

**DEEP-LEARNING FOR AUTOMATED DIATOM DETECTION AND
IDENTIFICATION FOR THE ECOLOGICAL DIAGNOSIS OF FRESH
WATER ENVIRONMENTS**

A Thesis
Presented to
The Academic Faculty

By

Pierre Faure--Giovagnoli

In Partial Fulfillment
Of the Requirements for the Degree
Master of Computer Science

Georgia Institute of Technology

August 2020

Copyright © Pierre Faure--Giovagnoli 2020

**DEEP-LEARNING FOR AUTOMATED DIATOM DETECTION AND
IDENTIFICATION FOR THE ECOLOGICAL DIAGNOSIS OF FRESH
WATER ENVIRONMENTS**

Approved by:

Dr. Cedric Pradalier, Tutor
School of Interactive Computing
Georgia Institute of Technology

Dr. Joseph Montoya
School of Biological Sciences
Georgia Institute of Technology

Dr. Ghassan AlRegib
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: July 24, 2020

I have learned to love diatoms during this project...
...and I hope you will like them a little more after this reading.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisors Dr. Pradalier and Dr. Laviale who have been of great support during all the thesis, providing valuable advice and motivation. Besides, thanks to my partner Souhila Founas for her positive energy and great work on this project. Equally, the completion of this project could not have been accomplished without the support of the UMI 2958 CNRS-GT and UMR 7360 CNRS-UL laboratories, providing funds and hardware support.

I would also like to thank my INSA Lyon tutor Dr. Solnon which have been providing precious feedbacks and help during the project. Thanks also to my schools Georgia Tech and INSA Lyon to give me such a great opportunity but also to provide me with the knowledge and tools to make it happen.

Finally, many thanks to Dr. Montoya and Dr. AlRegib for taking the time to review this project.

TABLE OF CONTENTS

Acknowledgments	iv
List of Tables	vii
List of Figures	ix
Summary	x
1 Introduction	1
1.1 What are diatoms ?	1
1.2 The project	2
1.3 Microscopy imaging	3
2 Related works	4
2.1 Detection	4
2.1.1 Microorganism detection	4
2.1.2 Object detection	5
2.1.3 Synthetic dataset	6
2.2 Classification	6
2.2.1 Diatom classification	7
2.2.2 Image classification	9
2.2.3 Hierarchical Classification	10
3 Materials	12
3.1 Diatom detection	12
3.1.1 Atlas	12
Construction	13
Content	14

3.1.2	Debris	14
3.1.3	DIC	15
3.2	Diatom classification	15
3.2.1	Aqualitas	15
3.2.2	ADIAC	16
4	Methods	17
4.1	Diatom detection	17
4.1.1	Synthetic microscope image generation	17
4.1.2	Datasets splitting	20
4.1.3	Training and evaluation	20
4.1.4	Metrics	21
4.2	Diatom classification	23
4.2.1	Datasets	23
4.2.2	Training and evaluation	24
4.2.3	A hierarchical approach	24
Similarity Matrix	25	
Artificial taxonomy	26	
Potential applications	26	
5	Results and discussion	27
5.1	Diatom detection	27
5.2	Diatom classification	31
5.3	Hierarchical classification	33
6	Conclusion	35
References		36

LIST OF TABLES

3.1	Summary of the datasets used in this paper	12
5.1	Results the diatom detection (6 COCO's metrics)	27
5.2	Results of diatom classification. The \pm denotes for standard deviation. <i>(1) If "Yes", the dataset has been augmented to compensate for class unbalance as explained in the Methods section. (2) This score is discussed in the Results section.</i>	31

LIST OF FIGURES

1.1	Example of diatoms with great variety of shapes and sizes (source: D. Heudre, DREAL Grand Est)	1
2.1	Simplified taxonomic tree of fresh water diatoms (source: [29])	8
3.1	Example of diatom atlases pages	13
3.2	Histogram of the dataset distribution <i>On the abscissa, number of images in the dataset grouped in bins. In ordinates, number of taxa with this number of images. Blue ticks at the bottom indicate the real number of image for each taxon.</i>	14
3.3	5 samples extracted from the Rhône-Alpes atlas (light microscopy). Associated species listed from left to right: <i>Luticola goeppertiana</i> , <i>Aulacoseira granulata</i> var. <i>angustissima</i> , <i>Gomphonema parvulum</i> var. <i>parvulum</i> , <i>Cyclostephanos invisitatus</i> , <i>Achnanthidium catenatum</i>	14
3.4	5 samples extracted from the Debris dataset	15
3.5	5 samples extracted from the DIC dataset with labeled diatoms	15
3.6	4 samples extracted from the Aqualitas dataset (light microscopy). Associated species listed from left to right: <i>Cyclostephanos dubius</i> , <i>Eolimna rhombelliptica</i> , <i>Gomphonema micropumilum</i> , <i>Cyclostephanos invisitatus</i> , <i>Humidophila contenta</i>	16
3.7	5 samples extracted from the ADIAC dataset (light microscopy). Associated species listed from left to right: <i>Achnanthes parvula</i> , <i>Amphora veneta</i> , <i>Cocconeis peltoides</i> , <i>Gomphonema ventricosum</i> , <i>Gomphonema ventricosum</i>	16
4.1	Example of real microscope image with labeled diatoms (DIC, X1200) .	17

4.2 Illustration of the artificial image generation procedure P. (1) <i>Patches are firstly selected randomly from the debris and diatom datasets and placed at a random position on a first image I' after applying histogram matching with T_{ref}.</i> (2) <i>Then we use borders B to create a smooth gradient on a new image I</i> (3) <i>to finally place all the diatoms in random order using seamless blending [34].</i>	19
4.3 Examples of synthetic images	19
4.4 Illustration of the Intersection over Union (IoU) measure. <i>The green bounding box can been seen as the ground truth and the red as the detected one.</i>	22
4.5 Illustration of the data augmentation pipeline. <i>Images are first converted to square and then randomly augmented. Possible augmentations are: rotation, horizontal and vertical flips, horizontal and vertical shifts, zooms and brightness variations.</i>	24
5.1 Comparisons of the 2 diatom detection pipelines with the groundtruth .	30
5.2 Full dendrogram and extract of the bottom-left part	33
5.3 Images of diatoms from NHEU and NIPF diatoms	34

SUMMARY

Diatoms are a type of unicellular microalgae found in all aquatic environments. Their great diversity and ubiquity make these organisms recognized bio-indicators for monitoring the ecological status of watercourses, notably in the frame of the implementation of the European Water Framework Directive. In this context, we propose a study on diatom detection on microscope images using a deep-learning object detection architecture. To reduce the number of manually labeled images needed for training, we use a synthetic dataset in pair with a real one and gain more than 10% of precision and 5% of recall. This synthetic dataset represents a significant time saving, especially as it is made from publicly available images provided by diatom atlases, avoiding the extensive task of microscopic image acquisition. Diatom detection can be used for many tasks and notably for further classification of the extracted thumbnails by hand or using machine learning. To illustrate this use, we will also propose an update on automatic diatom classification using the latest advances in image classification. Finally, we will also discuss the applications of artificial taxonomy in the case of hierarchical diatom classification.

Chapter 1

Introduction

1.1 What are diatoms ?

Diatoms are a type of microscopic algae found in all aquatic environments. They are essential living organisms as they are estimated to represent 50% of oceanic primary production. Those unicellular organisms have the particularity of being surrounded by an external silica cell wall called a frustule. These frustules can adopt a great variety of shapes (Figure 1.1) that, in conjunction with the various properties observed among those organisms, led the biologists to create a complex taxonomy of thousands of taxa. The main distinctive features to identify the taxa are the general shape, the number, shape and places of the raphes (slits on the frustule's surface) and the various shell's ornamentations.



Figure 1.1. Example of diatoms with great variety of shapes and sizes (source: D. Heudre, DREAL Grand Est)

The study of diatoms has become increasingly important these last decades, especially in research on climate change and water quality assessments as diatoms are extremely responsive organisms. Indeed, small changes in their environment (pollution, temperature...) can impact the taxa proportions greatly. From those observations, diatoms have been at the center of multiple ecological analysis, the proportions of taxa

being used as variables for water quality indices, especially in the context of the implementation of the European Water Directive (WFD:2000/60/EC) [2, 8, 36, 38].

Note *In this study, we will knowingly use the words "family", "species", "genus" or "variety" to denote specific taxonomic ranks. The term "taxon" is generic and can refer to any rank of this classification.*

1.2 The project

In this project, we focus on the detection of diatoms in microscope images. This operation is essential to study diatom populations and notably for a following classification work made manually or automatically. In particular, we want to explore the latest advances in object detection architectures but also the use of a synthetic dataset in pair with real images. The goal of this process is to reduce the number of manually labeled images which are time-consuming to make. We will also address the use case of diatom classification by comparing the results of the latest Convolutional Neural Network (CNN) image classifiers to previous works.

Detection By detection, we understand the localization of diatoms on a microscope image. Hence, our objective is to apply a state of the art object detection algorithm to detect diatoms in light microscopy images. We will train a Faster R-CNN architecture in two stages: (1) we will construct a generic diatom detector model using a synthetic dataset and (2) we will fine-tune our model on a selection of real images. To our knowledge, this application of object detection including the use of synthetic microscope images has never been seen literature and the benefits of this approach will be thoroughly evaluated qualitatively and using common Object Detection (OD) metrics. To demonstrate the contribution of the synthetic dataset, we will notably compare the results to a control pipeline trained without the synthetic images.

Classification Diatom classification is an essential task for many works, notably ecological monitoring. If automatic diatom classification has been an active topic for a few decades now, the use of CNN classifiers is almost absent and most of the studies use different datasets, making comparison difficult. This is why we also propose

an update on previous works, using their datasets with the Xception architecture, a state of the art CNN image classifier. We will notably show that CNNs can achieve scores as good as those obtained using regular classifiers and handcrafted features (morphological and textural).

Moreover, we will lay the foundation for a hierarchical classification approach based one the model's prediction errors. We will notably detail a method to generate an artificial taxonomy and will discuss the possible application in a classification scheme.

1.3 Microscopy imaging

As most diatoms measure between 10 and $200\mu\text{m}$, we will be working with microscopic imagery. Among the microscopy techniques currently available, we want to make our study suitable for light microscopy in general, including the multiple illumination techniques such as brightfield, phase or differential interference contrast (DIC). If the diatom images we will use to create the synthetic dataset are made using brightfield microscopy, the fine-tuning phase will allow to adapt the model to other illumination techniques. We support this claim by using a real image dataset made using DIC.

Microscopic images have some specificities which seem worth pointing out, among them: cells can be totally or partially transparent which can lead to visible overlappings, images have no reading direction (ie. up or down) and some artifacts like debris (fragments of frustules, sediment particles...) or focusing problem can appear. We will address these issues in the later parts, especially in the data augmentation section.

