applied sciences

MDPI

Article

Automated Diatom Classification (Part A): Handcrafted Feature Approaches

Gloria Bueno [1,*], Oscar Deniz [1], Anibal Pedraza [1], Jesús Ruiz-Santaquiteria [1], Jesús Salido [1], Gabriel Cristóbal [2], María Borrego-Ramos [3] and Saúl Blanco [3]

[1] VISILAB-University of Castilla-La Mancha, Av. Camilo José Cela s/n, 13071 Ciudad Real, Spain; Oscar.Deniz@uclm.es (O.D.); Anibal.Pedraza@uclm.es (A.P.); Jesus.RAlegre@uclm.es (J.R.-S.); Jesus.Salido@uclm.es (J.S.)

[2] Institute of Optics, Spanish National Research Council (CSIC), Serrano 121, 28006 Madrid, Spain; gabriel@optica.csic.es

[3] The Institute of the Environment, University of Leon, E-24071 León, Spain; mborr@unileon.es (M.B.-R.); saul.lanza@unileon.es (S.B.)

[*] Correspondence: gloria.bueno@uclm.es

Received: 31 May 2017; Accepted: 18 July 2017; Published: 25 July 2017

Abstract: This paper deals with automatic taxa identification based on machine learning methods. The aim is therefore to automatically classify diatoms, in terms of pattern recognition terminology. Diatoms are a kind of algae microorganism with high biodiversity at the species level, which are useful for water quality assessment. The most relevant features for diatom description and classification have been selected using an extensive dataset of 80 taxa with a minimum of 100 samples/taxon augmented to 300 samples/taxon. In addition to published morphological, statistical and textural descriptors, a new textural descriptor, Local Binary Patterns (LBP), to characterize the diatom's valves, and a log Gabor implementation not tested before for this purpose are introduced in this paper. Results show an overall accuracy of 98.11% using bagging decision trees and combinations of descriptors. Finally, some phycological features of diatoms that are still difficult to integrate in computer systems are discussed for future work.

Keywords: feature analysis; textural features; morphological features; automatic classification; handcrafted approaches; diatoms

1. Introduction

Diatoms are a major group of algae and are among the most common microorganisms in marine and freshwater habitats. They are important contributors to the primary production in aquatic ecosystems, placed at the bottom of the food chain. The diatoms have been shown to be increasingly important worldwide in studies related to climate change, as well as in the development of functions that allow the modeling of such change. Moreover, they are good indicators of environmental conditions and are commonly used in water quality assessment [1,2].

Diatom indices are known to correlate more significantly with water chemical variables, within continental waters, while macroinvertebrate- or plant-based methods are more sensitive to changes affecting structural parameters [3].

Diatoms have several advantages over other indicators that make them ideal as indicators of water quality. These features are: (a) their ability to spread over a variety of habitats; (b) they are relatively easy to sample, and such sampling has no impact on the ecosystem during collection; (c) they have a quick response to variation in environmental conditions; and (d) they are sensitive to changes in environmental conditions that may not be observed in other communities.

Diatom cells are enclosed within a unique silica cell wall known as a frustule made up of two valves, which fit into each other like a pill box. The frustules show a wide diversity of shapes and sizes, though mainly, they can be divided into centric diatoms (radial symmetry, rounded) and pennate diatoms (bilateral symmetry, elongated). The main taxonomic features used in diatom identification are related to the morphology and ornamentation of the frustule. The presence of raphe and the ornamentation of the frustule, stigmata and other features are important in identifying these organisms [4].

The diatom size, in the range of 2–2000 µm, is suitable for observation of most species using an optical microscope. However, diatoms require specialized skills for their classification, that is trained diatomists (phycologists specializing in diatom taxonomy). As stated in [5], more than 200,000 diatom species are estimated to exist, although just half of them have been described. Several intercalibration tests have shown that the results of biological indices based on diatoms are highly sensitive to the level of accuracy in taxonomic classification [6]. The identification task is very difficult due to the huge number of species estimated to exist [7].

