	0.50	0.67
Gymnodiniales	41	13	0.25	15	0.92	0.85	0.88
Gymnodinium like	94	31	0.21	33	0.40	0.45	0.42
Heterocapsa rotundata	368	122	0.45	124	0.76	0.52	0.61
Heterocapsa triquetra	1965	655	0.22	656	0.95	0.96	0.96
Heterocyte	157	52	0.59	54	0.83	0.83	0.83
Katablepharis remigera	32	10	0.17	12	0.35	0.60	0.44
Licmophora sp.	44	14	0.29	16	0.92	0.86	0.89
Lingulodinium polyedrum	1064	354	0.21	356	0.99	0.94	0.97
Melosira arctica	25	8	0.32	10	0.89	1.00	0.94
Merismopedia sp.	58	19	0.45	21	0.86	0.63	0.73
Mesodinium rubrum	1312	437	0.32	439	0.93	0.94	0.93
Monoraphidium contortum	196	65	0.69	66	1.00	0.98	0.99
Nitzschia paleacea	39	13	0.45	13	1.00	0.85	0.92
Nodularia spumigena	101	33	0.23	35	1.00	0.88	0.94
Oocystis sp.	505	168	0.40	169	0.90	0.90	0.90
Oscillatoriales	2664	888	0.46	888	1.00	1.00	1.00
Pauliella taeniata	71	23	0.36	25	1.00	1.00	1.00
Pennales sp. thick	126	42	0.41	42	0.76	0.81	0.78
Pennales sp. thin	468	156	0.19	157	0.98	0.75	0.85
Peridiniella catenata chain	115	38	0.49	40	0.94	0.84	0.89
Peridiniella catenata single	539	179	0.24	181	0.89	0.85	0.87
Prorocentrum cordatum	165	55	0.36	56	0.88	0.84	0.86
Prorocentrum micans	541	180	0.42	181	0.97	0.99	0.98
Prorocentrum triestinum	1339	446	0.23	447	0.91	0.92	0.91
Pseudopedinella sp.	227	75	0.69	77	0.48	0.45	0.47
Pyramimonas sp.	734	244	0.56	246	0.29	0.27	0.28
RhizoPseudosolenia	80	26	0.33	28	1.00	0.88	0.94
Scrippsiella pair	3	1	1.00	1	0.00	0.00	0.00
Scrippsiella CPX	12496	4165	1	4166	1.00	1.00	1.00
Skeletonema marinoi	2476	825	0.62	827	0.98	0.98	0.98
Snowella-Woronichinia	1770	590	0.21	590	0.97	0.97	0.97
Strombidium sp.	1372	457	0.19	458	0.99	0.98	0.98
Thalassiosira levanderi	1522	507	0.42	508	0.88	0.89	0.88
Torodinium robustum	49	16	0.79	18	0.80	1.00	0.89
Uroglenopsis sp.	309	103	0.07	104	0.88	0.04	0.07
Sum of classified images	62004	20650		20738			
N of Unclassifiable images		20644		20738			

3.1.4 Threshold Determination

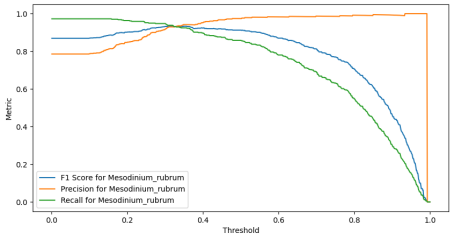
The optimal probability thresholds are presented in Table 1. To give more insight into the threshold determination process, the precision, recall, and F1 score was plotted over threshold size for four image classes in Figure 10. The F1 scores



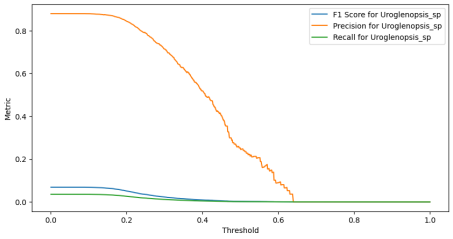
(a) *Centrales* sp. Threshold selected: 0.21



(b) Cryptophyceae-Teleaulax sp. Threshold selected: 0.61



(c) *Mesodinium rubrum*. Threshold selected: 0.32



(d) *Uroglenopsis* sp. Threshold selected: 0.07

Figure 10: Effect on classifier performance metrics from probability threshold for four image classes.

are computed with the validation dataset with added unclassifiable images from SYKE-plankton_IFCB_2021. It can be observed that the precision is increasing with threshold size, while the recall decreases for *Centrales* sp., Cryptophyceae-Teleaulax sp., and *Mesodinium rubrum*. However, for *Uroglenopsis* sp., precision unexpectedly decreases with increasing threshold size. In each run of the code, the training results in different weights in the network and, as a result, in different thresholds (data not shown).

3.2 Classification of IFCB Images from R/V Svea

Figure 11 illustrates the classification of images from R/V Svea. It can be observed that a majority of images were in the Unclassified class. For the plot, only the classes which had relative abundance larger than 0.03 in the images from one sample were included. These were *Uroglenopsis* sp., Cryptophyceae-Teleaulax, *Thalassiosira levanderi*, Cryptomonadales, *Heterocapsa rotundata*, *Mesodinium rubrum*, and *Skeletonema marinoi*. These classes are all from the SYKE dataset, except *Mesodinium rubrum* which is present in both datasets, and *Licmophora* sp. which is from SMHI. It can be observed that there were fewer images captured per sample in Öresund compared to the Baltic Proper and Kattegat-Skagerrak area.

