A data integration pipeline towards reliable monitoring of phytoplankton and early detection of harmful algal blooms

Bruna Guterres
Computer Science Center
Federal University of Rio Grande
Rio Grande, Brazil
guterres.bruna@furg.br

Sara Khalid
University of Oxford
United Kingdom
sara.khalid@ndorms.ox.ac.uk

Marcelo Pias
Computer Science Center
Federal University of Rio Grande
Rio Grande, Brazil
pias.marcelo@furg.br

Silvia Silva da C. Botelho
Computer Science Center
Federal University of Rio Grande
Rio Grande, Brazil
silviacb@furg.br

Abstract

Climate change is making oceans warmer and more acidic. Under these conditions phytoplankton can produce harmful algal blooms which cause rapid oxygen depletion and consequent death of marine plants and animals. Some species are even capable of releasing toxic substances endangering water quality and human health. Monitoring of phytoplankton and early detection of harmful algal blooms is essential for protection of marine flaura and fauna. Recent technological advances have enabled in-situ plankton image capture in real-time at low cost. However, available phytoplankton image databases have several limitations that prevent the practical usage of artificial intelligent models. We present a pipeline for integration of heterogeneous phytoplankton image datasets from around the world into a unified database that can ultimately serve as a benchmark dataset for phytoplankton research and therefore act as an important tool in building versatile machine learning models for climate adaptation planning. A machine learning model for early detection of harmful algal blooms is part of ongoing work.

1 Introduction

Climate change is causing progressive warming, acidification, and de-oxygenation of oceans[8]. These conditions produce Harmful Algae Blooms (HAB) and diminish the ability of marine fish and plants to survive, thereby endangering the entire marine ecosystem. Excessive proliferation of phytoplankton species causes HABs and may produce hazardous toxins, adversely affecting human health and economic activity. Phytoplankton monitoring and early detection of HABs are vital for protecting marine life, restoring oceans, and developing climate-resilient economies.

Artificial intelligence (AI) is widely used in image-based classification of phytoplankton species [9, 14]. AI models in turn need to be trained on in-situ phytoplankton images that are sufficiently representative in terms of volume and variety to support early detection and monitoring of HABs. Currently there is no unified database that fulfills this need. Over time, a number of databases have emerged, however they are heterogeneous in multiple ways, e.g. they may capture different types of

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

phytoplankton species from different parts of the world. Hence, there is an imperative need for data integration and standardization.

Additionally the quality of images can vary across databases. The image resolution in some datasets is too low to extract distinct features in detail [10]. Most public datasets comprise gray-scaled images hence lack RGB representations that typically retain more details on phytoplankton characteristics. Public image databases also do not necessarily cover target phytoplankton species for in-situ applications (e.g. aquaculture farms). For instance, the WHOI database, the world's largest plankton classification database, includes over 3.4 million expert-labeled low-resolution gray-scaled images across 70 classes [13]. The RGB image dataset PMID2019 has higher-resolution images, but it only includes 10,819 labeled images for 24 distinct classes [10]. For reliable and representative machine learning models suitable for real-world application, a unified, benchmark database of sufficient quality, variety, and volume is required.

1.1 Proposed Solution and Climate Impact

This work presents an approach towards developing a pipeline for integration and standardization of image databases. The key output is a geographically representative, unified, labeled phytoplankton database.

The proposed pipeline is generic and can be applied widely for curation of real-world data from natural environments. It can support training of AI models such as those for early detection of HAB outbreaks, ultimately contributing to climate resilience and adaptation efforts.

2 Methodology

Aquaculture sites from Brazil, South Africa, Ireland, Argentina and Scotland have provided a list of target phytoplankton species considering the organisms usually encountered in local monitoring campaigns. The list is organized by genus to support research on public databases and further AI development (Figure 1).

Figure 1: Target phytoplankton organisms within aquaculture farms of Brazil, South Africa, Argentina, Ireland, South Africa and Scotland. The information is organized by genus.