Classical taxonomic diagnosis is performed using key features or by visual comparison with type samples or reference iconographies [8]. The conventional approach is to analyze these microalgae using light microscopy (brightfield, DIC, RIC, etc.). Diatom-based metrics are calculated based on the relative abundance of different taxa in an assemblage and the autecological parameters characterizing each species. However, the current manual analysis of images is tedious, requiring highly qualified staff, and it is time consuming. This is the case when diatoms are used in the context of water quality, as in this study. According to a European directive, in order to compute an index score, the identification of a minimum of 400 valves per sample is required [9].

In the case of transparent specimens such as diatoms, brightfield microscopy presents some difficulties. Some details are barely distinguishable from the background, and some other alternative modalities such as phase contrast, DIC or dark field need to be considered. There are few species of the *Nitzschia*, which are good examples of that (*Nitzschia costei*, *Nitzschia frustulum var frustulum* and *Nitzschia inconspicua*). This is probably due to a not well-developed silicification process.

Phase contrast microscopy is a suitable technique for visualizing transparent specimens. However, it produces some artifacts (halos and shading-off) that limit its usefulness in some applications. Halos are unresolved images with reverse contrast, and shading-off is a contrast-decreasing effect from the edge of the specimen towards the center of it [10]. DIC microscopy is also a popular method for improving the contrast of unstained specimens. DIC is absent of the halo effect found in phase contrast microscopy producing pseudo-relief images that can be understood as the derivative of the optical path length defined as product of the index of refraction times the specimen thickness. In dark field microscopy, the specimen is illuminated with a hollow cone of light, which is too wide to enter the objective lens. Dark field is a suitable modality for diatom visualization that can be a substitute of phase contrast or DIC in many cases, although both are out of the scope of this work. Scanning electron microscopy is another suitable modality, especially in taxonomy for revealing structural details [11], which is also out of the scope of this paper. On the other hand, most of the diatom databases that are publicly available use brightfield microscopy.

Other difficulties attached to microscopy are: images partially focused and multiple orientations (views). Both are related to the projection of 3D objects into 2D images. The main challenges faced by automatic identification and classification methods are due not only to the high number of species to recognize, but to the great similarities between them and even the presence of polymorphisms within species. In some cases, analysis is simply unfeasible due to the huge amount of information and images involved. Advances in digital microscopy and image analysis systems offer a potentially advantageous solution compared to manual methods of counting and classification.

Before going further, it is worth pointing out the difference between classification and identification in terms of biology and pattern recognition terminology. In biology, it may be said that the term identification answers the question: ‘What is the name of the taxon in front of me?’. However, classification answers questions of the sort: How is this taxon related to other taxa? According to the pattern recognition terminology, a class is a set of objects that within a given context is recognized as similar. Such a class has usually a unique name, the class name. The individual objects within a class have a label that refers to this name. Additionally, classification is the assignment of a class name to an object by evaluating a trained classifier for that object. Therefore, in this study, our objects or classes are the diatoms, and we will classify them to give them a label with their taxon name. Henceforth, a diatom taxa will be referred to as a class.

A good review of morphometric methods for shape analysis and landmark-based analysis used in diatom research is presented in Pappas et al. [12]. Another attempt to describe the morphology and geometry of diatoms is the work of Kloster et al. [13], who developed a system that allows the segmentation and feature extraction from diatom contours, although it does not provide textural information about the frustule. As mentioned by Pappas et al. [12], two areas of study may be identified with regard to the methods used in diatom research, namely shape analysis and pattern recognition. We will focus on pattern recognition or machine learning methods, but notice that the traditional machine learning methods are based on a previous definition of a set of features describing the objects to be classified that may or may not have a biologically-meaningful interpretation.

Methods for diatom detection and identification have been studied in Cairns et al., 1979 [14], Culverhouse et al., 1996 [15], Pech-Pacheco and Alvarez-Borrego, 1998 [16], Pech-Pacheco 2001 [17]. Cairns et al. [15] have proposed some diatom identification methods based on coherent optics and holography. However, such work did not have any impact on diatom research mainly due to the fact that the identification system was too specialized and probably too expensive to be used by a diatomist community [18]. Culverhouse et al. [15] derived some methods for phytoplankton identification based on neural networks, but again, they do not provide a fully-automatic method. Pech-Pacheco et al. [16] have proposed a hybrid optical-digital method for the identification of five different species of phytoplankton through the use of operators invariant to translation rotation and scale. Pappas and Stoermer 2003 [19], used form descriptors by Legendre polynomials and principal component analysis in the identification of the *Cymbella cistula* species.