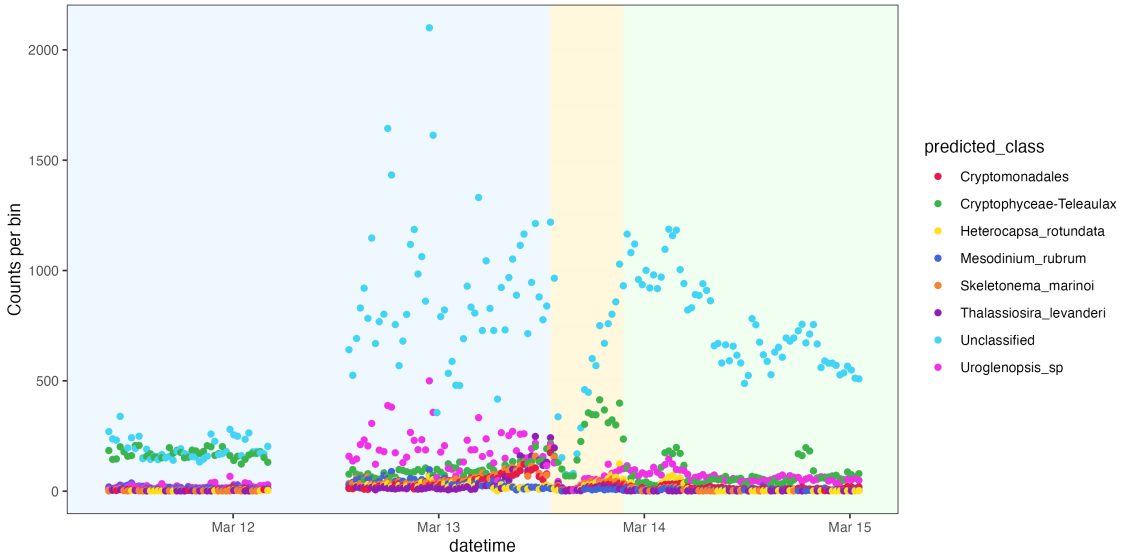


Figure 11: Graphs displaying the classification of IFCB images from the cruise on R/V Svea (by image class). The colored regions indicated (from left to right) Baltic Proper, Öresund, and Kattegat-Skagerrak

3.3 Metabarcoding

A total of 22 549 amplicon sequence variants (ASV:s) were found, and after the exclusion of Macroalgae and Metazoa 18 992 ASV:s remained. The number of reads per sample differed between 185 981 and 577 037 after the exclusion of Macroalgae and Metazoa. 104 samples were sequenced successfully.

3.4 Correlation of Metabarcoding and IFCB

The total number of images captured per sample and the DNA concentration after extraction are illustrated in Figure 12. It can be seen that the cell counts vary drastically during the cruise and that it roughly covaries with DNA concentration.

In Figure 13 the abundance of some image classes is illustrated together with the metabarcoding-based counts. The metabarcoding-based counts are calculated by multiplying the DNA concentration in the sample with the relative abundance of the taxa. The presented image classes follow the following criteria:

- Are detectable with 18S barcoding (does not include Cyanobacteria)
- Are detected in more than 3 IFCB samples
- Have a correlation coefficient R>0.8

For image classes with the species' epithet being "sp.", they were matched with the corresponding genus in the metabarcoding data. For example, the image class *Licmophora* sp. was plotted together with the taxon *Licmophora* as it was uncertain which species of *Licmophora* SMHI and SYKE referred to.

It can be observed that *Thalassiosira levanderi*, *Chaetoceros* sp., *Chateoceros* sp. single, and *Skeletonema marinoi* are the detected taxa with the strongest correlation between barcoding and IFCB classification. They are diatoms with very recognizable features (see Figure 5 and 4). However, there are more taxa with clearly recognizable features in the datasets. The other top detected classes in IFCB were *Uroglenopsis* sp., Cryptophyceae-Teleaulax, Cryptomonadales, *Heterocapsa rotundata*, *Mesodinium rubrum*. Of these, *Mesodinium rubrum* is the largest and most recognizable to humans. However, it is known to have poor coverage in PR^2. In the manual inspection of random images classified as *Mesodinium rubrum*, it is clear that the classification was correct. For *Licmophora* sp. on the other hand, the images collected in the Öresund strait are misclassified. The images depict diatoms of the genus *Guinardia*, which to the human eye look very different from *Licmophora* sp. In fact, the correlation between *Licmophora* sp. from image classification and ASV32, annotated to *Guinardia delicatula*, has R=0.94.

The other classes detected with IFCB but not present in barcoding are small and not as recognizable (see Figure 5). *Uroglenopsis* sp. has the second lowest F1 score: 0.07 with recall 0.04 and precision 0.88 (Table 1. As it is the recall and not precision that is impaired, it is unexpected to observe a high classification rate but low detection in barcoding.