Genus	Aquaculture farm
Alexandrium	
Anabaena	
Azadinium	
Centric	
Chaetoceros	
Cilliates	
Dinophysis	
Euglena	
Fragilaria	
Gonyaulax	

Genus	Aquaculture farm
Karenia	
Katodinium	
Leptocylindrus	
Lingulodinium	
Mesodinium	
Nematodinium	
Nodularia	
Paralia	
Pennate	
Prorocentrum	

Genus	Aquaculture farm
Protoceratium	
Pseudo-nitzschia	
Rhizosolenia	
Scrippsiella	
Skeletonema	
Tetraselmis	
Thalassiosira	
Tripos	

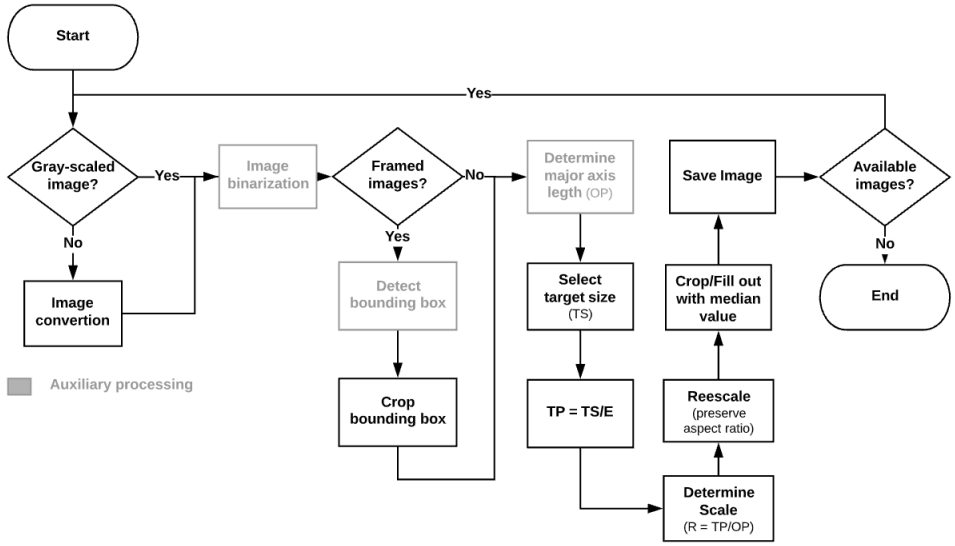
*Argentina ( ), Brazil ( ), Ireland ( ), South Africa ( ) and UK ( )

Recent articles (from 2017 onwards) covering the construction of phytoplankton datasets or image-based models for phytoplankton classification have been studied. The availability of images for each target phytoplankton genus is analyzed in each cited or provided database. The images are organized by genus for further integration.

2.1 Data Integration Pipeline

The proposed methodology uses the most comprehensive public dataset (WHOI) as basis for data processing. Output data are gray-scaled images of fixed size. Phytoplankton organisms are represented at a fixed scaled considering expected size ranges within target genus. Figure 2 illustrates the data integration pipeline.

Figure 2: Pipeline for dataset integration. Gray operations represent auxiliary processing. Target Size (TS) is a random selected value between minimum and maximum expected size of each phytoplankton specie. TP represents the target pixel size considering an output scale E ($\mu m/pixel$).



Although image scale varies within public datasets, the proposed pipeline maintains consistent size among target phytoplankton genus. It considers a fixed output scale (E) for the integrated dataset. For each image, a target size (TS $[\mu m]$) is randomly selected considering minimum and maximum expected sizes of each genus.

Image regions properties are considered to enable consistent output representation. The major axis length (OP) is defined as the number of pixels between the extreme points of longest line along the length of a phytoplankton organism. It is used to determine the ratio coefficient ($R = TP \div OP$) necessary to represent the phytoplankton organisms at a target size in pixels (TP) and fixed scale (E). The ratio coefficient is used to resize the image and maintain size consistency.

Some datasets provide framed images for further usage. The pipeline removes it considering automatically detected image regions. The bounding box of the region with biggest area is considered for frame removal. The Matlab software is used to employ the proposed integration pipeline. Output images are organized in a file system at genus-level to support further development of AI models.