An important attempt to automate diatom classification was conducted for the ADIAC project (Automatic Diatom Identification And Classification) [18,20]. Several accuracy results were reported with a database composed of different numbers of diatom taxa ranging from 37–55 classes. In ADIAC, 171 features were used for diatom classification. These features are intended to describe the diatom symmetry, shape, geometry and texture by means of different descriptors, such as: rectangularity, circularity, compactness, shape of poles, length, width, length-width ratio, size, stria density orientation, horizontal frequency, Gray-Level Co-occurrence Matrix (GLCM), moment invariants, Gabor wavelets, Fourier and Scale-invariant Feature Transform (SIFT) descriptors. The classifiers that perform better are bagging of decision trees and random forest of predictive clustering trees, all of them evaluated with 10-fold cross-validation (10 fcv). The best results, up to 97.97% accuracy, were obtained with 38 classes using Fourier and SIFT descriptors with random forest. Performance decreased down to 96.17% when classifying 55 classes with the same descriptors and classifier [21].

New techniques based on Convolutional Neural Networks (CNN) have also been explored to classify sea plankton (Kuang 2015 [22], and Dai et al., 2016 [23]). Note, however, that these images are different than those studied in this paper since this type of plankton varies from phytoplankton (diatoms). In [22], a database of 30,000 images belonging to 121 classes was used. The results were poor with a maximum performance of 73.90%. In [23], a database of 30,000 images belonging to 33 classes was used. They obtained an accuracy up to 96.3%. The other work related to CNN applied to diatoms is the one published by the authors (see the next paper companion [24]). The work presented here and the methodology have been compared to the CNN approach [24].

Thus, most of the efforts are still carried out with handcrafted approaches or "hand-designed" methods where a set of fixed features is used. That is, the methods rely on expert knowledge to extract the most relevant features versus CNN approaches that learn features from data. However, still, the handcrafted methods present limited results as in [25], where 14 classes were classified with Support Vector Machine (SVM) 10 fcv, using 44 GLCM features that describe only geometric and morphological properties. They obtained an accuracy of 94.7%.

Therefore, the automated classification of diatoms (in terms of pattern recognition) or taxon identification remains a challenge. At present, there is no system capable of taking into account variations in both the contour and the texture in a relatively large number of species. One of the reasons is the difficulty in acquiring a big dataset of tagged data with a sufficient number of samples per species. The classification of diatoms is tedious and laborious, even for an expert diatomist.

In this paper, we present a complete study of relevant features to describe and classify diatoms. The main purpose is to define the most discriminant features and to make a comparison of classifiers based on these features versus the CNN approach. For that, we collected an important database of 80 diatom taxa with 300 samples per taxon described in Section 2. The database was composed of an average of 100 distinct diatoms per class and augmented by means of computational simulations up to 300 samples per class. In order to extract the main diatom features, we propose in Section 3 a segmentation to apply descriptors to the contour and inner diatom regions. In Section 4, a complete list of descriptors is provided. Those are handcrafted features that describe, in terms of computer vision, the discriminant properties of diatoms. Sections 5 and 6 describe classification strategies and some classifiers. Experimental results are presented in Section 7 where an overall accuracy of 98.11% is presented, which improves previous related works. Finally, Section 8 concludes the paper addressing unresolved challenging problems.

2. Materials: Dataset Preparation

Once having collected the diatom samples from the rivers, the chemical treatment of the sample is carried out in the laboratory with hydrogen peroxide (120 vol.), which causes the digestion of the organic matter and allows one to obtain suspensions of frustules and valves free of organic remains. The process is done at a temperature of 70–90 °C, to accelerate the reaction. A few drops of the sample are taken and deposited in a round coverslip. After evaporation of the water, the diatom frustules remain in the cover-objects. Then, using a synthetic resin (Naphrax) with an optical refractive index of 1.7, diatoms are attached to the glass slide for later classification under brightfield microscopy following standard protocols [9].