Chapter 2

Related works

2.1 Detection

In this this study, we propose a new method to detect diatoms that could be easily extended to microorganism detection in general. Only a very few works on microorganism detection using CNNs are present in literature and, to our knowledge, we are the first to apply this approach to diatom detection. In particular, our work relies on the use of a dataset of synthetic microscope images to diminish the number of hand-labelled images and we found no equivalent in previous literature. In this first part of the survey, we will start by studying the previous methods used for microorganism detection and will also discuss previous uses of deep-learning OD. Finally, we will address the use of synthetic datasets in machine learning.

2.1.1 Microorganism detection

Automatic analysis of microscope images has been an active topic since the 1950s. At that time, we were not yet talking about detection but mainly of segmentation, with extensive work on cell segmentation in the medical field (blood cells, neurons...). Those approaches relied on numerous methodologies including thresholding (watershed, intensity, shape...), feature detection or morphometry [28]. Later, [18] proposes a segmentation based on a Gaussian prior and a few other papers tried region-based classification approaches [20, 44]. While these studies have had good results, they required the extensive intervention of biologist experts (for modeling or manual segmentation), were sensitive to debris and overlays or relied on images taken using specific processes like fluorescent markers or multi-spectral imagery, not making them easily generalizable. A few works on diatom segmentation have also been made but the main goal was not detection but the creation of morphological descriptors for classification. Some famous systems like the FlowCam also allow microorganism detection along with morphological features extraction. However, in addition to their software being proprietary, it relies on specific imaging techniques (water with microorganisms flows

through a tube which allows to separate entities more easily) which do not correspond to the generic microscope images generally used.

If OD has made a real breakthrough with the use of deep-learning architectures around 2014, using high-level features instead of handcrafted ones, we found only very few works using those technologies for microorganism detection. In 2019, [27] proposes a solution for phytoplankton detection in microscope images using a hand-labeled dataset they created to train state of the art OD deep-learning architectures (Faster R-CNN, YOLOv3, RetinaNet...). To our knowledge, this is the first time that such architectures are used for microorganism detection.

2.1.2 Object detection

The early apparitions of OD in literature relied on traditional Computer Vision (CV) tools mostly based on handcrafted features. In 2001, the iconic Viola-Jones face detector based on very simple descriptors and a cascade of classifiers may be one of the greatest application example [42]. More complex image descriptors like SIFT also appeared around that time, allowing to describe complex object structures and to match them using techniques like bag of words or RANSAC. More advanced methods like the HOG (Histogram of Oriented Gradients) Detector in 2005 [9], introducing the use of sliding windows, or the Deformable Part-based Model [16] in 2008 have led to great improvements. Since 2010, OD using traditional CV tools did not know any significant progress with only few improvements [46].

2014 marked the beginning of a new era for OD. The use of deep-learning architectures to extract robust and high level features allowed significant advances. The goal of those architectures is to produce a set of Bounding Boxes (BB) enclosing objects on the images. From there, we distinguish two main architectures: One-stage, and Two-stage. The One-stage family (YOLO, SSD, RetinaNet...) generally uses an unique network to extract and classify the BBs. Its Two-stage counterpart (mostly the R-CNN lines) first uses a Region Proposal Network to propose multiple Region of Interest (RoI) (potential BBs) which are subsequently classified using a second generic backbone architecture (any image classification network). One stage architectures are

generally faster but less robust and therefore better suited for real-time applications. For a more comprehensive OD history, see [46].

As time performances are not an issue in our project, we will use the Faster-RCNN architecture as it is a robust architecture of the OD field. Moreover, many implementations are available, making it easily reusable. As a backbone architecture, Resnet-101 is a good trade-off as it is a reasonably sized and robust network considering that we only need binary classification (the presence or absence of diatom).

2.1.3 Synthetic dataset

The primary component of any machine learning problem is data, the ideal dataset being large and diversified, theoretically showing to the model all the examples it could possibly encounter. Therefore, gathering a quality dataset can be laborious and time consuming. In the context of object detection, the difficulty comes from the image acquisition and from the labeling. Capturing a comprehensive set of images can be tedious (especially in the frame of biology where finding the organisms can sometimes already be tricky) and drawing the BBs is most of the time manual or at best semi-manual. That is why for the past few years we have seen an increasing use of synthetic data in machine learning. Numerous techniques are used for this purpose (3D modeling, generative models...) and a good overview of the recent developments and uses of synthetic data in machine learning can be found in [30].

In this study, we are going to generate synthetic images of microscopic slides containing various diatoms but it is also essential to incorporate debris to teach the model to avoid them. As we start with cropped diatom images with a wide range of resolutions, brightness and background colors, it is important to create a seamless smooth image to avoid misleading the network with polluted data. To achieve that, we will use various CV techniques like histogram matching or seamless blending [34].

2.2 Classification

Once the diatom is detected and cropped from the microscope image, we propose a classification process relying on a state of the art CNN. If microorganism classification

is far from new, the use of CNNs is much rarer and only one study mentions its use for a diatom application. In this second part, we will summarize the diatom classification approaches found in literature and then will study the latest advances in CNNs for image classification to justify our architecture choice.

2.2.1 Diatom classification

Due to their ecological interest and their ubiquity on Earth, diatoms have been a subject of interest for many researchers. In particular, their classification turned out to be a real challenge given the thousands of taxa currently referenced and many solutions have been studied to automate the process. If the diatom taxonomy can appear relatively simple in the upper taxonomic ranks (Figure 2.1), it can become quite deep and complex in lower levels, especially for some specific species and genus. Taking the example of water quality assessment, it can be crucial to classify at those lower ranks to obtain meaningful indicators and it is important to understand the diversity of classification schemes possible. Most of previous works focus on classification on the specie level but biologists work on every stage, even on deformations at the individual level [26]. As morphological variations reduce at each taxonomic rank, we understand that it is easier to classify at higher levels and, for a given group, some taxa are harder to differentiate than others.

From those observations, we realize that the comparison of results made with different datasets does not make sense, the diatoms composing the datasets impacting greatly the classification results. Nonetheless, none of the works we found during our survey have worked on the same exact datasets and one of the goal of this project is to study the use of CNNs on diatom classification by comparison with other methods and therefore, with previously used datasets.

If some of the early techniques were based on holographic filters [5], absorbance measurements [21] or harmonic decomposition [35], the pilot study ADIAC [15] (Automatic Diatom Identification and Classification) might have been the first large scale project purely dedicated to automatic diatom identification using machine learning techniques. The research focused on building a large labeled database of diatoms, automatizing image acquisition of microscope slides and testing the identification methods avail-

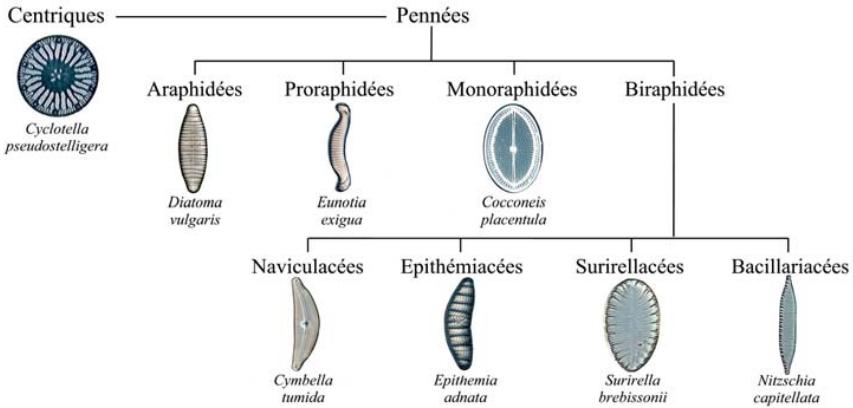


Figure 2.1. Simplified taxonomic tree of fresh water diatoms (source: [29])

able. Those methods were mainly based on diatom morphometric descriptors (shape, symmetry, compactness..). They managed to gather 4700 images of 500 diatom taxa and obtain 96.9% of accuracy with a tree forest using 321 mathematical descriptors to differentiate 37 taxa [14].

Many following works focused on the use of morphometry for automatic diatom classification. An important part of this research focused on automatic diatom segmentation, taking a diatom image as an input and automatically finding its contours [19, 22]. If most of those approaches were successful, the contours of diatoms are far from sufficient for their classification as a lot of distinctive features lie in the frustule's ornamentation and this is why a lot of effort was also put in the improvement of morphometric and texture descriptors [32]. In 2011, a new study on the ADIAC dataset showed an improvement with 97.97% for an ADIAC subset of 38 taxa using Bagging with 230 descriptors. This accuracy falls to 96.17% for 55 taxa in the same conditions [10]. In 2017, a 98.11% of accuracy was reached using a collection of refined morphological descriptors [4]. They obtained their score using 10-fold cross-validation (fcv) on a dataset of 80 species with at least 100 samples per taxon that we will name the "Aqualitas" dataset [3]. If this score is noticeably higher than previous ones, the use of a different dataset makes it irrelevant to compare with ADIAC scores. As pointed out by the authors, their evaluation process is also subject to bias as they perform data augmentation before splitting the dataset for 10-fcv, therefore showing augmented versions of the exact same source image for training and validation.

The rise of the computing power opened up new opportunities for automatic microorganism classification using CNNs. The major difference with morphometric classification is that the network learns by itself the most distinctive morphological characteristics using chains of convolutions. In 2015, following a Kaggle competition on automatic plankton classification, one of the first use of CNN for microorganism classification was released. They reached an accuracy of 73.90% with a dataset of \approx 30000 images of 121 taxa [25]. A few publication on microorganism followed using purely CNNs [24] or mixing them with other models [43]. To our knowledge, no work on diatom classification using CNNs has been made before 2017. It was that year that the same team behind the "Aqualitas" dataset also published a paper on automatic diatom classification using CNNs [33]. They obtained 99.51% of accuracy for the same dataset.

2.2.2 Image classification

If automatic image classification appeared before OD, the early techniques involving traditional CV were about the same, using mostly handcrafted features [23]. Around 2000s, LeCun introduces LeNet-5, a pioneer 7-layers convolutional architecture tested successfully on handwritten digits recognition [45]. Thereafter, architectures have gradually become more and more complex, introducing new types of layers and concepts (pooling layers, residual networks, inception modules...). For a comprehensive overview of the subject, see [39].

Nowadays, numerous CNN architectures are available (VGG, Inception, Xception...) and there is no absolute rule to choose one. Multiple parameters must be considered in that choice: accuracy on reference datasets, size and portability, prediction time, number of parameters... Given the problem, Xception seems to be a good trade-off. As the portability and time efficiency are not an issue, the great number of parameters of Xception will allow to extract the many features needed to effectively differentiate many classes of diatoms and its excellent scores on ImageNet have proven its robustness on complex problems. It seems to us that Xception is a good choice to show the current advances in the field of CNNs for image classification [6].

