AUTOMATIC RECOGNITION OF DIATOMS USING A DEEP-LEARNING APPROACH FOR THE ECOLOGICAL DIAGNOSIS OF FRESHWATERS

Pierre FAURE--GIOVAGNOLI[1], Aishwarya VENKATARAMANAN[1,2,3], Cécile FIGUS[2,3], David HEUDRE[4],
Thibault de GARIDEL-THORON[5], Camille NOÛS[6], Philippe USSEGLIO-POLATERA[2], Cédric PRADALIER[1,3], Martin LAVIALE[2,3]

[1]GeorgiaTech Lorraine-UMI 2958, GeorgiaTech-CNRS, Metz, France
[2]Université de Lorraine, CNRS, LIEC, F-57000 Metz, France
[3]LTSER-"Zone Atelier Moselle", France
[4]Direction Régionale de l'Environnement, de l'Aménagement et du Logement Grand Est, Metz, France
[5]CEREGE, CNRS, Aix-Marseille Université, Collège De France, IRD, INRAE, Aix en Provence, France
[6]Laboratoire Cogitamus

CONTEXT

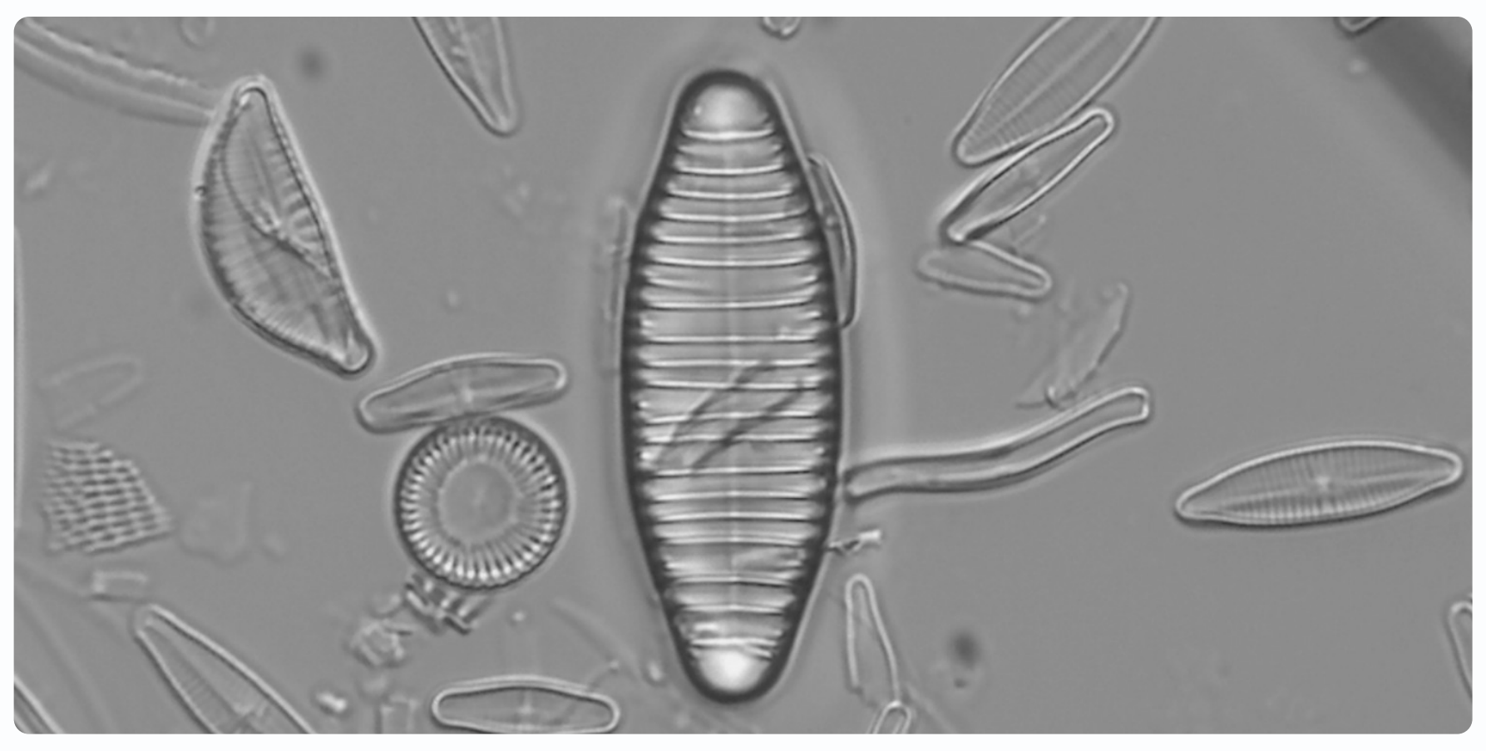
Figure 1. *Examples of diatoms.*

Diatoms are **ubiquist microalgae** that are commonly used for monitoring the **ecological status of watercourses** in the context of the European Water Framework Directive implementation. Current diatom-based biological indices rely on morphological criteria (shape and ornamentation of the siliceous exoskeleton, the frustule) that are sometimes difficult to characterize. Routine identification is **time-consuming**, often subject to **multiple biases** (operator experience, image quality) and sometimes requires a **high level of expertise**. This justifies the development of a robust tool for the **automatic detection and classification** of diatoms using microscopic images. The recent development of **deep learning approaches** allowing the automatic learning of the main characteristics of the image seems promising.