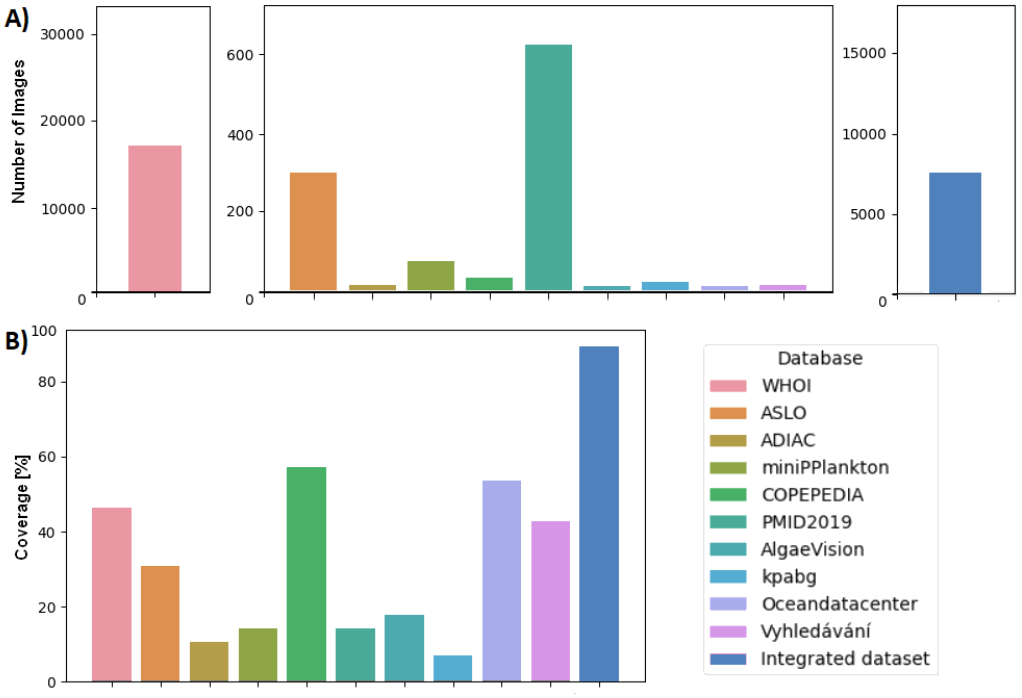
3 Results and Discussions

Fourteen public phytoplankton image datasets were identified from the literature. Tables 1 and 2 summarize the complete list of public databases, main characteristics and the number of images provided for each target phytoplankton genus. Figure 4 illustrates some image examples of distinct databases to showcase the data variability.

Some databases (29%) do not contain genus-level images for any target phytoplankton. Figure 3 illustrates the coverage (i.e. the percentage of genus variety contained in the database) and the average number of images within represented genus for each dataset. The most comprehensive one (Ocean data center) encompassed 54% of target genus with an average of 38 images per not empty class.

Most databases (79%) covered only up to 50% of target phytoplankton genus with an average of 94 images per genus. A significant class imbalance was therefore identified within each database. It

Figure 3: Number of images (A) and Coverage (B) within original and integrated databases. Coverage was measured as the percentage of target phytoplankton genus with at least one image. The number of images is presented as the average and standard deviation of the number of images within covered genus.



reflects natural imbalances within the aquatic environment on which dominant species are imaged more frequently than rare taxa.

The integrated dataset yielded by the proposed pipeline covered 89% of target phytoplankton genus with an average of 7,400 images per genus. It has succeeded on data integration towards a more representative and suitable dataset for aquaculture applications. However, it still struggled with class imbalance since only 57% of target phytoplankton genus achieved at least 20 images. The proposed pipeline considered gray-scaled characteristic of most public databases. It may be basis for new data integration pipelines towards colorful, comprehensive and representative phytoplankton image databases for in-situ applications.