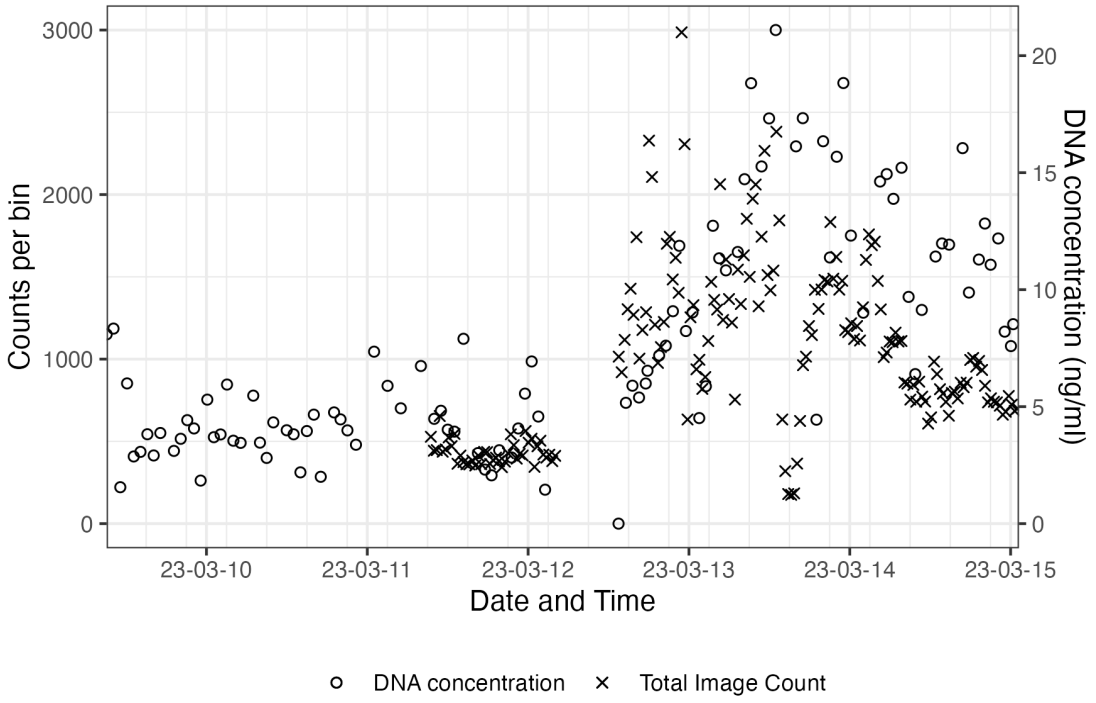
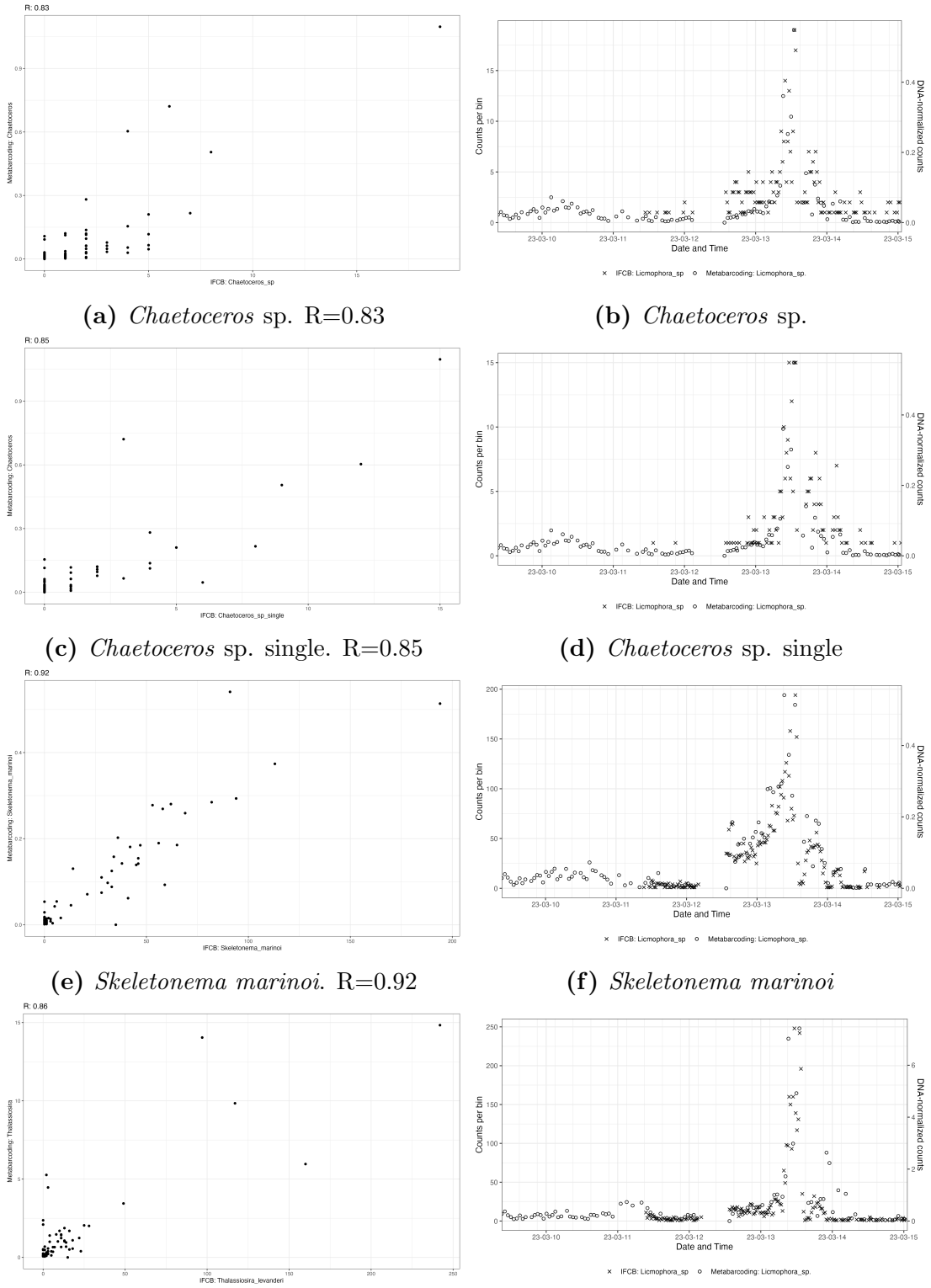