2.2.3 Hierarchical Classification

In the papers we studied, the number of taxa used for classification never got past 80 and the precision decreases inversely to the number of taxa for a given study. Moreover, depending on amount of data available and the species we want to classify, automatic diatom classification can be very difficult. This is why, in this paper, we will also focus on an alternative classification approach based on Hierarchical Classification (HC) to propose adaptive classification levels based on the dataset challenges. To our knowledge, no work about hierarchical diatom classification has been published yet and this survey will therefore be about HC in general.

When usual classification schemes seen in literature use a flat organization of classes with the assignation of a single label to the model input, the HC approach organizes classes with a set of relationships. The output of hierarchical classifiers can be of various forms but it must take the classes relationships into account to be called as such. When working on HC, it is essential to first study the problem itself and its hierarchical properties. We distinguish mainly two types of hierarchical structures, trees and DAGs (Directed Acyclic Graph), the later allowing the nodes to have more than one parent. In addition to those categories, it is common to qualify the type of nodes used to make the classification: a *real tree/DAG* only allows leaf nodes to be assigned as labels when in a *virtual tree/DAG*, internal nodes can also be used to do so [40, 41]. For a comprehensive overview of the subject, see [17].

We understand from the HC definition that a hierarchy of classes has to be chosen and used to construct the model. In the frame of a biological application, using the taxonomy of the studied organism can be a viable option. However, some alternatives try to take the specificities of image classification into account, notably by creating the hierarchy based on the errors made by a given model on a test dataset grouping notions of visual similarity (classes objectively hard to differentiate), completeness of the dataset (lack of samples for training) and learning bias (sub-optimal model for learning specific features). In this study, we will mainly discuss the creation of this artificial taxonomy and apply a method based on the confusion matrix [37]. Nonetheless, we must keep in mind that many alternatives have been proposed to address this problem

(see [31]).

For our study, the question of the importance of HC in large taxonomic problems like diatoms has two main answers. (1) Firstly, by having multiple levels of classification, we can choose at which level to stop the classification process depending on the amount of data on specific branches and therefore provide an adaptive answer based on the current state of the dataset. This analysis is based on the trivial observation that lower levels are the levels trained with the lower number of examples, making them less robust than their higher level counterparts. (2) The second reason depends on the model architecture but lots of hierarchical solutions propose various classification stages in which classifiers become more and more specialized, proposing a better overall accuracy.

Chapter 3

Materials

We will use five datasets for this paper. The "Atlas" and "Debris" datasets are composed respectively of cropped diatoms and debris and will be used to create the synthetic images. The "DIC" dataset is a set of 187 manually labeled real microscope images made using DIC illumination and will be used to fine-tune and test our model. Finally, the "Aqualitas" and "ADIAC" datasets, which are also composed of cropped diatoms, are datasets from previous studies and will be used for comparison in the classification section.

Table 3.1. Summary of the datasets used in this paper

Name	Description	Original paper(s)
Atlas	206 taxa	this one
Debris	600 debris samples	this one
DIC	187 microscope images	this one
Aqualitas	100 taxa	[4, 33]
ADIAC55	55 taxa	[10]
ADIAC48	48 taxa	[10]
ADIAC38	38 taxa	[10]

3.1 Diatom detection

3.1.1 Atlas

The "Atlas" dataset will be the main diatom dataset used for the rest of this project. To set some boundaries and given the study context, we focused on the species found in the hydrographic basin Rhin/Meuse. Building a robust dataset is an essential task in a machine learning context and many factors like balance and diversity must be considered. Within this paper, we will discuss those parameters extensively as well as the procedures we followed to make the most of the dataset we gathered.

Construction

Diatom atlases generally gather diatom image samples retrieved in the rivers of some delimited area labeled with their identified taxon. We used diatom atlases provided by the public french organism DREAL (Direction régionale de l'Environnement, de l'Aménagement et du Logement) to construct our primary dataset. Those atlases are generally given as PDF files of various formats and contents. Hence, the construction of the dataset mainly required various PDF scrapping and CV tasks to extract automatically the numerous diatoms and their labels. Example of pages of such atlases are presented in Figure 3.1.

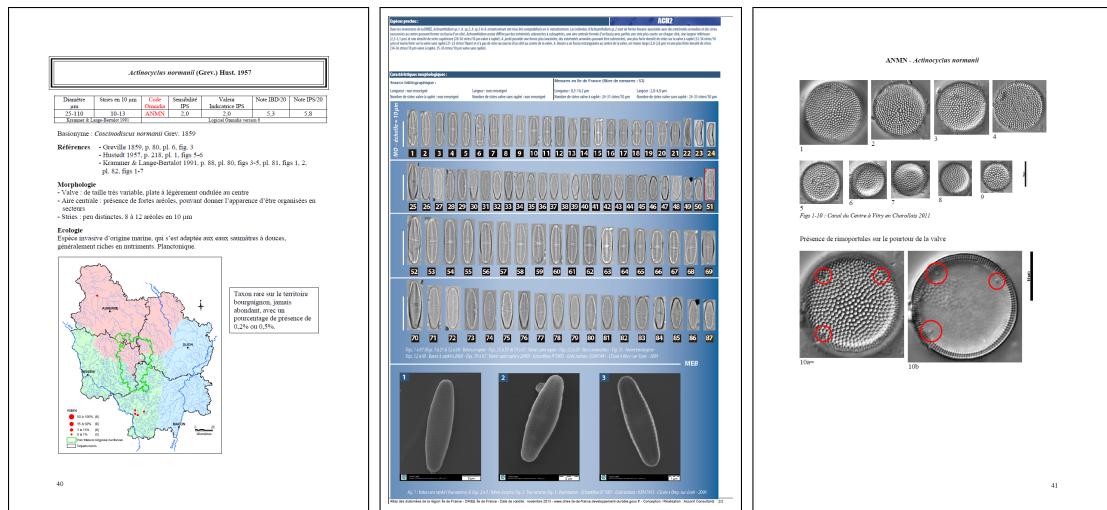


Figure 3.1. Example of diatom atlases pages

The main challenge of this process was to extract the right images with their respective labels, some atlas needing extensive segmentation tasks and many filters to reduce manual post-processing. Among the atlases we found, some would not have been profitable enough to extract and, given the project time-length, we settled on the following three: Rhône-Alpes [11], Ile-De-France [13] and Bourgognes [12]. Finally, we eliminated the remaining extraction errors visually (scale bars, embedded images, icons, logos...).

We also want to highlight the interest of using such sources. Indeed, by using diatom atlases, we are using already made images of great quality which would have been very time-consuming to make otherwise. Giving a new usage to such documents seems to us very beneficial to the scientific community as their public availability allows to

facilitate researches in the field of bioinformatics.

Content

Our final dataset consists in 9230 grayscale images of 206 taxa and its distribution is plotted in Figure 3.2. It is highly unbalanced with a median of 51 samples per taxon. If this property is generally unwanted in machine learning datasets, manually gathering the samples from microscopic slides generally leads to same results as diatoms are not evenly represented in water. Hence, working with this dataset puts us closer to a real life situation and we will discuss our methods for each of the applications. In Figure 3.3, one can observe five samples of different taxa extracted from the dataset.

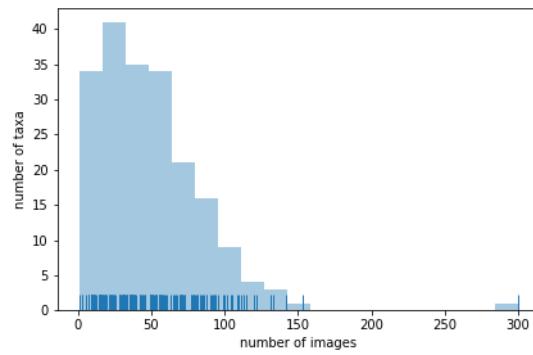


Figure 3.2. Histogram of the dataset distribution *On the abscissa, number of images in the dataset grouped in bins. In ordinates, number of taxa with this number of images. Blue ticks at the bottom indicate the real number of image for each taxon.*

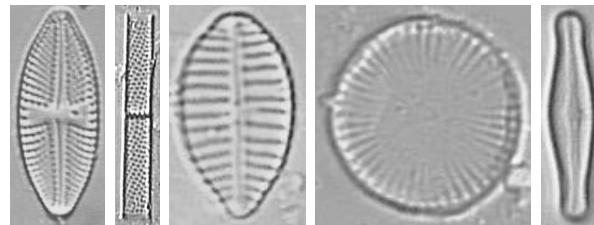


Figure 3.3. 5 samples extracted from the Rhône-Alpes atlas (light microscopy). Associated species listed from left to right: *Luticola goeppertiae*, *Aulacoseira granulata* var. *angustissima*, *Gomphonema parvulum* var. *parvulum*, *Cyclostephanos invisitatus*, *Achnanthidium catenatum*

3.1.2 Debris

To create the synthetic images of microscope slides, we need to take into account the various occurrences of debris generally present on microscope images. Therefore, we manually gathered 600 debris samples of various sizes, shapes and illuminations to

cover as many cases as possible. Figure 3.4 shows a sample of this dataset that we will name "Debris" in this paper.

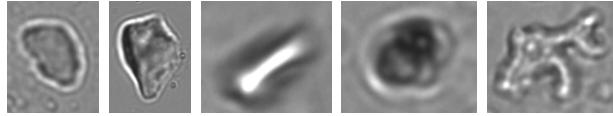


Figure 3.4. 5 samples extracted from the Debris dataset

3.1.3 DIC

To fine-tune and evaluate our models, we built a set of 138 real microscope images using DIC illumination and X1200 magnification. We then labeled every "full" diatom, avoiding the broken ones or those with too little area visible. A sample of the dataset is presented in Figure 3.5.

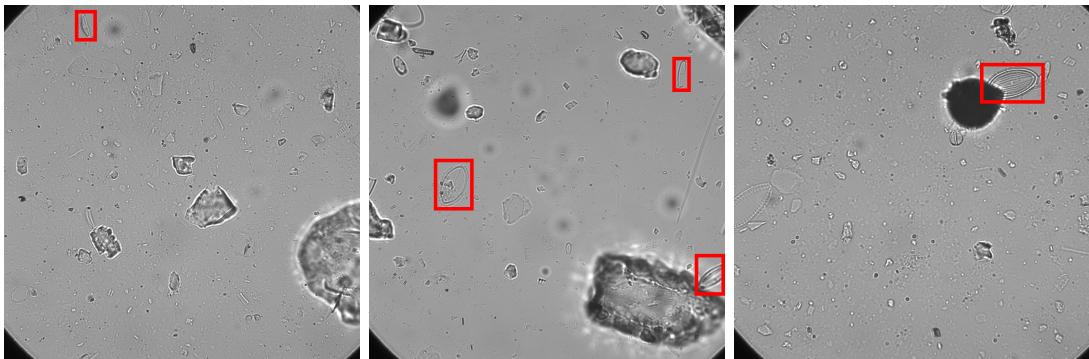


Figure 3.5. 5 samples extracted from the DIC dataset with labeled diatoms

3.2 Diatom classification

3.2.1 Aqualitas

In 2017, [4] and [33] proposed an update on diatom classification reaching respectively 98.11% of accuracy with morphological descriptors (bagging decision trees) and 99.51% of accuracy with CNNs (AlexNet). They achieved those scores with their own dataset created in partnership with the Institute of Optics (Spanish National Research Council). The full dataset was released on march 2020 under the name Aqualitas. [3] It is composed of 100 diatom taxa with about 100 specimen per taxon using oblique illumination and X600 magnification. A sample from the Aqualitas dataset can be seen in Figure 3.6.

