Design and Implementation of an Automated Image Classification Workflow for Phytoplankton Monitoring

Decrop Wout[a], Rune Lagaisse[b], Mortelmans Jonas[a], Carlota Muñiz[a], Deneudt Klaas[a]

[a]Flanders Marine Institute (VLIZ), InnovOcean site, Jacobsentstraat 1, 8400 Oostende, Belgium
[b]UGent Sint-Pietersnieuwstraat 25, 9000 Gent, Belgium
Corresponding author: Wout Decrop; wout.decrop@vliz.be

Introduction – Phytoplankton plays a pivotal role in both marine and freshwater ecosystems. They provide over 45 percent of the global net primary production, play a pivotal role in the global carbon and nutrient cycles and they fuel entire aquatic foodwebs through their photosynthetic activity (Ducklow et al., 2001; Hays et al., 2005; Pierella Karlusich et al., 2020). Moreover, phytoplankton has the ability to quickly respond to environmental and anthropogenic disturbances through their short generation times. The Water Directive Framework and the Marine Strategy Framework directive recognize phytoplankton as an indicator for monitoring aquatic ecosystem health. (European Commission, 2008) (EEC, 1991). Marine and freshwater plankton represent a highly diverse group of organisms, both taxonomically and morphologically, spanning numerous phyla and tens of thousands of species (Sournia et al., 1991; De Vargas et al., 2015). They are generally small in size, going from less than 1 µm up to 1 mm (Winder & Sommer, 2012). Traditionally, phytoplankton identification has relied on expert taxonomists manually classifying specimens under a microscope. This process is not only time-consuming and resource-intensive but also demands highly skilled specialists to distinguish subtle morphological differences between taxa (Benfield et al., 2007).
To address these issues, researchers are increasingly adopting specially designed high-throughput digital imaging systems. These systems, such as the Video Plankton Recorder (Sournia et al., 1991; Ollevier et al., 2022) and the FlowCam (Sieracki et al., 1998), are employed both in situ and in laboratory settings respectively. These next-generation imaging systems significantly speed up analysis time in the lab by capturing thousands of particles in a matter of minutes. While these systems generate vast volumes of plankton image data, the lack of automated, reliable classification techniques means that manual image classification is still necessary, creating a bottleneck in the processing workflow (Kerr et al., 2020; Sosa-Trejo et al., 2023). Recent studies have turned to automated methods, particularly Convolutional Neural Networks (CNNs) (Krizhevsky et al. 2012), to perform plankton image classification.

Despite the promising advantages of deep learning on phytoplankton images, according to Lumini and Nanni (2019), challenges persist for three main reasons: (i) plankton images are often low-resolution, making classification difficult even for human experts; (ii) the wide range of phylogenetic species in plankton images presents specific challenges for taxonomy; and (iii) class imbalance between datasets and data drift between training and test sets.

In this study, phytoplankton images were captured using FlowCam technology (Fluid Imaging Technologies) and annotated by a human expert. This automated high-throughput device combines the principles of flow cytometry, microscopy, and imaging to generate high-resolution images of particles in liquid samples. As part of the LifeWatch observatory, FlowCam instruments are routinely deployed to monitor microphytoplankton communities in the Belgian Part of the North Sea. This long-term monitoring program yields approximately 300,000 to 400,000 annotated particle images annually, contributing to a continuously expanding archive of quality-controlled biodiversity data. Based on this archive, a curated subset was created for training purposes. The resulting training dataset consists of 95 common taxa from the Belgian Part of the North Sea (BPNS).

In this work, we focus on developing and evaluating a deep learning model for phytoplankton image classification, leveraging high-throughput FlowCam data to address the challenges of automated taxonomic identification and support scalable ecological monitoring. The model was initialized using pre-trained weights from the EfficientNetV2B0 architecture, a member of the EfficientNetV2 family developed by the Google Brain team (Tan and Le, 2021). The model predicts the five most probable classes along with their associated confidence scores. The model reached an accuracy of 86.3% and a top-5 accuracy of 98.8%.