Our pipeline for automatic diatom identification using a deep-learning approach consists in 3 main steps (Figure 2): DETECTION → EXTRACTION → CLASSIFICATION.

DATASET

First, a large number of individual images of diatoms should be acquired. The dataset of images should be representative of the taxa encountered in the field both in terms of specific diversity and intra-specific variability (size, deformations). This dataset can be adapted to different study areas, from local (a specific river watershed) to larger spatial scales. For our proof of concept, a dataset representative of the Moselle watershed is currently under development.

PROOF OF CONCEPT

Three parallel approaches are used: **i/** a large set of permanent microscopic slides using an automatic images acquisition platform (Figure 1), each diatom present on every images being hand-labelled; **ii/** pre-labelled individual images of diatoms (Figure 3a) were retrieved from the literature, enabling us to gather 20,000 images representing 200 diatom taxa; **iii/** an image dataset of debris (sediment particles, pieces of frustule) is developed (Figure 3b).

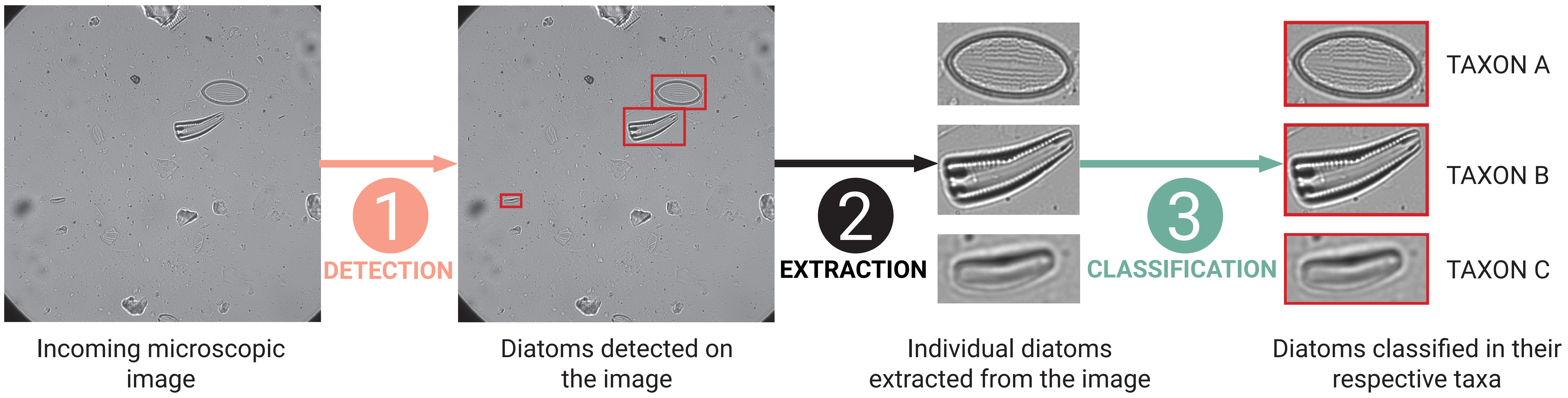
1
DETECTION
2
EXTRACTION
3
CLASSIFICATION
TAXON A
TAXON B
TAXON C
Incoming microscopic image
Diatoms detected on the image
Individual diatoms extracted from the image
Diatoms classified in their respective taxa