Figure 12: The number of images captured per bin and DNA concentration after extraction from the samples collected on R/V Svea.



(a) *Chaetoceros* sp. R=0.83

(b) *Chaetoceros* sp.

(c) *Chaetoceros* sp. single. R=0.85

(d) *Chaetoceros* sp. single

(e) *Skeletonema marinoi*. R=0.92

(f) *Skeletonema marinoi*

(g) *Thalassiosira levanderi*. R=0.86

(h) *Thalassiosira levanderi*

Figure 13: IFCB classification and metabarcoding results from the Svea cruise. The left graphs display the correlation between the barcoding and IFCB. The right graphs display time series. For IFCB the value plotted is counts per bin, and for barcoding the value plotted is the relative abundance corrected by sample volume and DNA concentration after extraction of DNA.

4 Discussion

4.1 Classifier Performance

The classifier's predictions on *Mesodinium rubrum*, *Dinophysis acuminata Strombidium* sp. had significantly higher recall in the classifier trained on both datasets than in the classifier trained on SYKE data but evaluated on SMHI images. This could be due to differences in instruments or instrument settings, or morphological differences in the different seas. From conversations with Ann-Turi Skjevik at the Oceanography unit at SMHI, I learned that taxonomists can see significant differences for the same taxa between the Baltic and Skagerrak/Kattegat. The morphological variation of *Mesodinium rubrum* is also mentioned in the literature, for example in the abundance and placement of cirri. [27] It points towards the importance of including training data from multiple locations if classifiers are to be used in larger areas. Even though the classifier trained on SYKE-plankton_IFCB_2022 had a weak performance on Tångesund data, the combined training led to good performance on the combined dataset (F1=0.93 for *Mesodinium rubrum* and F1=0.94 for *Dinophysis acuminata*). The CNN can thus learn to recognize the diverse morphologies if it is exposed to them during training.

One issue that could impact the performance of SMHI classes is that SMHI classes could be frequent in the Unclassified images from SYKE-plankton_IFCB_2021. They were deemed Unclassified as the images were not classified to the classes in SYKE-plankton_IFCB_2022. An overlap in taxa between the unclassified images and the classes would disrupt the threshold determination process significantly, as the precision would appear lower in the metrics than it actually is. I believe that threshold determination when datasets are combined is an important challenge. Since class-specific thresholds are needed to optimize the performance of each class, the comparison with images that do not belong to any of the classes is crucial. However, Kaisa Kraft, one of the creators of the dataset, states that if some of the unclassified images could be annotated to a taxon not present in the image classes, that class would typically be created. Therefore, the unclassified image mostly consists of images that are not classifiable by humans. However, I believe that the misclassification of images to *Licmophora* sp. could have been avoided if the Unclassifiable set included *Guinardia*, or if *Guinardia* was a class in the classifier. Inclusion of *Guinardia* as a separate class would allow the classifier to learn differentiation between them during training, and inclusion in the Unclassifiable set would adapt the probability thresholds to avoid misclassification of *Licmophora* sp. to *Guinardia*.

In general, the F1 scores of the classes on the testing dataset are high, but the cross-dataset compatibility test shows that the performance in other settings (instrument, season, sea, etc) is not as good as the performance in testing condi-

tions on the same dataset as the model was trained on. To evaluate performance in settings more similar to the sampling on R/V Svea, a dataset more similar to natural samples from R/V Svea would be beneficial. For the SYKE classes, the dataset SYKE-plankton_IFCB_2021 could be used, but to my knowledge, there is no other annotated dataset of IFCB images covering the SMHI classes. This type of testing dataset could be developed by selecting random IFCB samples gathered on R/V Svea and manually annotating the images.