To encourage reproducibility, transparency, and further research, we make the complete image library, the sampled training dataset, and the fully trained classifier openly available. The raw FlowCam image collection, containing phytoplankton observations from the Belgian Part of the North Sea, is accessible via the VLIZ Marine Data Archive Flanders Marine Institute. The annotated subset used for training the model has been published on Zenodo (Decrop et al., 2024), along with the final version of the trained classifier (Decrop and Lagaisse, 2025).

These contributions were developed in the context of the IMAGINE project, which supports open and modular tools for ecological monitoring. The classification pipeline is compatible with the IMAGINE infrastructure and is already available as a module in the IMAGINE Marketplace (https://dashboard.cloud.imagine-ai.eu/marketplace). This framework includes alternative annotation tools that complement or can replace internal systems, providing flexibility for future developments and external use.

Methods

Here, we describe a data acquisition and image processing procedure that includes preprocessing, segmentation, classification, and postprocessing for the accurate identification of 95 classes of phytoplankton using CNNs.

Time series are built as part of the LifeWatch marine observatory in the Belgian Part of the North Sea (BPNS). Several fixed stations are visited regularly using the RV Simon Stevin. A grid of nine coastal stations is sampled on a monthly basis, while eight offshore stations are sampled seasonally. Samples are collected on board using an Apstein net with a 55 µm mesh size and preserved in Lugol's iodine solution. In the laboratory, samples are analyzed using a VS-4 FlowCAM system at 4X magnification, capturing particles in the 55–300 µm size range. Image classification is performed using an automated classifier, followed by manual validation. Since May 2017, this dataset has provided microplankton and phytoplankton data—primarily comprising diatoms, dinoflagellates, and ciliates—for the BPNS. The lab protocol and pipelines are described here Lagaisse (2024); Lagaisse et al. (2025).

Raw FlowCam output data consisting of image collages and a number of .txt files with lab based parameters and measurements are processed via in-house Python data pipeline. The manufacturers VisualSpreadsheet software was used solely for FlowCam operation and data acquisition during the lab run. Raw and binary images are not saved during the FlowCam run; instead, only the image collages created at the end of the run are retained and used for downstream processing and identification. Individual Regions Of Interest (ROIs) are extracted from the image collages based on the coordinates (width and height) from the raw textual data file, using custom Python code. The background of the ROIs remains untouched. These single ROIs are then uploaded to an internal processing database to enable en-masse prediction and subsequent human annotation.

The dataset consists of 95 classes, you can see an overview of the distribution of images training set in Figure 1.

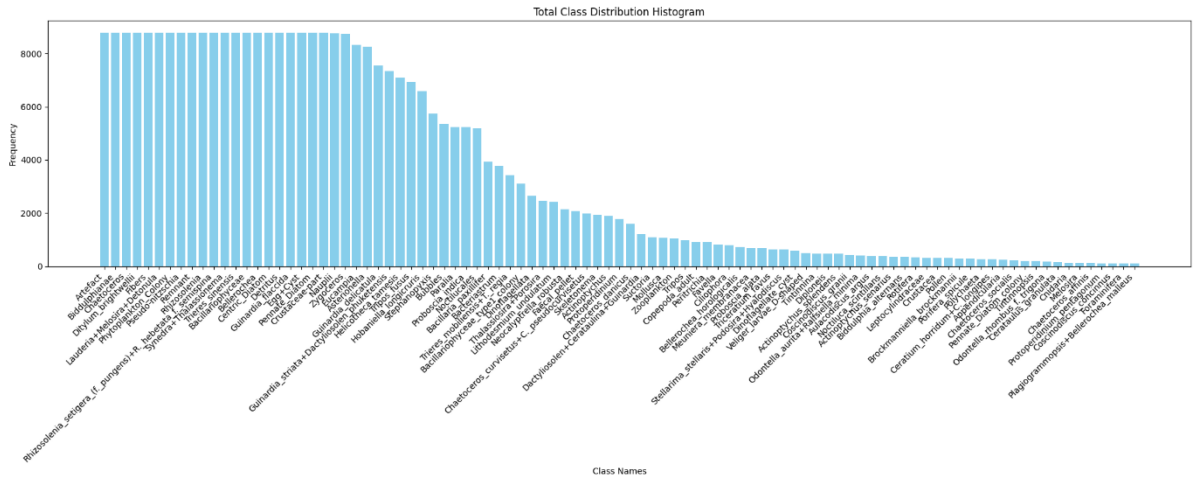

Figure 1. Histogram showing the class distribution across the full phytoplankton dataset.