The dataset curation design embedded a standardized mechanism for label-assignment so that the dataset can be used efficiently for machine learning modeling. The output phytoplankton images were organized at genus-level to support the development of AI models suitable for in-situ monitoring programs within aquaculture farms.

4 Conclusions

In this paper an image database integration pipeline was presented, which resulted in a unified, benchmark database covering all publicly available phytoplankton images with an increased coverage from an average of 26% to 89% considering species in the natural marine environment. It can serve as an important tool in building versatile machine learning models for planning protection and resilience of marine ecosystems in the face of climate change. Data quality assessment, and application to early detection machine learning models is part of ongoing work.

Acknowledgments

This work was developed as part of the ASTRAL (All Atlantic Ocean Sustainable, Profitable and Resilient Aquaculture - `https://www.astral-project.eu`) project. This project has received

funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N^0 863034.

References

[1] Algaevision. http://algaevision.myspecies.info.

[2] Copepedia: The database of taxonomy, distribution maps, photos, and biometric traits. https://www.st.nmfs.noaa.gov/copepod/about/about-copepedia.html.

[3] Kaggle national data science bowl. https://www.kaggle.com/c/datasciencebowl.

[4] Ocean data center. http://oceandatacenter.ucsc.edu/PhytoGallery/toxigenic.html.

[5] Planktonmkl. https://github.com/zhenglab/PlanktonMKL.

[6] Vyhledávání. http://galerie.sinicearasy.cz/galerie.

[7] Robert K Cowen, S Sponaugle, K Robinson, and J Luo. Planktonset 1.0: Plankton imagery data collected from fg walton smith in straits of florida from 2014–06-03 to 2014–06-06 and used in the 2015 national data science bowl (ncei accession 0127422). *NOAA National Centers for Environmental Information*, 2015. Dataset. https://doi.org/10.7289/v5d21vjd. Accessed: March, 2021.

[8] Christopher J Gobler. Climate change and harmful algal blooms: insights and perspective. *Harmful algae*, 91:101731, 2020.

[9] Thomas Kerr, James R Clark, Elaine S Fileman, Claire E Widdicombe, and Nicolas Pugeault. Collaborative deep learning models to handle class imbalance in flowcam plankton imagery. *IEEE Access*, 8:170013–170032, 2020.

[10] Qiong Li, Xin Sun, Junyu Dong, Shuqun Song, Tongtong Zhang, Dan Liu, Han Zhang, and Shuai Han. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES Journal of Marine Science*, 77(4):1427–1439, 2020.

[11] Jessica Y. Luo, Jean-Olivier Irisson, Benjamin Graham, Cedric Guigand, Amin Sarafraz, Christopher Mader, and Robert K. Cowen. Data from Automated plankton image analysis using convolutional neural networks. October 2018.

[12] AV Melechin, DA Davydov, SS Shalygin, and EA Borovichev. Open information system on biodiversity cyanoprokaryotes and lichens cris (cryptogamic russian information system). *Bulleten MOIP. Otdel biologicheskiy*, 118(6):51, 2013.

[13] Eric C Orenstein, Oscar Beijbom, Emily E Peacock, and Heidi M Sosik. Whoi-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. *arXiv preprint arXiv:1510.00745*, 2015.

[14] Rene-Marcel Plonus, Jan Conradt, André Harmer, Silke Janßen, and Jens Floeter. Automatic plankton image classification—can capsules and filters help cope with data set shift? *Limnology and Oceanography: Methods*, 2021.

[15] Rene-Marcel Plonus, Jan Conradt, André Harmer, Silke Janßen, and Jens Floeter. Automatic plankton image classification - can capsules and filters help coping with data set shift? January 2021.

[16] Heidi M Sosik and Robert J Olson. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, 5(6):204–216, 2007.

[17] Xin Sun, Hongwei Xv, Junyu Dong, Huiyu Zhou, Changrui Chen, and Qiong Li. Few-shot learning for domain-specific fine-grained image classification. *IEEE Transactions on Industrial Electronics*, 68(4):3588–3598, 2020.

5 Appendix

Figure 4: Images of some phytoplankton species identified in six different public databases.