By F1 score, the classes with the poorest performance are *Chlorococcales* (0.46), *Gonyaulax spinifera* (0.50), *Gymnodinium* like (0.42), *Katablepharis remigera* (0.44), *Pseudopedinella* sp. (0.47), *Pyramimonas* sp. (0.28), ScrippCPX pair (0), and *Uroglenopsis* sp. (0.07). These classes are all from the SYKE-plankton_IFCB_2022 dataset except Scrippsiella and *Gonyaulax spinifera*. For Scrippsiella pair, the issue is that the total number of images is four, and none were assigned to the validation set. This means that no threshold could be determined, and the default of 1 was used. In comparison to the other classes, the poorly performing ones have low to medium sizes amounts of training data (57, 6, 94, 32, 227, 734, and 309 images, respectively). It can also be noted that the poorly performing classes have images with smaller dimensions than average. *Gonyaulax spinifera* is around average image size but has an unusually small training set (6 images).

The gradual decrease in precision observed in Figure 10 for *Uroglenopsis* sp. is not observed in any other classes. Most class precision curves end with a drop from 1 to 0, as observed in the other three subfigures in Figure 10. This is since, at high thresholds, only the most normative images of the class have probabilities over the threshold, and the prediction is very selective. However, in the case of *Uroglenopsis* sp., the fraction of correct images in the images assigned to the class decreases with threshold size. The reason for this remains to be investigated.

During training, incorrect classifications are punished by the loss function. For the classifier developed in this project, a correct classification was defined as the classifier assigning the same class as the image was assigned to. However, since the classes are at varying taxonomic levels, an image can belong to different image classes. For example, assigning *Strombidium* sp. to Ciliata shouldn't be punished in the loss function as *Strombidium* is a genus of ciliates. However, a classifier assigning all *Strombidium* images to Ciliata has less utility, so classification to a more detailed taxonomic level should be encouraged. The first step to designing classifiers with this feature is using image classes that align well with taxonomy, like connecting each class to a WoRMS ID. The problem with overlapping image classes also affects measures of classifier performance, as the performance measures currently do not take class overlap into account.

4.2 Correlation between Classifier and Metabarcoding

Figure 13 shows that the correlation between the image classes and their corresponding taxa is strong for some image classes. However, the figure only displays a subset selected to have strong correlation. Many of the most abundant image classes on the cruise (Figure 11) were not detected in enough barcoding samples to compute a correlation coefficient. These discrepancies between barcoding and IFCB are good starting points for the improvement of these techniques. For example, that *Guinardia* is classified as *Licmophora* sp. highlights a weakness in the classification. The fact that *Mesodinium rubrum* is detected with IFCB but not barcoding points towards a weakness in the coverage of PR^2 or in the primers used.

It is also surprising that no *Ceratium* were detected with IFCB, as the genus is very recognizable (see Figure 4. The species under the *Ceratium* genus in this report belong in the *Tripos* genus in WoRMS and PR^2. DNA from the *Tripos* genus was detected in 27 samples. I suspect this is related to the small sample sizes. 8 samples with *Ceratium fusus* (*Tripos fusus*) and 27 samples with *Ceratium furca* (*Tripos furca*). No images were classified to these classes. I suspect it is due to the small training sets (4 and 9 images, respectively). It can be noted that the F1 scores are 1 for both of these classes when using the test set, but the real-sample performance was not aligned with the barcoding results.

4.3 Conclusion

The classifier developed performed well in most classes on the test set. However, this is not similar to real-world conditions. The cross-dataset performance test showed that the performance was significantly worse when training on one dataset and testing on the other. This drop in performance can mainly be attributed to an increase in the number of Unclassified images.

The parallel collection of samples with IFCB and metabarcoding has highlighted weaknesses in both techniques. There is a strong correlation for some of the taxa (*Chaetoceros*, *Skeletonema marinoi*, *Thalassiosira levanderi*), and discrepancies in the techniques for other taxa (*Licmophora*, *Mesodinium rubrum*, *Ceratium*). The sampling was performed during a single cruise, and many taxa in the image classes were undetected. For these taxa, the correlation between metabarcoding and IFCB classification is unknown. The strong correlations to IFCB cell counts for some image classes display how metabarcoding can be used as a quantitative measure, not only qualitative.

5 Future Perspectives

For a full comparison of the two methods used in this report, it would be preferable to detect a wider range of image classes in the IFCB. The sampling was limited to one cruise, which took place in March. Many taxa were not detected with either method and can therefore not be evaluated. Also, the analysis of the barcoding sequences was limited by the timeframe of the project, as they were delivered in the last month of the project. However, the dataset collected will be used in future projects to develop a deeper knowledge of DNA metabarcoding and IFCB image classification.

6 Acknowledgements

I want to thank my supervisor Anders Andersson for letting me join this exciting project and for always being available to support me. Krzysztof Jurdzinski for giving great advice and taking the night shifts on R/V Svea. Bengt Karlson and the Oceanography group at SMHI for answering all my curious questions. Fernando Delgado, Emma Bell, and everyone in the PhD lab on Alfa 3 for making this semester more fun & full of opportunities to learn. The crew at National Genomics Infrastructure (NGI) for the swift generation of 18S metabarcoding libraries. I'd also like to thank Josephine Sullivan and Minchong Li for inspiration in the classifier design. The CNN computations were enabled by resources provided by the National Academic Infrastructure in Sweden (NAISS) at Berzelius.