Figure 2. *Overview of the identification pipeline.*

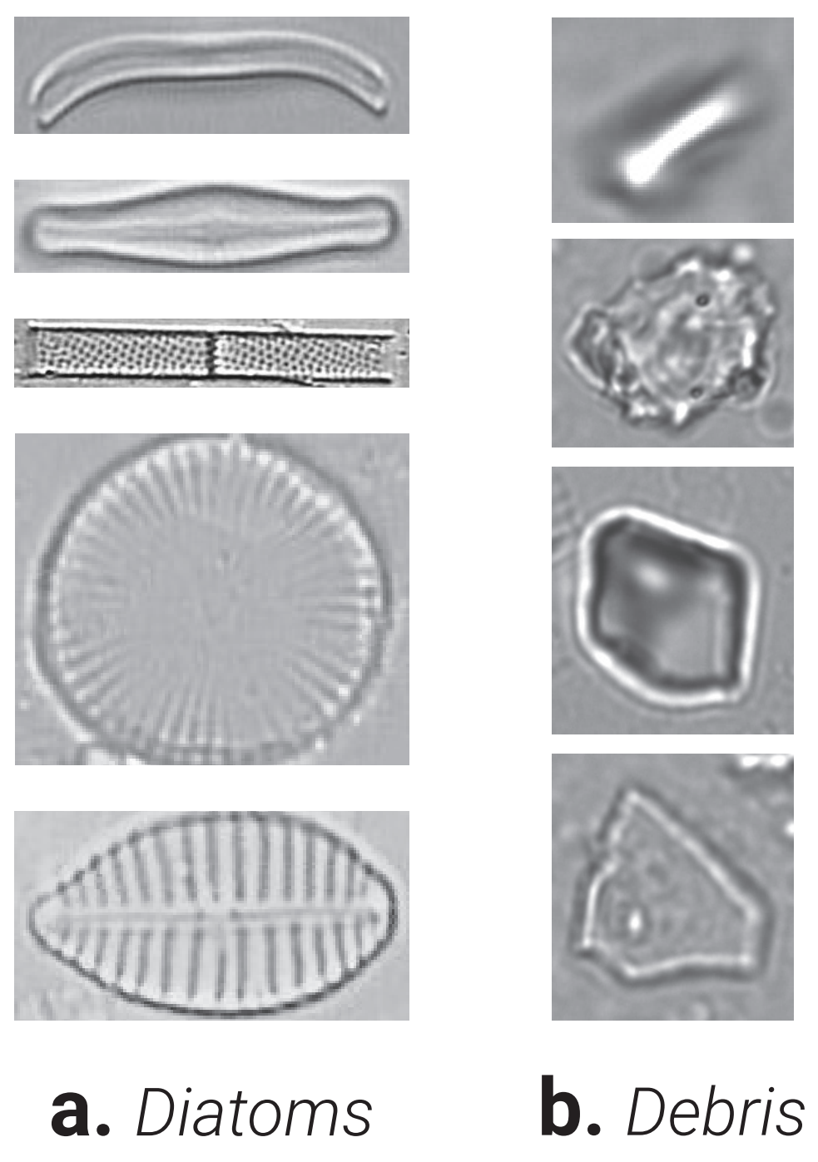
a. *Diatoms* **b.** *Debris*

Figure 3. *Examples of individual diatoms retrieved from the literature and hand extracted debris.*