The 80 diatom species studied here were collected during the years 2003 and 2015 from the Duero Basin water in Spain [26]. Those are the 80 dominant taxa in terms of relative abundance and occurrence. A Westbury SP/40 Brunel microscope and a Brunel AMA 050 camera were employed to capture the images at 60× magnification, with a numeric aperture of 0.85 and a physical resolution of 7.91 pixels/µm. An average of 100 distinct diatom valves per each diatom class were then manually cropped and labeled by an expert diatomist. The exact number of the cropped diatom valves is shown in Table 1. To complete the dataset with up to 300 image samples per taxon, a data augmentation was performed by means of applying rotations of 90°, 180°, 270° and up-down and right-left flips to the cropped images. These 6 transformations were performed on the original images and only if needed, to obtain up to 300 image samples per class. Thus, we end up with 24,000 images to classify into 80 diatom taxa.

Data augmentation aims at increasing the number of images in the dataset by representing image data in different orientations. That is, different copies of the same image are made, but from different perspectives or visual angles. Rotating and mirroring the data in different orientations may eventually help with identifying a similar object in different orientations. A more robust classifier will be obtained if the data are randomly rotated in multiple orientations.

In Figure 1, a capture with a manually-selected diatom is shown. The elements that form the ornamentation of the frustule in a diatom are illustrated in the cropped diatom. Features of the stria are key in diatom taxonomy, such as: areola and lineolae. Areola is a perforation (or pore) in the diatom valve, and lineolae are areola elongated in the apical direction. The lineolae density is calculated in ppm (pixels per micron). Every sample was manually cropped to ensure the best diatom samples avoiding as much as possible nearby samples or debris. The list of the 80 species with the number of original selected images is indicated in Table 1, and some examples are depicted in Figure 2. The number of the taxon corresponding to the one listed in Table 1 is shown in the upper left corner of each picture. The database can be obtained by request (see the contact at http://aqualitas-retos.es/en/), and it will be publicly available at the end of the project. Figure 3 shows the same diatom species at different views ("valvar view" and "girdle view") and sizes. Due to the deposition process of the sample, in most situations, the diatoms appear in valvar view, although sometimes appear in lateral view (less than 10% of cases).

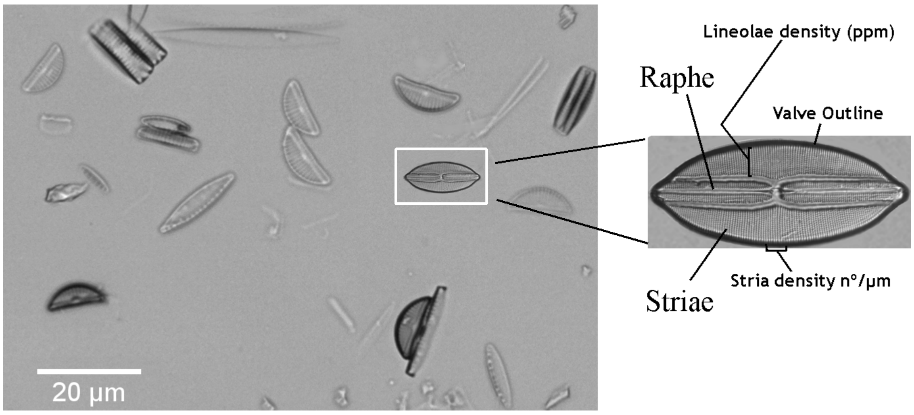


Figure 1. Diatoms observed by a microscope at 60× magnification and their main elements. The original image size is 903 × 614 pixels, and the selected diatom sample is 138 × 85 pixels.

Table 1. List of the 80 diatom species analyzed in the current study, showing the number of valves per class.