References

[1] Swedish Agency for Marine and Water Management. *Växtplankton, bakterieplankton, primärproduktion och blomning*. 2019. URL: https://www.havochvatten.se/overvakning-och-uppfoljning/miljoovervakning/marin-miljoovervakning/vaxtplankton-bakterieplankton-primarproduktion-och-blomning.html. (accessed: 24.01.2023).

[2] Swedish Meteorological and Hydrological Institute. *Alger*. URL: https://www.smhi.se/data/oceanografi/algsituationen. (accessed: 25.01.2023).

[3] Heidi M. Sosik and Robert J. Olson. "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry". eng. In: *Limnology and oceanography, methods* 5.6 (2007), pp. 204–216. ISSN: 1541-5856.

[4] Lisa Campbell et al. "Continuous automated imaging-in-flow cytometry for detection and early warning of *Karenia brevis* blooms in the Gulf of Mexico". In: *Environmental Science and Pollution Research* 20 (2013), pp. 6896–6902.

[5] Schuyler C Nardelli, Patrick C Gray, and Oscar Schofield. "Developing a convolutional neural network to classify phytoplankton images collected with an Imaging FlowCytobot along the West Antarctic Peninsula". In: *OCEANS 2021: San Diego–Porto*. IEEE. 2021, pp. 1–7.

[6] Johan Kronsell. *Rapport från SMHIs utsjöexpedition med R/V Svea*. June 2022. URL: https://www.smhi.se/polopoly_fs/1.184491!/exp_202122.pdf. (accessed: 26.01.2023).

[7] Pablo González et al. "Automatic plankton quantification using deep features". eng. In: *Journal of plankton research* 41.4 (2019), pp. 449–463. ISSN: 0142-7873.

[8] *Deep Convolutional Neural Networks*. URL: https://www.run.ai/guides/deep-learning-for-computer-vision/deep-convolutional-neural-networks. (accessed: 04.05.2023).

[9] Mathworks. *What Are Convolutional Neural Networks? — Introduction to Deep Learning*. URL: https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html. (accessed: 05.05.2023).

[10] David Stutz. *Illustrating (Convolutional) Neural Networks in LaTeX with TikZ*. URL: https://davidstutz.de/illustrating-convolutional-neural-networks-in-latex-with-tikz/. (accessed: 05.05.2023).

[11] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[12] Peter Flach and Meelis Kull. "Precision-recall-gain curves: PR analysis done right". In: *Advances in neural information processing systems* 28 (2015).

[13] Zacchaeus G Compson et al. "Metabarcoding from microbes to mammals: comprehensive bioassessment on a global scale". In: *Frontiers in Ecology and Evolution* 8 (2020), p. 581835.

[14] Yue OO Hu et al. "Diversity of pico-to mesoplankton along the 2000 km salinity gradient of the Baltic Sea". In: *Frontiers in Microbiology* 7 (2016), p. 679.

[15] Agneta Andersson. *DNA-metabarcoding of marine phytoplankton*. URL: https://www.umu.se/en/research/projects/dna-metabarcoding-of-marine-phytoplankton/. (accessed: 31.01.2023).

[16] I. Puillat et al. *Progress Report #2*. Joint European Research Infrastructure network for Coastal Observatory – Novel European eXpertise for coastal observaTories - JERICO-NEXT, 2016. URL: https://archimer.ifremer.fr/doc/00406/51702/52267.pdf.

[17] Laure Guillou et al. "The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy". In: *Nucleic acids research* 41.D1 (2012), pp. D597–D604.

[18] Kaisa Kraft et al. "Towards operational phytoplankton recognition with automated high-throughput imaging, near-real-time data processing, and convolutional neural networks". eng. In: *Frontiers in Marine Science* 9 (2022). ISSN: 2296-7745.

[19] National Supercomputer Centre. *Berzelius*. URL: https://www.nsc.liu.se/systems/berzelius/. (accessed: 10.05.2023).

[20] *resnet18*. URL: https://pytorch.org/vision/stable/models/generated/torchvision.models.quantization.resnet18.html#torchvision.models.quantization.ResNet18_QuantizedWeights. (accessed: 26.04.2023).

[21] Agneta Andersson et al. *DNA extraction protocol for DNA-metabarcoding of marine phytoplankton using Zymobiomics DNA miniprep kit (Zymo Research; D4300)*. URL: https://www.protocols.io/view/dna-extraction-protocol-for-dna-metabarcoding-of-m-n92ld9q27g5b/v1. (accessed: 12.04.2023).

[22] Sergio Balzano, Elsa Abs, and Sophie C Leterme. "Protist diversity along a salinity gradient in a coastal lagoon". In: *Aquatic Microbial Ecology* 74.3 (2015), pp. 263–277.

[23] Meike AC Latz et al. “Short-and long-read metabarcoding of the eukaryotic rRNA operon: evaluation of primers and comparison to shotgun metagenomics sequencing”. In: *Molecular Ecology Resources* 22.6 (2022), pp. 2304–2318.

[24] Travis C Glenn et al. “Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext)”. In: *PeerJ* 7 (2019), e7755.

[25] Travis C Glenn et al. “Adapterama II: universal amplicon sequencing on Illumina platforms (TaggiMatrix)”. In: *PeerJ* 7 (2019), e7786.