To balance representation and maintain a realistic distribution, we applied minimum and maximum caps on the number of samples per class, setting the range between 100 and 8800 images. This approach was chosen after several iterations to avoid overfitting on classes with very few samples while preventing dominant classes from overwhelming the model. Through trial and error, this capped distribution provided the best trade-off between model robustness and generalization performance.

The model used for automated image classification is based on EfficientNetV2B0, a deep convolutional neural network known for its use of residual connections to facilitate the training of very deep architectures. EfficientNetV2B0 is part of the EfficientNetV2 family developed by the Google Brain team (Tan and Le, 2021).

Once trained, the classifier produces a ranked list of the top five predicted phytoplankton species for each input image, along with associated confidence scores. These scores represent the model's certainty in each prediction, calculated as softmax probabilities from the final classification layer. Presenting the top five predictions enables users to evaluate multiple plausible labels, which is especially important in ecological datasets where visual similarities between species can introduce ambiguity. The confidence scores also allow users to assess the reliability of each prediction and guide decisions for further expert verification or automated downstream analysis.

Results

The model achieved a Top-1 accuracy of 86.34%, meaning that the correct class was the model's most confident prediction in the vast majority of cases. The Top-5 accuracy increased substantially to 98.76%, suggesting that for almost all samples, the correct label was among the five most probable predictions. This highlights the model's reliability, particularly in scenarios where ranking the most likely classes is sufficient.

Table 1. Grouped classification performance metrics across different Top-K values (percentages)

Group	Metric	Top-1	Top-2	Top-3	Top-4	Top-5
Accuracy	Accuracy	86.34	94.59	97.12	98.19	98.76
Precision	Weighted Precision	86.24	94.61	97.13	98.20	98.77
	Micro Precision	86.34	94.59	97.12	98.19	98.76
	Macro Precision	81.08	92.91	95.93	97.44	98.18
Recall	Weighted Recall	86.34	94.59	97.12	98.19	98.76
	Micro Recall	86.34	94.59	97.12	98.19	98.76
	Macro Recall	77.62	89.39	93.55	95.44	96.64
F1 Score	Weighted F1	86.25	94.57	97.11	98.19	98.76
	Micro F1	86.34	94.59	97.12	98.19	98.76
	Macro F1	78.76	90.84	94.56	96.32	97.35

Table 1 provides a comprehensive overview of the Top-K classification metrics, including accuracy, precision, recall, and F1-score, using micro, macro, and weighted averaging strategies.

• Micro-averaged metrics aggregate contributions from all classes and are generally more influenced by performance on frequent classes. These scores were high across all metrics, with Top-1 micro precision, recall, and F1-score all reaching 86.34%.

• Macro-averaged metrics compute unweighted means across all classes and are more sensitive to class imbalance. At Top-1, macro precision was 81.08%, macro recall was 86.34%, and macro F1-score was 78.76%, indicating solid generalization to both frequent and infrequent classes.

• Weighted metrics fall between the two, accounting for label imbalance while preserving per-class performance. For Top-1 predictions, weighted precision, recall, and F1-score were 86.24%, 86.34%, and 86.34%, respectively—closely tracking the micro scores and indicating balanced overall performance.

As expected, all evaluation metrics improved as the number of considered top predictions (K) increased. For instance, the Top-3 weighted F1-score was 97.12 %, and the Top-5 weighted F1-score rose to 98.76%. These results confirm that the model often assigns high probabilities to the correct class, even when it is not ranked first.

As shown in Figure 2, increasing the probability threshold results in a higher proportion of correct predictions, while also reducing the total number of predictions made.