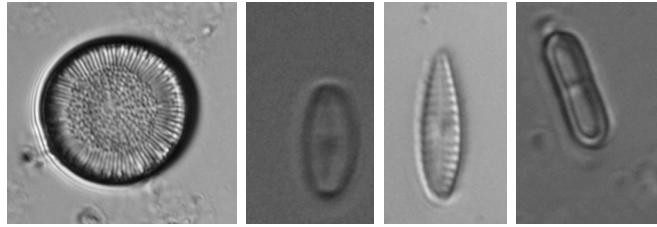


Figure 3.6. 4 samples extracted from the Aqualitas dataset (light microscopy). Associated species listed from left to right: *Cyclostephanos dubius*, *Eolimna rhombelliptica*, *Gomphonema micropumilum*, *Cyclostephanos invisitatus*, *Humidophila contenta*

3.2.2 ADIAC

ADIAC sets the first state of the art reference for automatic diatom identification and made available a robust public diatom dataset that we will name the "ADIAC" dataset in this paper [1]. A sample of this dataset is presented in Figure 3.7. This dataset has been splitted in (i) the public dataset [1] composed of 3400 images (2500 actually usable) and (ii) the restricted dataset which is no more available online. If the initial ADIAC studies used mainly three subsets composed of 6, 37 and 48 taxa, the restriction over some data forced the following studies to use new subsets. Among those studies, [10] used three subsets of 38, 48 and 55 taxa from the ADIAC public dataset that we will name respectively ADIAC38, ADIAC48 and ADIAC55. Those three datasets are fully detailed in their paper and we gathered the same subsets to make our classification comparisons.

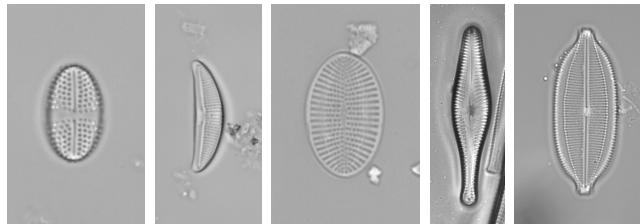


Figure 3.7. 5 samples extracted from the ADIAC dataset (light microscopy). Associated species listed from left to right: *Achnanthes parvula*, *Amphora veneta*, *Cocconeis peltoides*, *Gomphonema ventricosum*, *Gomphonema ventricosum*

Chapter 4

Methods

4.1 Diatom detection

4.1.1 Synthetic microscope image generation

We want to generate a synthetic microscope image I similar to Figure 4.1 from images of individual diatom or debris ("patches", eg. Figures 1.1,3.6,3.4) contained in the diatom and debris datasets (D_{diatom} and D_{debris} , respectively).

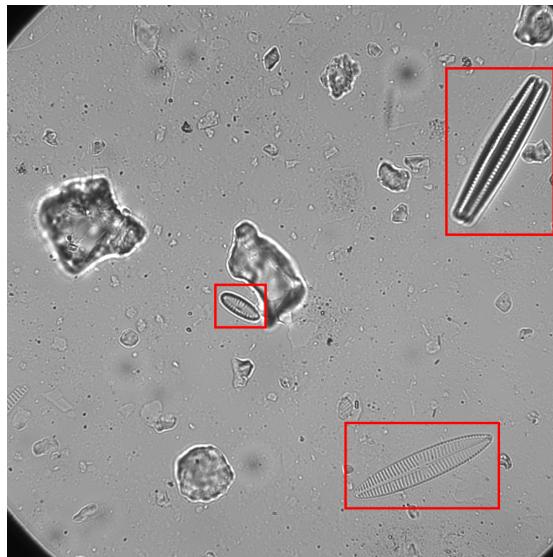


Figure 4.1. Example of real microscope image with labeled diatoms (DIC, X1200)

First, let's define a generic procedure $P(W, H, D, T_{ref})$ to create a seamless image I of width W and height H with patches from datasets $\{D_1, \dots, D_n\} = D$ and a reference patch T_{ref} :

1. Init I' , empty image of dimensions (W, H) . For each dataset $D_i \in D$,
 - Pick a subset $S_i \subset D_i$ composed of X_i patches with $X_i \sim U(a_i, b_i)$, a_i and b_i being respectively the minimal and maximal numbers of patches possible from D_i on the final image.
 - For each patch $T_{ij} \in S_i$ of width w_{ij} and height h_{ij} ,

- Use histogram matching to uniformize brightness and contrast of T_{ij} to reference image T_{ref} .
 - Resize T_{ij} while preserving the aspect ratio such that $\text{argmin}(w_{ij}, h_{ij}) = s_i + \epsilon$, with s_i being the minimal size possible for the shortest border of T_{ij} and $\epsilon \sim \text{Exp}(\lambda_i)$ with λ_i a parameter.
 - Apply data augmentation (rotation, flip, blur...).
 - Find a random position P_{ij} for patch T_{ij} on I' such that it does not overlap any previously added patch or the borders of I' more than a certain threshold percentage. This condition is crucial to avoid hidden subjects that would bias training and evaluation.
2. Init I , empty image of dimensions (W, H) . Let B be the set of pixels $p^{I'}$ of I' composing the edges delimiting $I'_{\neg empty}$ and I'_{empty} . Set the value of each pixel p_i^I of I as a weighted arithmetic mean such that:

$$p_i^I = \frac{\sum_{p_j^{I'} \in B} val(p_j^{I'}) w(p_i^I, p_j^{I'})}{\sum_{p_j \in B} w(p_i^I, p_j^{I'})} \quad (4.1)$$

$$w(p, p') = e^{-\frac{\|p' - p\|_2}{\lambda}} \quad (4.2)$$

where $val(p)$ is the gray value of pixel p and λ a parameter.

3. Finally, randomly pick a patch T_{ij} and place it at P_{ij} on I using the Poisson Editing method for seamless blending [34] until there is no patch left. This way, we shuffle the order of the patches on I .

Therefore, by running P with $D = \{D_{diatoms}, D_{debris}\}$ and the right set of parameters, we can create a synthetic image containing a mix of diatoms and debris as illustrated in Figure 4.2. For the rest of this study, for D_{diatom} and D_{debris} we will respectively use the "Atlas" and "Debris" datasets. In addition, we will generate images of 1000px by 1000px (W and H) using the reference image T_{ref} (also visible in 4.2) which has been chosen visually for its good contrast and brightness. Examples of generated synthetic images are visible in Figure 4.3.

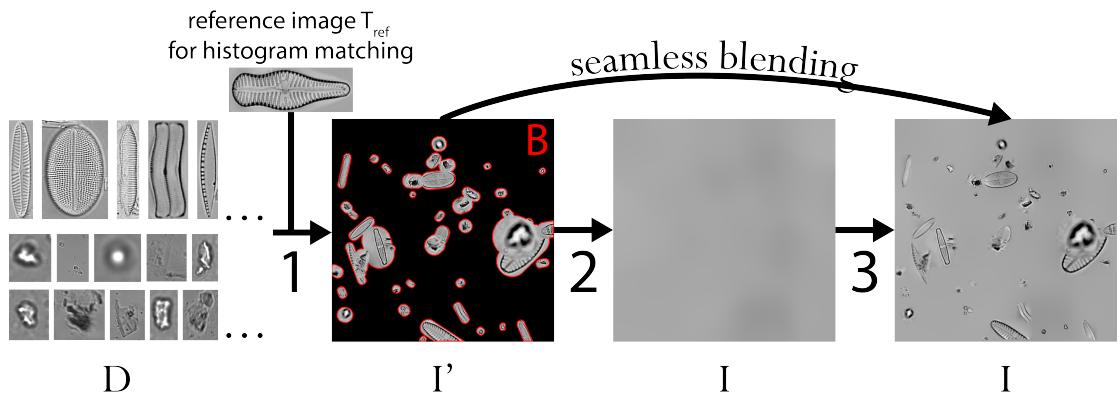


Figure 4.2. Illustration of the artificial image generation procedure P. (1) Patches are firstly selected randomly from the debris and diatom datasets and placed at a random position on a first image I' after applying histogram matching with T_{ref} . (2) Then we use borders B to create a smooth gradient on a new image I (3) to finally place all the diatoms in random order using seamless blending [34].