DETECTION

The detection phase consists in automatically finding the diatoms on incoming microscope images while avoiding the **false positives** such as sediments or diatom debris. The main technical challenges for diatom detection on microscopic slides arise both from **optical properties** (focus, type of illumination…), **diatom morphology** (intra and inter-specific variability, orientation of the frustule) and the **quality of the microscopic slide** (overlapping, debris…).

PROOF OF CONCEPT

A first dataset of hand-labelled microscopic images was gathered in order to train and compare several commonly used object detection architectures (Faster R-CNN, YOLO...). To minimize the number of manually labelled images required, we propose the addition of a complementary **artificial dataset** (Figure 4): patchworks made from images of individual diatoms (Figure 3a) and debris (Figure 3b). Those images are used for an initial training step to create generic diatom detector.
For our proof of concept, these images resulted in gains up to **12% in accuracy** (from 70% to 82%) and **6% in recall** (from 71% to 77%) compared to the use of the real images alone.

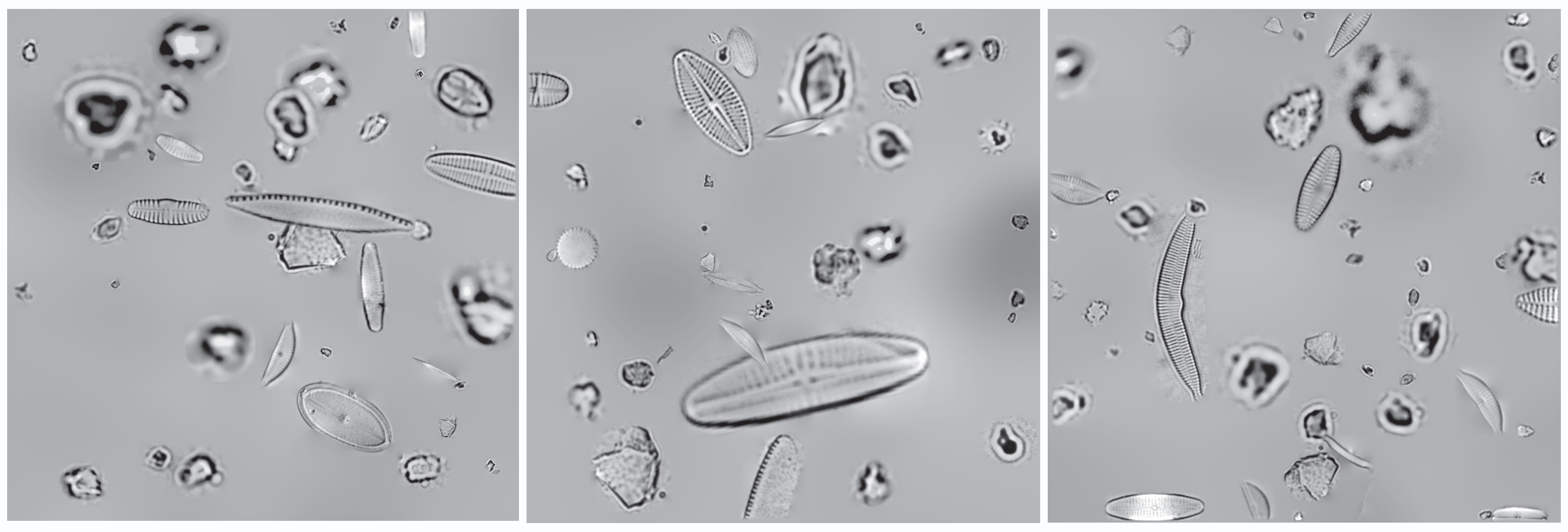
Figure 4. *Examples of artificial microscope slides.*

CLASSIFICATION

Automatic diatom classification dates back to the 2000's but the use of deep convolutional networks is much more recent, starting around 2017 in literature. Still, the number of diatom species needed to characterize a given area can be counted in hundreds which can be particularly challenging, even for the most recent models.