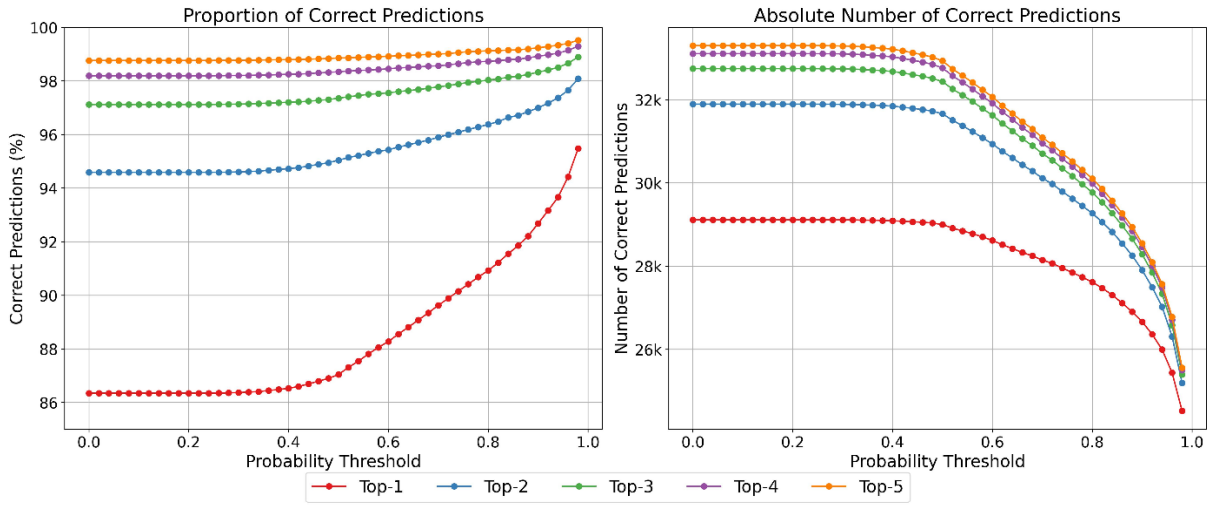

Figure 2. Progression of correct predictions across increasing probability thresholds. Left: Proportion of correct predictions. Right: Absolute number of correct predictions. Each curve corresponds to a different Top-k setting.

Discussion

These results indicate that the model is highly reliable in predicting the correct class within its top few guesses and performs especially well when broader prediction tolerance is allowed. The alignment between micro and weighted scores reflects the model's robustness across the full dataset. The slightly lower macro scores suggest variability in performance among classes, especially those with fewer samples.

The balance between precision and recall, as indicated by the F1 scores, confirms that the model maintains a good trade-off between identifying the correct classes and avoiding false positives. This supports the model's suitability for practical applications, particularly in scenarios where identifying the top few likely classes is sufficient.

Figure 2 illustrates a key trade-off: increasing the probability threshold improves prediction certainty but reduces the number of samples considered. Including more candidate predictions (e.g., Top-2 or Top-3) improves the correct prediction rate while discarding fewer data points. This highlights a practical tuning parameter depending on the intended application—favoring either higher confidence or broader coverage.

This pattern reflects the challenge of class imbalance and taxonomic complexity in plankton image classification and suggests that model improvements—either through data augmentation, hierarchical classification, or targeted data collection—may be beneficial for rare or difficult taxa.

The strong performance at Top-3 and Top-5 levels aligns well with practical workflows in marine ecology, where model predictions can serve as decision support for expert taxonomists. Presenting multiple likely taxa can accelerate annotation while retaining human oversight.
A crucial aspect of this work is the open and free access provided to the entire image library, the annotated training dataset, and the fully trained classifier model. Making these resources publicly available supports reproducibility, transparency, and further research. This accessibility enables other researchers to benchmark and improve upon the classification models, fostering collaboration and accelerating advances in plankton image analysis.
Furthermore, the integration of the classifier and datasets within the Imagine infrastructure enhances the accessibility and usability of these tools. The Imagine platform offers a modular marketplace where the classifier module is available alongside complementary tools, including annotation software that can serve as an alternative to internal systems. This feature is particularly important to address reviewer concerns and to promote standardized, scalable workflows in plankton image annotation and classification.