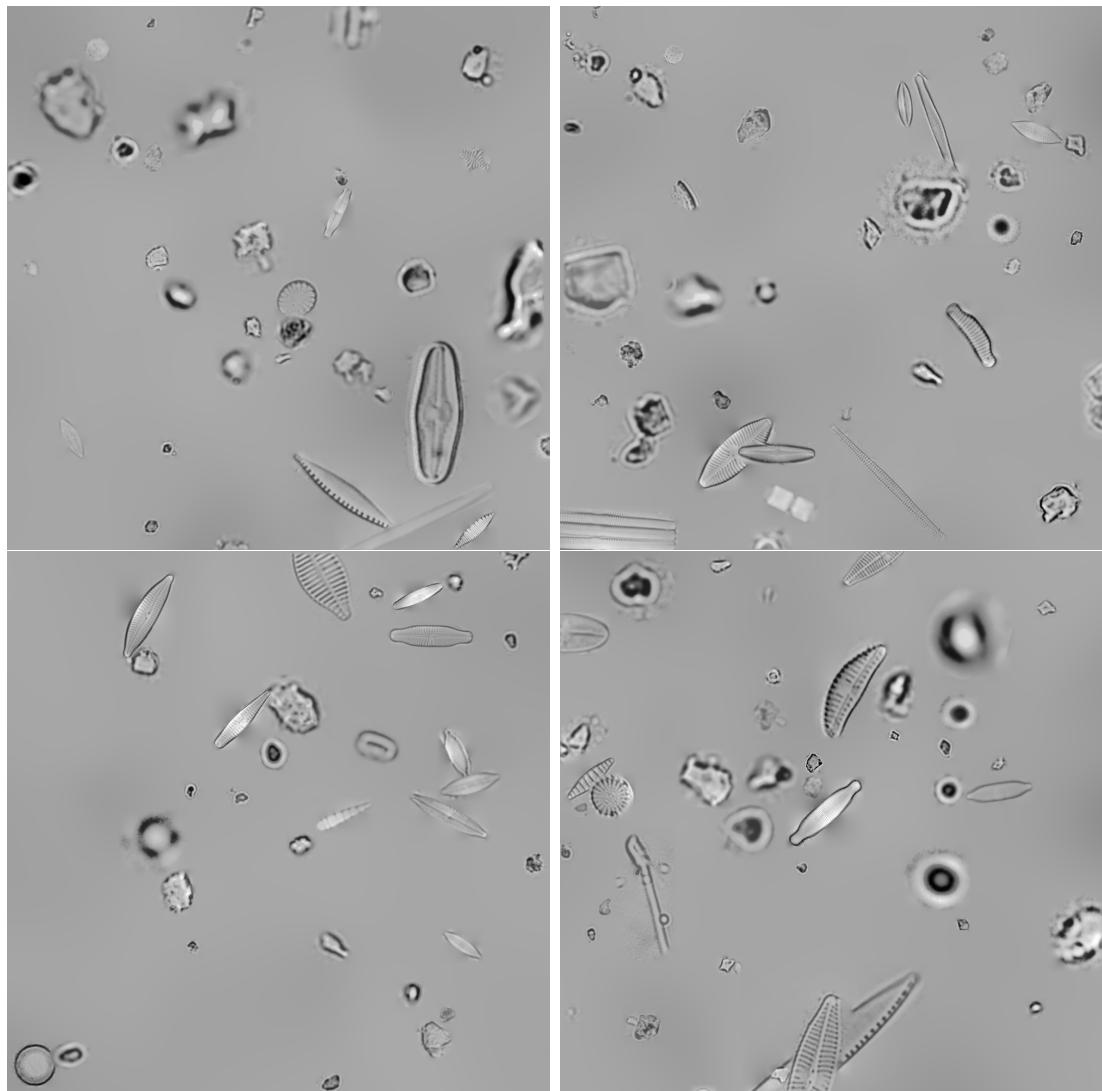


Figure 4.3. Examples of synthetic images

Our goal is to be able to generate thousands of images easily with various sets of parameters, making the generation time crucial. Indeed, we found that the implementation of P can critically influence the running time. To generate one image, our first implementation ran in about 180s when after our diverse improvements, that time fall to 1.6s on average. The two major improvements we made are the use of GPU for matrix operations (CuPy) and a weight map to store values of w in advance.

4.1.2 Datasets splitting

Note *For the sake of clarity, datasets of real and synthetic images containing multiple diatoms will be denoted respectively D^{*r} and D^{*s} .*

Synthetic For training and evaluation, we need to generate two distinct synthetic datasets: one for training D_{train}^{*s} (36,000 artificial images) and one validation D_{val}^{*s} (4,000 artificial images). Firstly, we split our patch datasets keeping 90% of the data for the generation of D_{train}^{*s} and 10% for D_{val}^{*s} such that $D_{diatoms} = D_{diatoms}^{90\%} \cup D_{diatoms}^{10\%}$ and $D_{debris} = D_{debris}^{90\%} \cup D_{debris}^{10\%}$. Then we generated 36,000 images for D_{train}^{*s} and 4,000 for D_{val}^{*s} using respectively $P(1000, 1000, \{D_{diatoms}^{90\%}, D_{debris}^{90\%}\}, T_{ref})$ and $P(1000, 1000, \{D_{diatoms}^{10\%}, D_{debris}^{10\%}\}, T_{ref})$.

Real To train and evaluate our models using the "DIC" dataset, we need to split it first. As the we have a limited number of images, we will use a 60/40 rule, keeping 60% of the images (83 images) for training (D_{train}^{*r}) and 40% (55 images) for evaluation (D_{val}^{*r}).

4.1.3 Training and evaluation

For training, we used the Object Detection API of Tensorflow to train a Faster R-CNN with a ResNet101 backbone, an architecture that we will abbreviate with FR101. To avoid unnecessarily long training times, we took advantage of the common transfer learning technique of fine-tuning, starting with a pretrained FR101 model and adjusting it with our own dataset. The Object Detection API proposes such a model pretrained on the COCO dataset that we will denote by M_{coco} . From there, only a few epochs (complete pass through the training data) are needed to obtain the optimal model. During training, the images are augmented using random horizontal and vertical flips.

To support the use of a synthetic dataset in addition to real images, we will compare the results of two pipelines: (1) in Pipeline 1, the model is first trained on synthetic images and real images are then used for fine-tuning, (2) Pipeline 2 is the control pipeline trained only with the real images. Our goal is to demonstrate that Pipeline 1 can obtain better scores than Pipeline 2, therefore proving that a synthetic dataset can be beneficial in the frame of diatom detection.

Pipeline 1 The pretrained M_{coco} model is first trained with the synthetic images from D_{train}^{*s} until we get the best validation metrics for D_{val}^{*s} . We will call this new model M_{gen} as it is a generic model, not specialized in any type of real microscope images. Subsequently, our goal is to specialize the model with real images such as the ones presented in the "DIC" dataset. For this purpose, we will train M_{gen} with D_{train}^{*r} until we get the best metrics for D_{val}^{*r} . We will denote this model by M_{spe}^1 as it is the model specialized for the "DIC" dataset created by Pipeline 1. It is essential to avoid that the second training on real images wipes out the first one made on synthetic ones. To ensure this, the first training needs to be made on far more iterations than the second one. In practice, we trained for about 100000 iterations with the synthetic dataset and 1000 with real one.

Pipeline 2 This control Pipeline 2 is only trained with the real images from the "DIC" dataset. We directly train the M_{coco} model with D_{train}^{*r} until we get the best metrics for D_{val}^{*r} . This second specific model will be denoted by M_{spe}^2 .

4.1.4 Metrics

Most of the metrics for evaluating a classification model are based on precision and recall. Precision measures the proportion of objects that are correctly classified by the model (exactness) when recall denotes the proportion of objects that are actually retrieved by the model (completeness). They can be expressed as such:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

with True Positive (TP), False Positive (FP), False Negative (FN).

However, the question of matching the model's detections with the ground truth or the threshold over the detection scores are non-trivial challenges specific to OD. Therefore, precision and recall alone do not take all the subtleties of OD into account and additional metrics have been designed. Among them, the Intersection over Union (IoU) measures the overlap percentage between two bounding boxes. Therefore, detected and ground truth bounding boxes will be considered as a match if their IoU is above a certain threshold (see Figure 4.4). By using together multiple metrics such as precision, recall or IoU, new indicators like Average Precision (AP) or Average Recall (AR) have been created. AP and AR compute respectively the common precision and recall metrics as presented above as the mean over various thresholds of IoU and detection scores.

If some metrics such as the ones mentioned above have become standard, their number and implementations may vary. Sets of metrics with their implementation are generally defined by OD challenges (VOC, COCO, Open Images...) and become commonly used afterwards. In this paper, we will focus on the COCO's metrics as they are comprehensive and have became a reference in the field. Here are the six COCO metrics we will use with their official description [7]:

- $AP_{IoU=0.50:0.95}$: AP over 10 IoU thresholds and considered as the primary metric of the challenge
- $AP_{IoU \geq 0.50}$: AP with $IoU \geq 0.5$
- $AP_{IoU \geq 0.75}$: AP with $IoU \geq 0.75$
- $AR^{max=1}$: AR given at most 1 detection per image
- $AR^{max=10}$: AR given at most 10 at max detections per image
- $AR^{max=100}$: AR given at most 100 at max detections per image

To compute those metrics, we will use the Object Detection API's implementations.

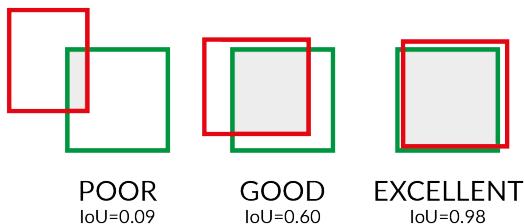


Figure 4.4. Illustration of the IoU measure. *The green bounding box can be seen as the ground truth and the red as the detected one.*

4.2 Diatom classification

4.2.1 Datasets

We conducted our experiments on the three datasets "Atlas", "Aqualitas" and "ADIAC38/48/55". Those datasets, as a reminder, contain images of single diatoms and an example is given in Figure 3.3. If the results we obtained on the "Atlas" dataset are new and therefore no comparison is possible, it is also the most challenging one with the greater number of different taxa. By also using the "Aqualitas" and "ADIAC" datasets, our goal is to make a comparison with previous works, respectively [4] [33] and [10].

For greater generalization and to compensate the uneven number of images per taxa in the datasets, we used data augmentation. Using Keras's `ImageDataGenerator`, images are augmented randomly directly during the training process, limiting overfitting as the network never sees twice the exact same augmented image. For all our tests, here are the augmentations we used: *rotation, horizontal and vertical flips, horizontal and vertical shifts, zooms and brightness variations*. All those augmentations are applied randomly on each incoming image as the training goes on. Shifts and zoom are interesting in our particular case as on cropped images by hand or an object detection system like FR101, diatoms can be highly off-centre or different-sized.

For some of the trainings (see (2) in the results table 5.2), we also used data augmentation to compensate for the uneven number of images per taxa in the datasets. To do that, we offset the number of images per taxon in the training datasets to match the one with highest number of images. We did that by duplicating the inputs but, as data augmentation is applied, the repeated images were never exactly the same when training the model.

As diatoms images are initially rectangles, applying rotations would crop them. To prevent that, we first normalized all the images in 256 by 256px squares as illustrated in Figure 4.5b. In Figure 4.5c, you can also observe a batch of nine randomly augmented images.

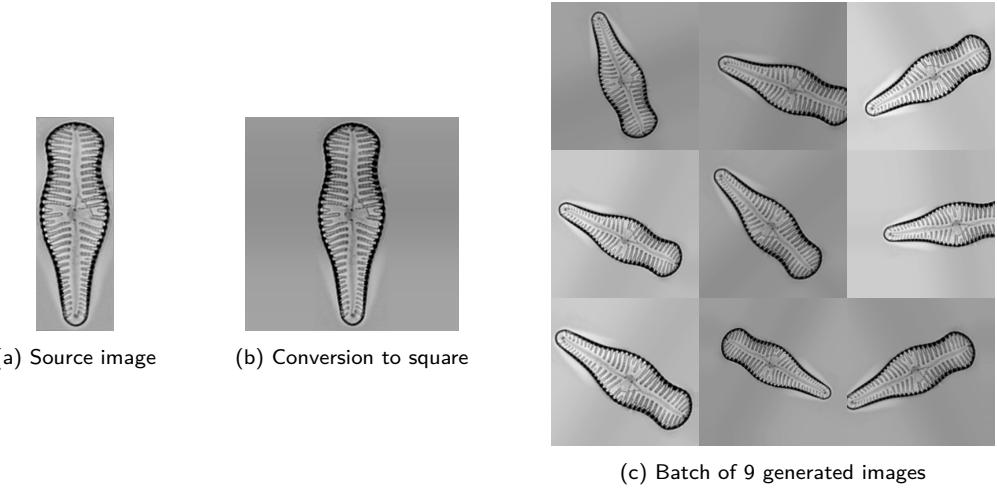


Figure 4.5. Illustration of the data augmentation pipeline. *Images are first converted to square and then randomly augmented. Possible augmentations are: rotation, horizontal and vertical flips, horizontal and vertical shifts, zooms and brightness variations.*

4.2.2 Training and evaluation

Relying once more on fine-tuning, we trained an Xception network pretrained on the ImageNet dataset. For all our tests, we used 10-fcv to match previous evaluation techniques, notably used by [4, 10, 33]. Note that every split of the dataset is made in a *stratified* fashion, meaning that when splitting the dataset in 90%/10%, each class will splitted independently following this rule.