[26] Benjamin J Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7 (2016), pp. 581–583.

[27] David W Crawford. “Some observations on morphological variation in the red-water ciliate Mesodinium rubrum”. In: *Journal of the Marine Biological Association of the United Kingdom* 73.4 (1993), pp. 975–978.

Appendix

Primer Sequences

The primers used in 18S amplicon library preparation were introduced by Balzano et al. in [22]. Phasing nucleotides are indicated in red. The sequence downstream of the phasing nucleotides targets the 18S rRNA gene.
Forward primers:
ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCAGCASCYGCGGTAATTCC
ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCAGCASCYGCGGTAATTCC
ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCCAGCASCYGCGGTAATTCC
ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCCAGCASCYGCGGTAATTCC

Reverse primers:
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTTTCGTTCTTGATYRR
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACTTTCGTTCTTGATYRR
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTTTCGTTCTTGATYRR
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACGACTTTCGTTCTTGATYRR
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTACGACTTTCGTTCTTGATYRR
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTACGACTTTCGTTCTTGATYRR

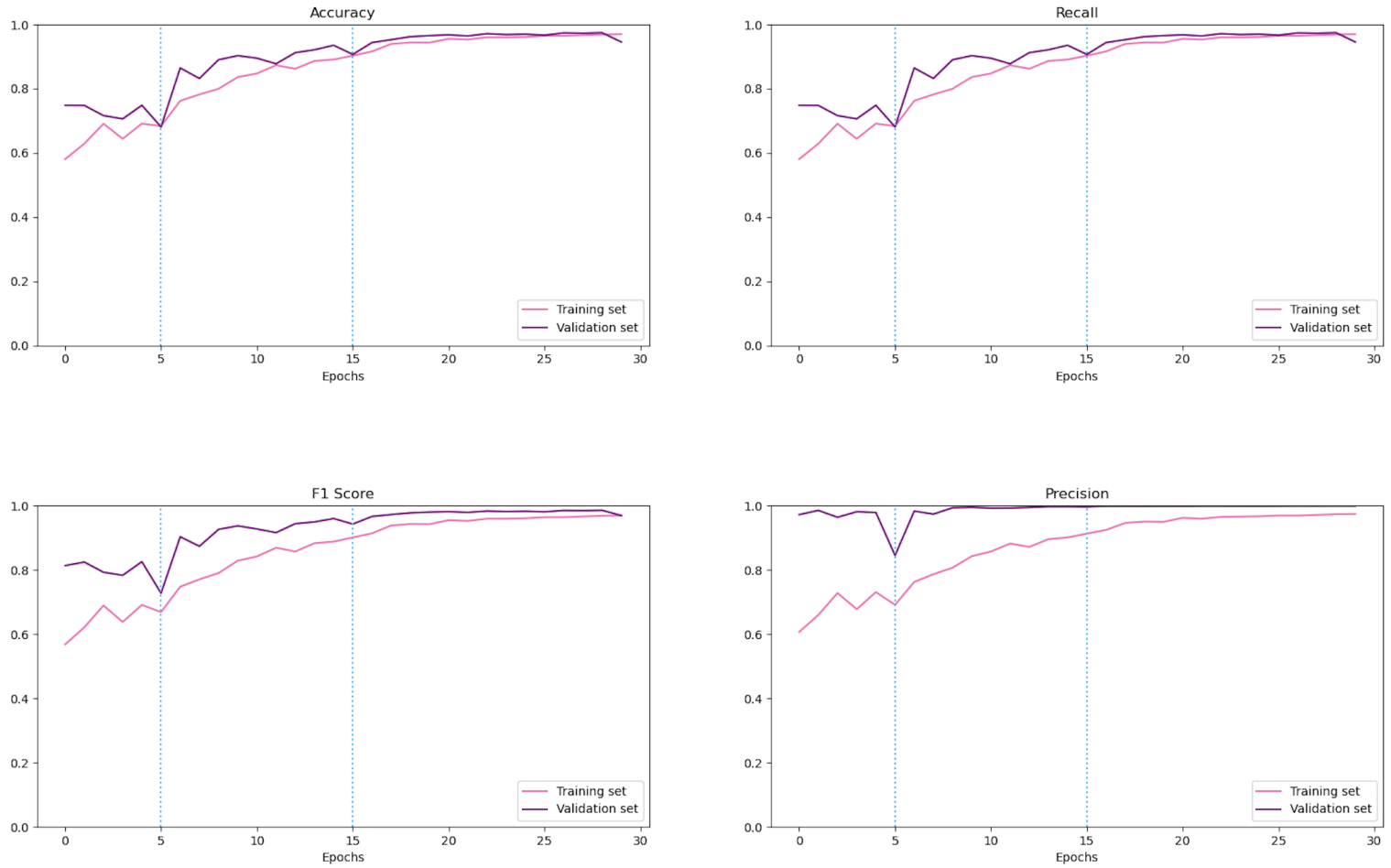


Figure A1: Curves displaying progress during training with the image classes weighted by abundance. The three training phases are indicated with blue lines and begin at Epoch 0, 5, and 15.