Conclusion

Improved automated classification of phytoplankton images has significant ecological implications. By accelerating the processing of microscopy data, such models enable more timely detection of plankton blooms, biodiversity shifts, and other critical ecological changes in response to environmental stressors such as climate change, eutrophication, or pollution events. The classifier's strong performance, particularly at Top-3 and Top-5 prediction levels, demonstrates its potential as a valuable tool in operational monitoring pipelines. Rather than replacing expert taxonomists, it offers a means of prioritizing and guiding expert validation, increasing throughput without sacrificing scientific rigor.
Moreover, the public availability of the image dataset, model weights, and source code fosters transparency and reproducibility, enabling researchers worldwide to replicate, validate, and build upon this work. The model only needs a labeled training set to start retraining models, which can be created through many different labeling methods. The integration within the Imagine infrastructure further supports the development of modular, standardized workflows for large-scale marine image annotation and analysis. Looking forward, future research could address remaining challenges such as class imbalance, inter-class similarity, and domain adaptation to microscopy variations.
Altogether, this work represents a significant step toward scalable, accessible, and high-performance plankton image analysis, opening new possibilities for long-term ecological monitoring and global marine biodiversity research.

References

Benfield, M. C., Grosjean, P., Culverhouse, P. F., Irigoien, X., Sieracki, M. E., Lopez-Urrutia, A., et al. (2007). Rapid: research on automated plankton identification. Oceanography 20, 172–187

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mah ´e, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605

[Dataset] Decrop, W. and Lagaisse, R. (2025). Pre-trained phytoplankton species classifier model. doi:10.5281/zenodo.15269453

[Dataset] Decrop, W., Lagaisse, R., Jonas, M., Muyle, J., Amadei Mart´ınez, L., and Deneudt, K. (2024). Lifewatch observatory data: phytoplankton annotated trainingset by flowcam imaging in the belgian part of the north sea (version v1). doi:10.5281/zenodo.10554845

Ducklow, H. W., Steinberg, D. K., and Buesseler, K. O. (2001). Upper ocean carbon export and the biological pump. Oceanography 14, 50–58

[Dataset] Flanders Marine Institute (VLIZ) (2024). Lifewatch observatory data: phytoplankton observations by flowcam imaging in the belgian part of the north sea. doi:10.14284/650

Hays, G. C., Richardson, A. J., and Robinson, C. (2005). Climate change and marine plankton. Trends in ecology & evolution 20, 337–344

Kerr, T., Clark, J. R., Fileman, E. S., Widdicombe, C. E., and Pugeault, N. (2020). Collaborative deep learning models to handle class imbalance in flowcam plankton imagery. Ieee Access 8, 170013–170032

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25

[Dataset] Lagaisse, R. (2024). Lifewatch belgium: Flowcam sampling and lab protocol for imaging microphytoplankton in the belgian part of the north sea. https://www.protocols.io/view/lifewatch-belgium-flowcam-sampling-and-lab-protoco-6qpvr8e6zlmk/v1. Flanders Marine Institute, DOI: 10.17504/protocols.io.6qpvr8e6zlmk/v1

Lagaisse, R., Dillen, N., Bakeev, D., Decrop, W., Focke, P., Mortelmans, J., et al. (2025). Advancing long-term phytoplankton biodiversity assessment in the north sea using an imaging approach. Manuscript submitted for review to *Scientific Data*, Manuscript #SDATA-25-02770

Lumini, A. and Nanni, L. (2019). Deep learning and transfer learning features for plankton classification. Ecological informatics 51, 33–43

Ollevier, A., Mortelmans, J., Vandegehuchte, M. B., Develter, R., De Troch, M., and Deneudt, K. (2022). A video plankton recorder user guide: Lessons learned from in situ plankton imaging in shallow and turbid coastal waters in the belgian part of the north sea. Journal of Sea Research 188, 102257

Pierella Karlusich, J. J., Ibarbalz, F. M., and Bowler, C. (2020). Phytoplankton in the tara ocean. Annual Review of Marine Science 12, 233–265

Sieracki, C. K., Sieracki, M. E., and Yentsch, C. S. (1998). An imaging-in-flow system for automated analysis of marine microplankton. Marine Ecology Progress Series 168, 285–296

Sosa-Trejo, D., Bandera, A., Gonz ´alez, M., and Hern ´andez-Le ´on, S. (2023). Vision-based techniques for automatic marine plankton classification. Artificial Intelligence Review 56, 12853–12884

Sournia, A., Chrdtiennot-Dinet, M.-J., and Ricard, M. (1991). Marine phytoplankton: how many species in the world ocean? Journal of Plankton Research 13, 1093–1099

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In International conference on machine learning (PMLR), 10096–10106