The only modifications we made on the Xception network are the input layer (to accept 256 by 256 images) and the output layer (to output the right number of classes). The training is divided in two phases: a few epochs on the new added layers only and the rest on the full model.

4.2.3 A hierarchical approach

In this section, we will describe a method to create an artificial taxonomy based on visual similarity. When a model M is evaluated with a test dataset D , the error is generally not spread uniformly and some sets of classes tend to get mixed up more easily. To find those clusters grouping classes with high confusion, one common way is to use a clustering algorithm on the associated similarity matrix. As the raw output of such evaluations is generally a confusion matrix, we will first describe the process used to convert this confusion matrix in a similarity matrix (based on [37]) and then

we will apply a hierarchical clustering algorithm to create the artificial taxonomy.

In view of the duration of the project, we will not apply this taxonomic approach to a concrete classification scheme. However, we will discuss the possible approaches and this will serve as a foundation for the following PHD.

Similarity Matrix

If we consider the confusion matrix C associated to a set of predictions $\langle y_{pred}, y \rangle$, here is the procedure to create a similarity matrix S :

1. Normalization: Divide each row of C by the total sum of the row to create "percentages" of confusion.

$$C_{ij}^{norm} = \frac{C_{ij}}{\sum_{i=1}^n C_{ij}}$$

2. Class overlap: If C_{ij}^{norm} and C_{ji}^{norm} represent two different types of errors (the direction of misclassification), this distinction is not relevant in terms of taxonomy and we just want to extract the information that class i is similar to class j . To do so, we will replace C_{ij}^{norm} and C_{ji}^{norm} by the mean of the two terms as such:

$$C_{ij}^o = \frac{C_{ij}^{norm} + C_{ji}^{norm}}{2}$$

3. Diagonal: An other essential step is to set the diagonal values to 1 as we obviously want the classes to be 100% similar to themselves.

$$C_{ij}^{o'} = \begin{cases} 1, & \text{if } i = j \\ C_{ij}^o, & \text{otherwise} \end{cases}$$

4. Similarity matrix: As the confusion denotes for error, we want the opposite for similarity. Therefore, S can be expressed as:

$$S = 1 - C^{o'}$$

Artificial taxonomy

Once the similarity matrix S is computed, we will use a clustering algorithm to group similar classes together. To create an artificial taxonomy based on similarity, the most flexible option is to use a hierarchical algorithm like the classic linkage algorithms (complete, single, average...). With their bottom-up approach, they can be used at any level of the generated hierarchy depending on the application. In the case of our project, the initial objects are the row (or columns) of the similarity matrix.

For this project, we will use a derivative of the average-linkage algorithm using the Ward's method (Scipy implementation). This iterative approach is an optimization process to reduce the clusters' variance. Each iteration, the objective distance function is given by:

$$d(u, v) = \sqrt{\frac{|v| + |s|}{T} d(v, s)^2 + \frac{|v| + |t|}{T} d(v, t)^2 - \frac{|v|}{T} d(s, t)^2}$$

with v a cluster, u a cluster composed of s and t (joined in a previous iteration) and $T = |v| + |s| + |t|$ the total cardinality. This very popular method is known to give great results, especially as it takes the previous clusters (s and t) in consideration to associate a new one.

Potential applications

In following works on the subject, one goal will be to exploit this generated taxonomy to create an adaptive classification scheme. The simpler approach possible would be to use local classifiers at each node of the tree. Those classifier would become more and more specialized and if the scores are not convincing enough, we just have to stop the classification at an upper level. On the negative side, this method can be really heavy on computing depending of the size of the local classifiers (using Xception at every step might be a bad idea...). Other approaches propose to use this hierarchy with a flat classifier. Among the many ways to do it, one will find the use of conditional probabilities or multi-label classification. Finally, the lead of custom network architectures (hierarchical CNN models) could be effective but would also time-consuming to design.

Chapter 5

Results and discussion

5.1 Diatom detection

The computed metrics for diatom detection are presented in Table 5.1.

Table 5.1. Results the diatom detection (6 COCO's metrics)

Evaluation name	Pipeline 1			Pipeline 2
	A	B	C	D
Evaluated model	M_{gen}	M_{gen}	M_{spe}^1	M_{spe}^2
Evaluation dataset	D_{val}^{*s}	D_{val}^{*r}	D_{val}^{*r}	D_{val}^{*r}
Type	Synthetic	Real	Real	Real
#images	4000	55	55	55
$AP_{IoU=0.50:0.95}$	0.897	0.223	0.662	0.553
$AP_{IoU \geq 0.50}$	0.989	0.565	0.904	0.788
$AP_{IoU \geq 0.75}$	0.970	0.127	0.823	0.700
$AR^{max=1}$	0.099	0.167	0.355	0.321
$AR^{max=10}$	0.889	0.387	0.757	0.704
$AR^{max=100}$	0.918	0.419	0.765	0.711

If we gave all the metrics for full disclosure, we feel it is important to point out that they are not to be considered all equally. Notably, the average recall using a maximum of 1 detection ($AR^{max=1}$) makes no sense as almost every image in the "DIC" dataset contains more than 1 diatom and synthetic images contain all between 10 and 15 diatoms (making also $AR^{max=10}$ irrelevant in that case). Therefore, the $AR^{max=100}$ is for us the most relevant metric in terms of recall. Moreover, it is interesting to analyze the roles of the three precision metrics. The higher the IoU, the more accurately we want our bounding box to frame the diatom. If high IoU is sometimes needed, it is not a major issue in diatom detection for two main reasons: (1) very effective segmentation algorithms exist to detect diatom contours and reframe

if needed and (2), the major application of such systems being classification (manual or automatic), different-sized and off-centered diatoms are not an issue neither for human eye nor for machine learning algorithms thanks to data augmentation. Taking all these elements into account, it seems to us that $AP_{IoU \geq 0.50}$ or $AP_{IoU \geq 0.75}$ are sufficiently relevant metrics for our work.

Pipeline 1 With evaluation \boxed{A} , we see that the model learned to exploit synthetic images very well. We obtain very good precision and recall on a dataset which have been specifically designed to be hard with good amounts of overlapping and highly blurred or desaturated diatoms. This particularly high precision means that the model learned to generalize very well over the differentiation of debris and diatoms. With evaluation \boxed{B} , we observe however that the generic model M_{gen} could not be used as such for diatom detection in that case. Indeed, diatoms from the "Atlas" dataset are visually too different from the ones present in the "DIC" dataset, especially as it is not the same illumination technique (brightfield against DIC). This is also why, in this particular case, the fine-tuning phase turns out to be essential. On a different dataset of real images, we could imagine better scores directly with M_{gen} . Finally, evaluations \boxed{C} shows a great improvement over \boxed{B} with up to 90% of precision and 76% of recall.

Pipeline 2 Evaluation \boxed{D} also shows good performances despite being trained with only 83 microscope images with up to 79% of precision and 71% of recall. Those scores offer a great perspective on the possibilities offered by such architectures for microorganism detection.

When comparing evaluations \boxed{C} and \boxed{D} , The most important gain from Pipeline 1 over Pipeline 2 is the precision with a significant gain of about 12%! It makes sense as one the major advantage of the synthetic dataset is the possibility to present the diatoms and debris with various data augmentation (rotation, scale, contrast, brightness...), training the model with more diversified examples helping it better generalize. When training directly with real microscope images like the ones from "DIC", you are really constrained in terms of image variety with really few possible augmentations (mainly flips and 90° rotations). With a gain of about 5%, the recall has also

been improved with the synthetic training. This is thanks to the wide diversity of diatoms present in the diatom dataset used to create the synthetic images. However, the "Atlas" diatoms are markedly different visually from the one present in the "DIC" dataset. By completing this diatom dataset with more diverse examples from various illumination techniques, we think that this score could get even better.

A visual comparison is presented in Figure 5.1. On the four examples, the groundtruth is presented on the left, detections of Pipeline 1 in the middle and detections of Pipeline 2 on the right. Firstly, we note that both pipelines obtain pretty good results, especially Pipeline 2 given the small amount used for training. In Figure 5.1a, we observe that Pipeline 2 missed a diatom. After studying the "DIC" training images, it appears that this type of diatom is almost absent from the training set, making it almost impossible to identify for Pipeline 2. However, it is present in the "Atlas" dataset, explaining notably the better recall. Figure 5.1b shows that the numerous overlappings present in the synthetic dataset have helped the network to better handle them and Figure 5.1c is an example of accuracy improvement as we see that Pipeline 2 confused a debris for a diatom. Finally, Figure 5.1d is also an example of recall improvement. It is also important to note that the hand-labelling we made can be subject to discussion. Should we label very tiny diatom even though they are unidentifiable (taxon-wise)? Should diatoms partially hidden or broken (at what percentage?) be identified? Given that diatomists themselves do not always agree on those questions, we tried to optimize consistency among our labelizations, emphasizing a possible later classification. Therefore, we made the decision to label any diatom that we could identify, at the condition it was sufficiently visible and complete. If we consider the small diatom on the groundtruth of Figure 5.1 that no pipeline has been able to detect, its size can make it hard to identify but we still considered it was worth including. The broken diatoms detected by Pipeline 2 on Figures 5.1c and 5.1d are still diatoms but we just considered that they were too broken to be considered, which is again a classification-oriented judgment.