1. *Achnanthes subhudsonis*	123	28. *Encyonema minutum*	120	55. *Gomphonema rhombicum*	64
2. *Achnanthidium atomoides*	129	29. *Encyonema reichardtii*	152	56. *Humidophila contenta*	105
3. *Achnanthidium caravelense*	59	30. *Encyonema silesiacum*	108	57. *Karayevia clevei varclevei*	84
4. *Achnanthidium catenatum*	187	31. *Encyonema ventricosum*	101	58. *Luticola goeppertiana*	136
5. *Achnanthidium druartii*	93	32. *Encyonopsis alpina*	106	59. *Mayamaea permitis*	40
6. *Achnanthidium eutrophilum*	97	33. *Encyonopsis minuta*	89	60. *Melosira varians*	146
7. *Achnanthidium exile*	98	34. *Eolimna minima*	174	61. *Navicula cryptotenella*	136
8. *Achnanthidium jackii*	125	35. *Eolimna rhombelliptica*	132	62. *Navicula cryptotenelloides*	107
9. *Achnanthidium rivulare*	305	36. *Eolimna subminuscula*	94	63. *Navicula gregaria*	50
10. *Amphora pediculus*	117	37. *Epithemia adnata*	72	64. *Navicula lanceolata*	77
11. *Aulacoseira subarctica*	113	38. *Epithemia sorex*	85	65. *Navicula tripunctata*	99
12. *Cocconeis lineata*	81	39. *Epithemia turgida*	93	66. *Nitzschia amphibia*	124
13. *Cocconeis pediculus*	49	40. *Fragilaria arcus*	93	67. *Nitzschia capitellata*	123
14. *Cocconeis placentula var euglypta*	117	41. *Fragilaria gracilis*	54	68. *Nitzschia costei*	72
15. *Craticula accomoda*	86	42. *Fragilaria pararumpens*	74	69. *Nitzschia desertorum*	71
16. *Cyclostephanos dubius*	85	43. *Fragilaria perminuta*	89	70. *Nitzschia dissipata var media*	81
17. *Cyclotella atomus*	99	44. *Fragilaria rumpens*	49	71. *Nitzschia fossilis*	76
18. *Cyclotella meneghiniana*	103	45. *Fragilaria vaucheriae*	82	72. *Nitzschia frustulum var frustulum*	226
19. *Cymbella excisa var angusta*	79	46. *Gomphonema angustatum*	86	73. *Nitzschia inconspicua*	255
20. *Cymbella excisa var excisa*	241	47. *Gomphonema angustivalva*	55	74. *Nitzschia tropica*	65
21. *Cymbella excisiformis var excisiformis*	142	48. *Gomphonema insigniforme*	90	75. *Nitzschia umbonata*	91
22. *Cymbella parva*	177	49. *Gomphonema micropumilum*	89	76. *Rhoicosphenia abbreviata*	94
23. *Denticula tenuis*	181	50. *Gomphonema micropus*	117	77. *Skeletonema potamos*	155
24. *Diatoma mesodon*	115	51. *Gomphonema minusculum*	158	78. *Staurosira binodis*	94
25. *Diatoma moniliformis*	134	52. *Gomphonema minutum*	93	79. *Staurosira venter*	87
26. *Diatoma vulgaris*	88	53. *Gomphonema parvulum saprophilum*	52	80. *Thalassiosira pseudonana*	70
27. *Discostella pseudostelligera*	82	54. *Gomphonema pumilum var elegans*	128		