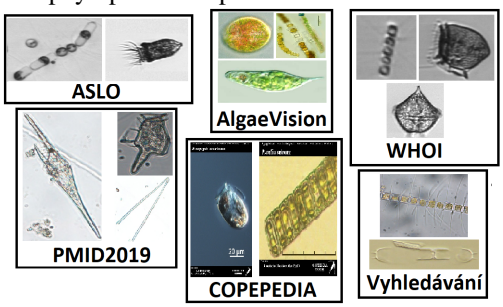


Table 1: Image-based databases for phytoplankton modeling and the number of images for target species within aquaculture farms. No target phytoplankton images were encountered for the dabases Kaggle [3], PlanktonSet [7], ISIIS [11] and VPR [15]

Item	**WHOI [16]**	**ASLO [5]**	**ADIAC**	**Mini PPlankton [17]**	**Vyhledávání [6]**
Representation	Gray Scale	Gray Scale	Gray Scale	RGB	RGB
Genus	**Number of Images**				
Alexandrium	0	0	0	0	0
Anabaena	0	0	0	0	19
Azadinium	0	0	0	0	0
Centric	0	0	0	0	0
Chaetoceros	45594	300	0	70	9
Cilliates	11613	300	0	0	0
Dinophysis	295	0	0	70	5
Euglena	542	300	0	0	14
Fragilaria	0	0	20	0	24
Gonyaulax	3	0	0	0	1
Karenia	0	0	0	0	0
Katodinium	0	0	0	0	1
Leptocylindrus	101375	0	0	0	0
Lingulodinium	0	0	0	0	0
Mesodinium	0	0	0	0	0
Nematodinium	0	0	0	0	0
Nodularia	0	0	0	0	18
Paralia	413	0	1	0	0
Pennate	4766	0	300	0	0
Prorocentrum	2590	0	0	0	5
Protoceratium	0	0	0	0	0
Pseudo-nitzschia	3220	300	0	0	0
Rhizosolenia	30.554	300	0	70	4
Scrippsiella	0	0	0	0	0
Skeletonema	12.323	300	0	70	6
Tetraselmis	0	0	0	0	0
Thalassiosira	11.025	300	5	0	0
Tripos	0	0	0	0	15
Coverage	46.43%	28.57%	10.71%	14.29%	42.86%

Table 2: Image-based databases for phytoplankton modeling and the number of images for target species within aquaculture farms. No target phytoplankton images were encountered for the dabases Kaggle [3], PlanktonSet [7], ISIIS [11] and VPR [15]

Item	**COPEPEDIA [2]**	**PMID2019 [10]**	**Algae Vision [1]**	**kpabg [12]**	**Ocean data center [4]**
Representation	Gray Scale	Gray Scale	RGB	RGB	RGB
Genus	**Number of Images**				
Alexandrium	0	0	0	0	11
Anabaena	0	0	0	29	0
Azadinium	0	0	0	0	0
Centric	0	0	0	0	0
Chaetoceros	300	0	70	0	7
Cilliates	300	0	0	0	0
Dinophysis	0	0	70	0	7
Euglena	300	0	0	0	0
Fragilaria	0	20	0	0	2
Gonyaulax	0	0	0	0	7
Karenia	0	0	0	0	6
Katodinium	0	0	0	0	0
Leptocylindrus	0	0	0	0	4
Lingulodinium	0	0	0	0	12
Mesodinium	0	0	0	0	0
Nematodinium	0	0	0	0	0
Nodularia	0	0	0	10	3
Paralia	0	1	0	0	0
Pennate	300	0	0	0	0
Prorocentrum	0	0	0	0	8
Protoceratium	0	0	0	0	1
Pseudo-nitzschia	300	0	0	0	6
Rhizosolenia	300	0	70	0	2
Scrippsiella	0	0	0	0	0
Skeletonema	300	0	70	0	7
Tetraselmis	0	0	0	0	0
Thalassiosira	300	5	0	0	11
Tripos	0	0	0	0	0
Coverage	28.57%	10.71%	14.29%	7.14%	53.57%