As a prospect, an other possible application of our synthetic imaging system could be as a data augmentation procedure on hand-labelled image. If we take the example of the "DIC" dataset, we could manually extract the diatoms and debris from the

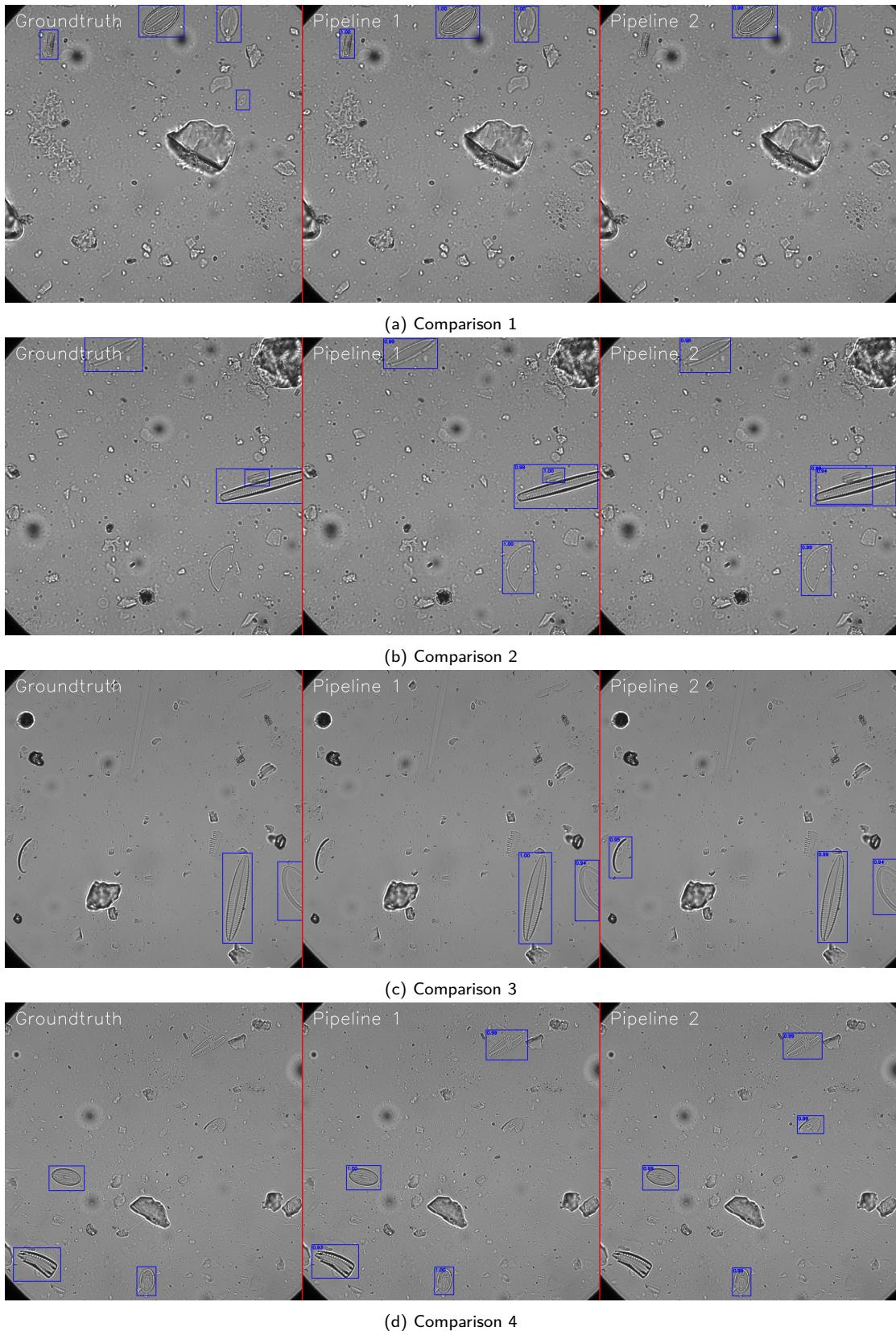


Figure 5.1. Comparisons of the 2 diatom detection pipelines with the groundtruth

training dataset and use them to create a great variety of synthetic images. This method would be different from the one shown in this project as the generated model

from the synthetic dataset would be directly specific and not generic. This interesting lead could be followed in a future study.

5.2 Diatom classification

The results of diatom classification are presented in Table 5.2.

Table 5.2. Results of diatom classification. The \pm denotes for standard deviation. (1) If "Yes", the dataset has been augmented to compensate for class unbalance as explained in the Methods section. (2) This score is discussed in the Results section.

	Atlas	Aqualitas	ADIAC55	ADIAC48	ADIAC38
#taxa	157	80	55	48	38
Median #images per taxon	51	94	20	20	21
Mean accuracy	0.9096 \pm 0.0192	0.9265 \pm 0.0160	0.9672 \pm 0.0223	0.9735 \pm 0.0131	0.9713 \pm 0.0203
Evaluation method	10fcv	10fcv	10fcv	10fcv	10fcv
Features	Xception	Xception	Xception	Xception	Xception
Classifier	Softmax	Softmax	Softmax	Softmax	Softmax
Solver	Adam	Adam	SGD	SGD	SGD
Balanced ⁽¹⁾	Yes	Yes	No	No	No
Previous best accuracy	Ø	0.9951 ⁽²⁾ [33]	0.9617 [10]	0.9715 [10]	0.9797 [10]
Evaluation method	Ø	10fcv	10fcv	10fcv	10fcv
Features and descriptors	Ø	AlexNet	30 Morphological + 200 Texture	30 Morphological + 200 Texture	30 Morphological + 200 Texture
Classifier	Ø	Softmax	Random forest	Random forest	Random forest

For the three ADIAC subsets, we get approximately the same scores as the ones presented in the original study. It reveals that Xception is able to produce high-level CNN features as least as good the handcrafted ones used in that case. Considering the similarity between our scores and the original ones, it even seems to us that we are reaching the limit score obtainable with this dataset. Nevertheless, the main advantage of using CNN networks for features extraction is the simplicity. As it is a fully automatic process, it removes the need of in-depth domain knowledge and makes the project easily reusable for an alternative dataset or other microorganisms. Of course, those networks require the use of GPUs for training and prediction but such hardware are now more and more common in laboratories.

On the three datasets, we obtained the best scores using the SGD solver with a batch size of 33 and constraining the data augmentation rotation between -20° and 20° (all the diatoms are more or less aligned in the ADIAC dataset).

As it is presented in Table 5.2, we got a score of 0.9265 for the "Aqualitas" dataset against 0.9951 in the original paper. If we performed less well, our evaluation technique differs and is, in our opinion, more representative of the actual performances. Indeed, in its original test, [33] performs balancing and data augmentation before splitting the dataset for 10fcv. Therefore, multiple augmentations of the same original image can be presented for training and evaluation which creates a bias. In our evaluation procedure, images are first splitted and then augmented, separating completely the training and evaluation datasets. As a result, both scores are not comparable. It is also important to mention that the training is performed on a subset of 80 taxa (over the 100 present in the dataset), similarly to their original paper. [33]

For Aqualitas, we obtained the best scores using the the ADAM solver with a batch size of 33, constraining the data augmentation rotation between -180° and 180° and by increasing slightly the horizontal and vertical shifts (Aqualitas diatom samples can be highly off-centre or different-sized).

Despite its notably higher number of taxa and a lower median number of images per class than "Aqualitas", the "Atlas" dataset get a reasonably good score with less than two points difference! We believe that this is due to the difference in quality between the two datasets. By comparing Figures 3.3 and 3.6, we observe that diatoms in the "Atlas" dataset are better defined and that we can distinguish a lot of details in the frustules. As such details are crucial to differentiate diatoms, the highest image quality means the better classification. However, we must bear in mind that comparing models made with different sets of taxon and images is difficult. Note that we only used 166 of the 206 taxa of the dataset as we eliminated taxa with too few images (strictly less than 20).

For Atlas, we obtained the best scores using the the ADAM solver with a batch size of 33 and constraining the data augmentation rotation between -180° and 180° .

Overall, we see that Xception does a really good job at diatom classification and can differentiate up to 166 taxa with a reasonable accuracy!

5.3 Hierarchical classification

In Figure 5.2a, you can observe the hierarchy generated by the Ward's Method on the 166 classes of the Atlas dataset illustrated as a dendrogram. Note that the red group far right gathers taxa which have been classified with no error. Therefore, they are all at the same similarity distance from each other.

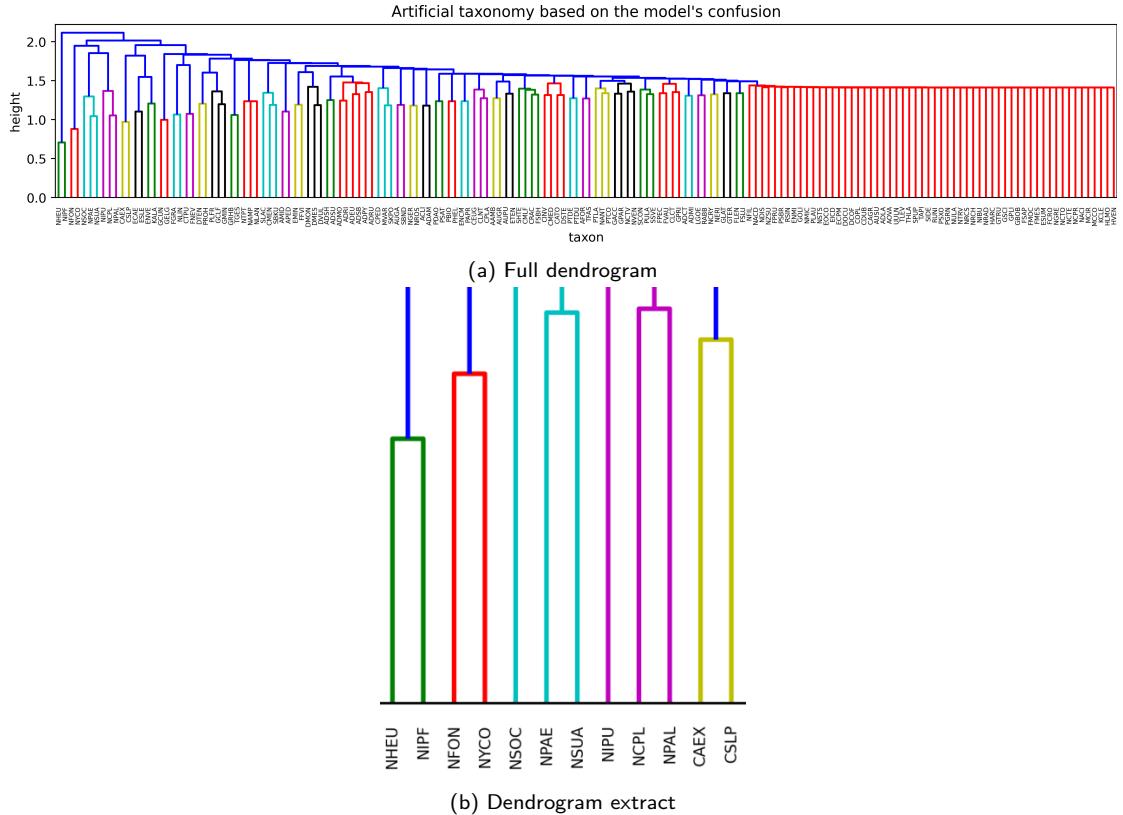
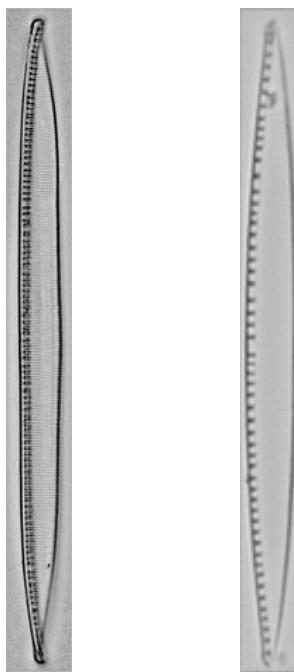


Figure 5.2. Full dendrogram and extract of the bottom-left part

To better analyze the results, we extracted a part of the dendrogram presented in Figure 5.2b as an example. The taxa are designated by their 4-letter symbol. The first letter of this symbol represents the genus of the taxon which is the highest level of the biological taxonomy. It is interesting to note that the clusters generally group together the diatoms from the same genus (**NHEU/NIPF**, **CAEX/CLSP...**), meaning that our artificial taxonomy partially joins the biological one. It does make sense as those taxonomic decisions are made, *inter alia*, on visual similarity.