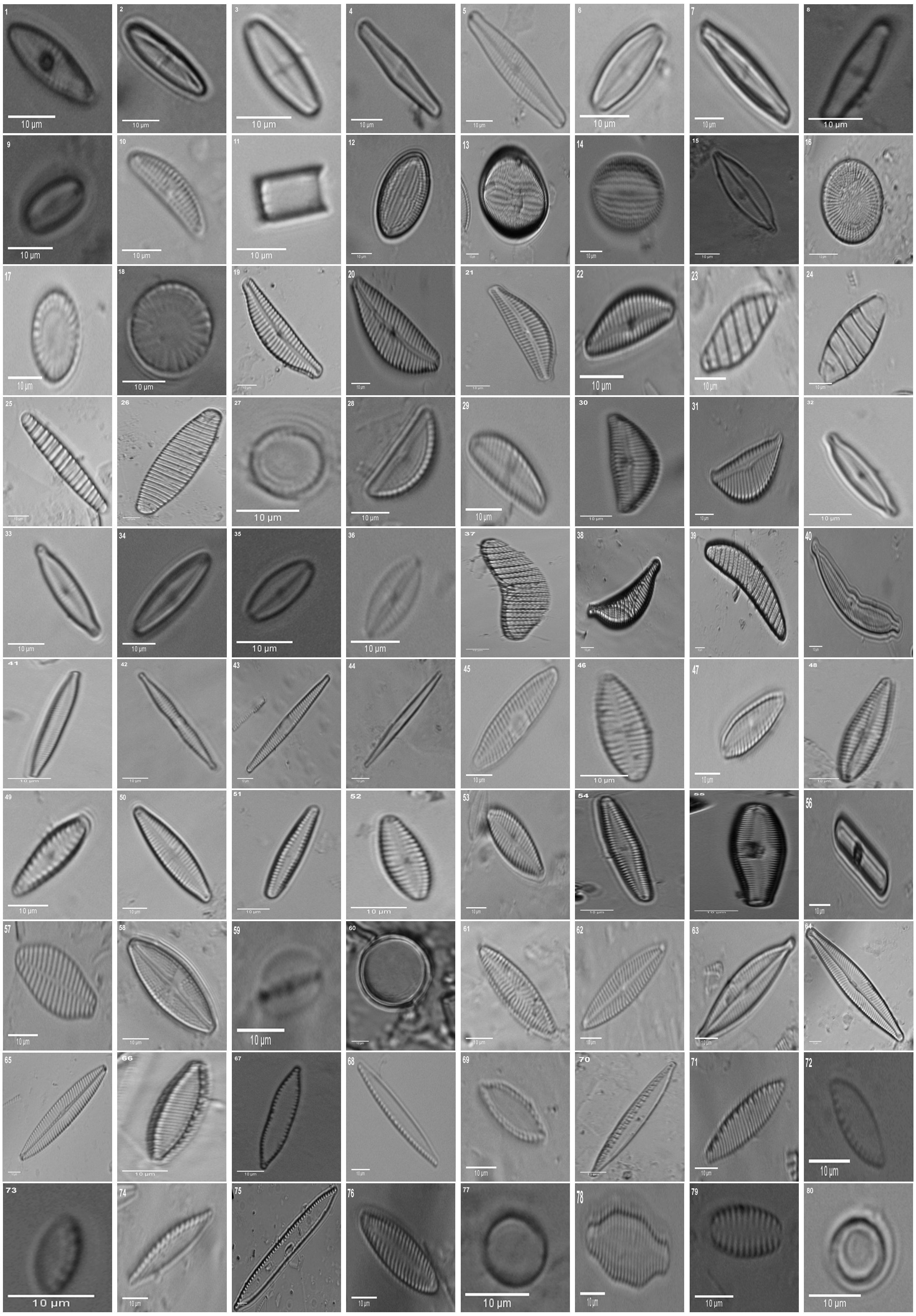
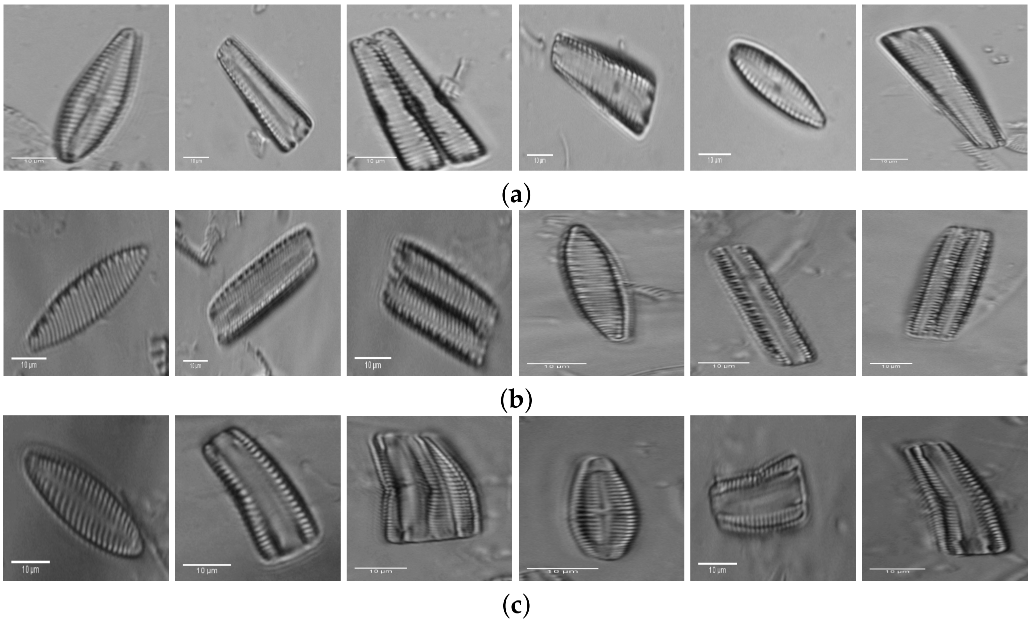


Figure 2. Examples of the 80 diatom taxa classified in this study. Sample images are stretched for visualization purposes.