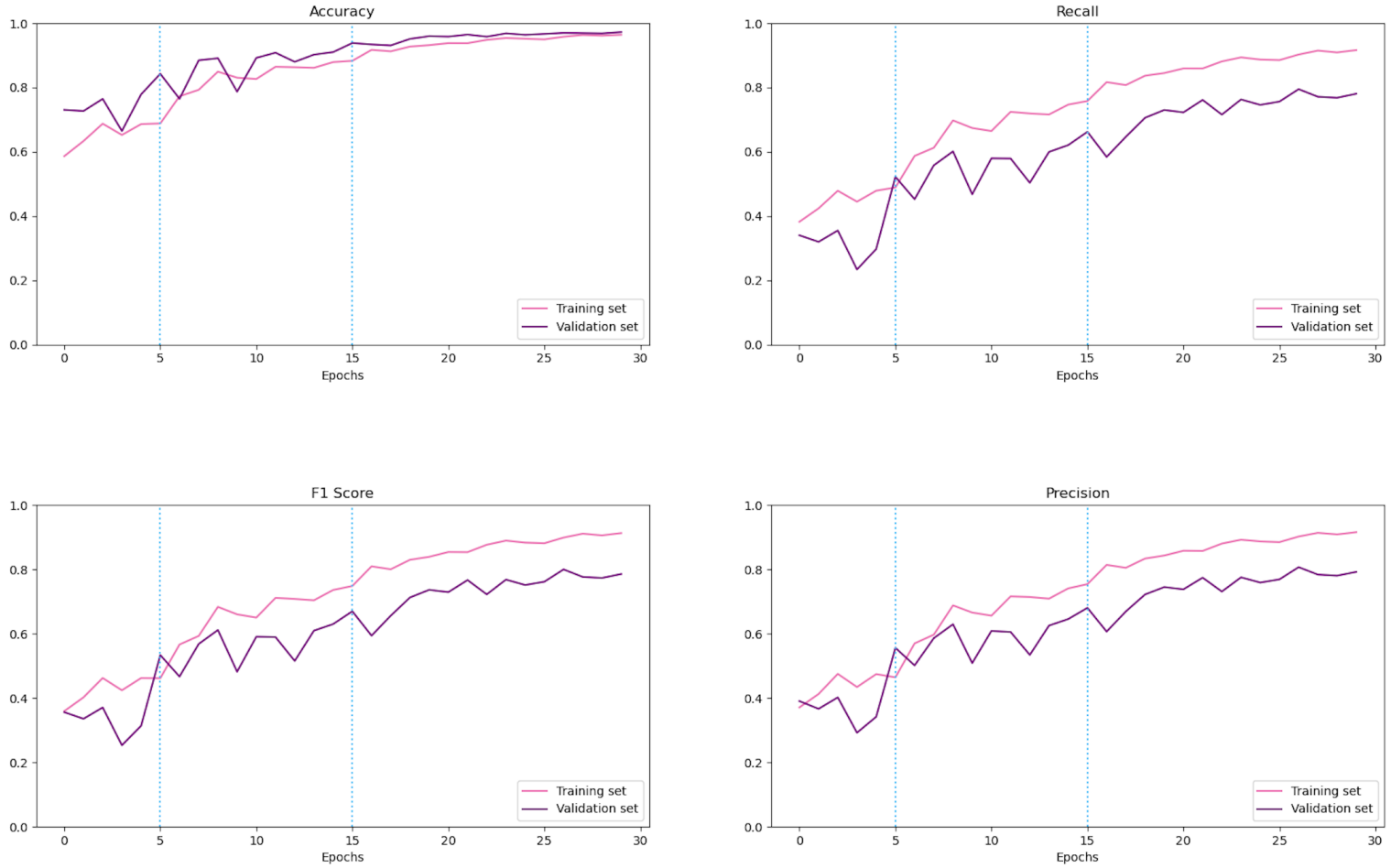


Figure A2: Curves displaying progress during training with all classes weighted equally. The three training phases are indicated with blue lines and begin at Epoch 0, 5, and 15. This graph was not produced during the same training session as A1.

Additional Tables

Table A1: Predictions on the *Dinophysis acuminata* images in the SMHI dataset, made by a classifier trained on images from SYKE.

Predicted class	**Instances**	**Fraction**
Dinophysis acuminata	48	0.64
Unclassified	24	0.32
Euglenophyceae	1	0.01
Licmophora sp.	1	0.01
Prorocentrum cordatum	1	0.01

Table A2: Predictions on the *Mesodinium rubrum* images in the SMHI dataset, made by a classifier trained on images from SYKE (top 5 most abundant classes)

Predicted class	**Instances**	**Fraction**
Unclassified	498	0.47
Mesodinium rubrum	254	0.24
Ciliata	169	0.16
Oocystis sp.	54	0.05
Heterocapsa triquetra	34	0.03

Table A3: Predictions on the *Strombidium sp.* images in the SMHI dataset, made by a classifier trained on images from SYKE (top 5 most abundant classes)

Predicted class	**Instances**	**Fraction**
Ciliata	1338	0.59
Unclassified	745	0.33
Chaetoceros sp.	64	0.03
Heterocapsa triquetra	34	0.01
Cymbomonas tetramitiformis	22	0.01