In Figure 5.3 are presented images of diatoms belonging to the NHEU and NIPF taxa. As one can observe on Figure 5.2b, they have been grouped together by the

algorithm and those images illustrate greatly that those two taxa present significant visual similarities. However, there are still some distinctions and the number of images in the dataset is also an important restraining factor. Indeed, NHEU and NIPF have among the lowest number of images in the dataset, respectively 21 and 22 which is far below the median of 51! This is actually an important quality of this type of artificial taxonomy: if two classes are subject to confusion, it can be due to the visual similarity but also to the number of samples in the training data. Therefore, this taxonomy can be used for an adaptive classification scheme and evolve as the dataset expands.



(a) NHEU diatom (b) NIPF diatom

Figure 5.3. Images of diatoms from NHEU and NIPF diatoms

Chapter 6

Conclusion

This project addresses the two main tasks of diatom detection and classification using deep learning architectures and we hope the results will prove to be useful for the following PhD.

For the detection, besides showing that deep-learning OD architectures can be used successfully to detect diatoms on microscope images, we also presented a novel approach using a synthetic dataset in pair with a real one allowing to gain more than 10% of precision 5% of recall.

To illustrate a use case of such system, we applied the latest improvements in image classification to datasets used in previous studies and a new one created for this project. We obtained results as good as the original ones made using non-CNN approaches, making diatom classification more widely available as it does not require in depth domain knowledge. We also discussed the evaluation process of a previous study and proposed our own score which is, to our eye, less subject to bias. Finally, we have shown with our own dataset that latest image classifiers can distinguish up to 166 taxa (more than twice the highest previous number in literature) with a reasonable accuracy of 91%!

In addition, we proposed a solution to generate an artificial taxonomy based on the model's classification mistakes which can be used to create an adaptive classification scheme based on class difficulty, dataset completeness and model bias.

REFERENCES

- [1] ADIAC. Public data adiac project. https://rbg-web2.rbge.org.uk/ADIAC/pubdat/downloads/public_images.htm, 2002. Accessed: 2020-04-20.
- [2] BENOISTON, A.-S., IBARBALZ, F., BITTNER, L., GUIDI, L., JAHN, O., DUTKIEWICZ, S., AND BOWLER, C. The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (09 2017), 20160397.
- [3] BUENO, C. S., BLANCO, S., BUENO, G., BORREGO-RAMOS, M., AND CRISTOBAL, G. Aqualitas Database (full release), 3 2020.
- [4] BUENO, G., DENIZ, O., PEDRAZA, A., RUIZ-SANTAQUITERIA, J., SALIDO, J., CRISTOBAL, G., BORREGO-RAMOS, M., AND BLANCO, S. Automated diatom classification (part a): Handcrafted feature approaches. *Applied Sciences* 7 (07 2017), 753.
- [5] CAIRNS, J., ALMEIDA, S. P., AND FUJII, H. Automated identification of diatoms. *BioScience* 32, 2 (1982), 98–102.
- [6] CHOLLET, F. Xception: Deep learning with depthwise separable convolutions, 2016.
- [7] COCO. Coco - evaluate - metrics. <http://cocodataset.org/#detection-eval>, 2019. Accessed: 2020-05-05.
- [8] COSTE, M., BOUTRY, S., TISON-ROSEBERY, J., AND DELMAS, F. Improvements of the biological diatom index (bdi): Description and efficiency of the new version (bdi-2006). *Ecological Indicators* 9, 4 (2009), 621 – 650.
- [9] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 1, pp. 886–893 vol. 1.

- [10] DIMITROVSKI, I., KOCEV, D., LOSKOVSKA, S., AND DŽEROSKI, S. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informatics* 7, 1 (2012), 19 – 29.
- [11] DREAL. Atlas des diatomées. <http://www.auvergne-rhone-alpes.developpement-durable.gouv.fr/atlas-des-diatomees-a3480.html>, 2014. Accessed: 2020-04-28.
- [12] DREAL. Atlas des diatomées des cours d'eau du territoire bourguignon. <http://www.bourgogne-franche-comte.developpement-durable.gouv.fr/atlas-des-diatomees-des-cours-d-eau-du-territoire-a7004.html>, 2017. Accessed: 2020-04-28.
- [13] DRIEE. Atlas des diatomées. <http://www.driee.ile-de-france.developpement-durable.gouv.fr/atlas-des-diatomees-a2070.html>, 2014. Accessed: 2020-04-28.
- [14] DU BUF, H., AND BAYER, M. *Automatic Diatom Identification*. Series in machine perception and artificial intelligence. World Scientific, 2002.
- [15] DU BUF, J., SHAHBAZKIA, H., CIOBANU, A., BAYER, M., DROOP, S., HEAD, R., JUGGINS, S., FISCHER, S., BUNKE, H., WILKINSON, M., ROERDINK, J., PECH PACHECO, J. L., AND CRISTOBAL, G. Diatom identification: A double challenge called adiac. *Proceedings - International Conference on Image Analysis and Processing, ICIAP 1999* (01 1999), 734–739.
- [16] FELZENZWALB, P., MCALLESTER, D., AND RAMANAN, D. A discriminatively trained, multiscale, deformable part model. vol. 8:.
- [17] FREITAS, A., AND DE CARVALHO, A. A tutorial on hierarchical classification with applications in bioinformatics. *Research and Trends in Data Mining Technologies and Applications* (01 2007).
- [18] GELAS, A., MOSALIGANTI, K., GOUAILLARD, A., SOUHAIT, L., NOCHE, R., OBHOLZER, N., AND MEGASON, S. G. Variational level-set with gaussian shape model for cell segmentation. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (2009), pp. 1089–1092.

- [19] JALBA, A. C., WILKINSON, M. H., AND ROERDINK, J. B. Automatic segmentation of diatom images for classification. *Microscopy Research and Technique* 65, 1-2 (2004), 72–85.
- [20] JENSEN, S. N., IRANI, R., MOESLUND, T. B., AND RANKL, C. General purpose segmentation for microorganisms in microscopy images. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)* (2014), vol. 1, pp. 690–695.
- [21] JOHNSTON, E.M. STOERMER, E. Computer analysis of phytoplankton cell images. *The Microscope* (1976), 663–665.
- [22] KLOSTER, M., KAUER, G., AND BESZTERI, B. Sherpa: An image segmentation and outline feature extraction tool for diatoms and other objects. *BMC bioinformatics* 15 (06 2014), 218.
- [23] KOHEI ARAI, S. K. Advances in computer vision. *Advances in Intelligent Systems and Computing* (2020).
- [24] KOSOV, S., SHIRAHAMA, K., LI, C., AND GRZEGORZEK, M. Environmental microorganism classification using conditional random fields and deep convolutional neural networks. *Pattern Recognition* 77 (2018), 248 – 261.
- [25] KUANG, Y. Deep neural network for deep sea plankton classification.
- [26] LAVOIE, I., HAMILTON, P. B., MORIN, S., TIAM], S. K., KAHLERT, M., GONÇALVES, S., FALASCO, E., FORTIN, C., GONTERO, B., HEUDRE, D., KOJADINOVIC-SIRINELLI, M., MANOYLOV, K., PANDEY, L. K., AND TAYLOR, J. C. Diatom teratologies as biomarkers of contamination: Are all deformities ecologically meaningful? *Ecological Indicators* 82 (2017), 539 – 550.
- [27] LI, Q., SUN, X., DONG, J., SONG, S., ZHANG, T., LIU, D., ZHANG, H., AND HAN, S. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES Journal of Marine Science* (09 2019). fsz171.
- [28] MEIJERING, E. Cell segmentation: 50 years down the road [life sciences]. *IEEE Signal Processing Magazine* 29, 5 (2012), 140–145.

- [29] MORIN, S. Bioindication des effets des pollutions métalliques sur les communautés de diatomées benthiques. approches *in situ* et expérimentales. 23.
- [30] NIKOLENKO, S. I. Synthetic data for deep learning, 2019.
- [31] ODABAI FARD, S. H. *Efficient multi-class objet detection with a hierarchy of classes*. Theses, Université Blaise Pascal - Clermont-Ferrand II, Nov. 2015.
- [32] PAPPAS, J., KOCIOLEK, P., AND STOERMER, E. Quantitative morphometric methods in diatom research. *Nova Hedwigia* (01 2014), 281–306.
- [33] PEDRAZA, A., BUENO, G., DENIZ, O., CRISTOBAL, G., BLANCO, S., AND BORREGO-RAMOS, M. Automated diatom classification (part b): A deep learning approach. *Applied Sciences* 7 (05 2017), 460.
- [34] PÉREZ, P., GANGNET, M., AND BLAKE, A. Poisson image editing. *ACM Trans. Graph.* 22, 3 (July 2003), 313–318.
- [35] ROHLF, F., AND SLICE, D. Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology* 39 (03 1990).
- [36] ROUND, F., CRAWFORD, R., MANN, D., AND PRESS, C. U. *Diatoms: Biology and Morphology of the Genera*. Cambridge University Press, 1990.
- [37] SILVA-PALACIOS, D., FERRI, C., AND RAMÍREZ-QUINTANA, M. J. Improving performance of multiclass classification by inducing class hierarchies. *Procedia Computer Science* 108 (2017), 1692 – 1701. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.
- [38] STEVENSON, R. Ecological assessments with algae: A review and synthesis. *Journal of Phycology* 50 (06 2014), 437–461.
- [39] SULTANA, F., SUFIAN, A., AND DUTTA, P. Advancements in image classification using convolutional neural network. *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICR-CICN)* (Nov 2018).
- [40] SUN, A., AND LIM, E.-P. Hierarchical text classification and evaluation. pp. 521–528.

- [41] SUN, A., LIM, E.-P., AND NG, W.-K. *Hierarchical Text Classification Methods and Their Specification*. Springer US, Boston, MA, 2003, pp. 236–256.
- [42] VIOLA, P., AND JONES, M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001), vol. 1, pp. I–I.
- [43] WAHID, M. F., AHMED, T., AND HABIB, M. A. Classification of microscopic images of bacteria using deep convolutional neural network. In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)* (2018), pp. 217–220.
- [44] WU, X., AND SHAH, S. A bottom-up and top-down model for cell segmentation using multispectral data. pp. 592–595.
- [45] YANN LECUN, LEO BOTTOU, Y. B. P. H. Gradient-based learning applied to document recognition.
- [46] ZOU, Z., SHI, Z., GUO, Y., AND YE, J. Object detection in 20 years: A survey, 2019.