PROOF OF CONCEPT

Our tests using the latest advances in image classification have proven successful with up to **91% accuracy in discriminating 157 diatom taxa**. This is already a good score which could easily be improved by completing the training dataset. Nonetheless, some species will always be harder to discriminate due to their **similar morphology** or **under-representation** in the dataset which led us to explore new avenues, including the creation of an **artificial taxonomy** based on classification errors (Figure 5).

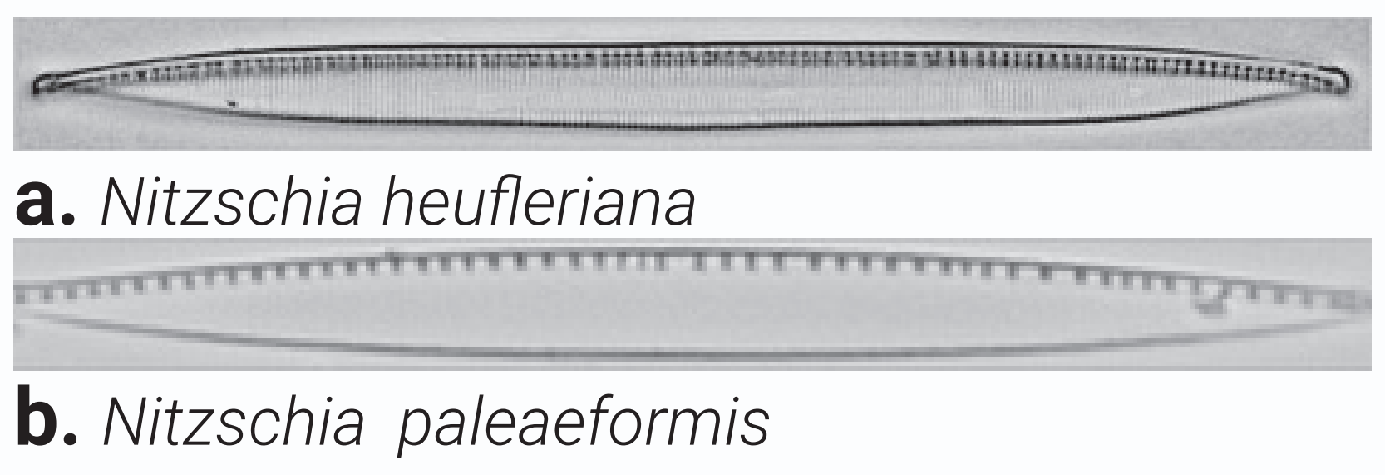
a. *Nitzschia heufleriana*
b. *Nitzschia paleaeformis*

Figure 5. *Examples of diatoms detected as difficult to discriminate by the algorithm.*

PERSPECTIVES

- Gathering more images for **completing the actual dataset** and expanding it to a larger spatial scale (Rhin-Meuse basin, national level…)
- Using **data augmentation** to artificially generate morphological variability within the learning dataset
- Implementing an algorithm able to take into account the **3D features of the diatom** frustule (z-stacks)
- Benefit from artificial taxonomy for **hierarchical classification**

ACKNOWLEDGMENTS

This work is funded by Université de Lorraine (IMADIA project), CNRS-INSU (Zone Atelier Moselle), Région Grand-Est and ANR (LOR-IA project) and by the Office Français de la Biodiversité-OFB (ex-ONEMA, ID: ONEMA-2016-02869). Many thanks to Souhila Founas for her help on the project.