(**a**)

(**b**)

(**c**)

Figure 3. Different views and sizes of the same species: (**a**) *Gomphonema insigniforme*; (**b**) *Nitzschia fossilis*; and (**c**) *Rhoicosphenia abbreviata*.

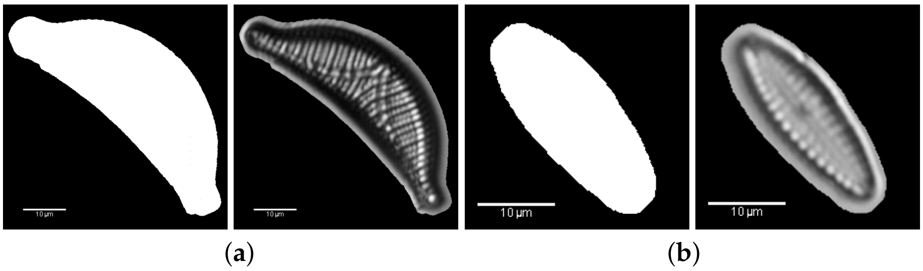
3. Valve Segmentation: Binary Thresholding

There are several works about diatom detection mainly related to the above-mentioned ADIAC project. It is out of the scope of this paper to present all of the segmentation methods; for further details, the reader is referred to the reviews mentioned in Section 1 [12,13,18]. In this work, we present an automatic method used to do an initial quick segmentation, which required visual supervision afterward to include only the correct segmented diatoms. Some of the images have been acquired with low contrast and background noise, which produces a poor segmentation in terms of valve overlapping with other structures. That means visual supervision is needed for discarding segmentation errors. The valve is the most significant region of the diatoms where structural differences can be distinguished. Therefore, the segmentation process should accurately extract such a region for extracting relevant features. A proper segmentation of the valve is expected to affect textural, frequential and statistical descriptors. This segmentation is done here by means of a binary thresholding where the segmented region is the binary masks where descriptors must be computed.

The process to obtain the binary masks consists of four steps:

1. Binary thresholding: automatic segmentation based on Otsu's thresholding.
2. Maximum area: calculation of the largest region (area).
3. Hole filling: interior holes are filled if present, using mathematical morphology operators.
4. Segmentation: the ROI is cropped with the coordinates of the bounding-box of the largest area (Step 2).

Due to the presence of debris and overlapping with different diatom taxa (or debris), the segmentation was afterward manually checked to discard bad segmented diatoms or finish correction; see some examples of the segmented diatoms and their masks in Figure 4.



(**a**) (**b**)

Figure 4. Segmented diatoms and their binary mask: (**a**) *Epithemia sorex*; (**b**) *Gomphonema minutum*.

4. Diatom Handcrafted Feature Descriptors

An effort must be made in translating the knowledge of the diatomists to distinguish the different diatom species and describe the most relevant features in terms of computer vision and pattern classification. The goal is to mimic human perception and the ability to recognize a 3D object from a 2D image, in this case diatoms. Even if it is sometimes unclear what features are used by the expert to distinguish among very similar diatom species, we proposed and described diatom features in terms of automatic pattern classification.

Along the next subsections, the handcrafted features are presented in groups according to their formulation with a brief explanation. The different groups of descriptors are indicated in Table 2. A total of 1460 descriptors is computed, and all of them are calculated uniquely in those pixels belonging to the segmented binary masks.

Table 2. List of handcrafted feature descriptors divided into categories. LBP, Local Binary Pattern.

CATEGORY	HANDCRAFTED FEATURE	TOTAL
Morphological	Area, eccentricity (3 eccentricities) Perimeter, shape, fullness	7 features
Statistical	1st order (histogram)	13 features
	2nd order (co-occurrence matrix)	19 features *distance* $= 1, 3, 5$ pixels *direction* $= 0°, 45°, 90°, 135°$
Texture space	LBP	Stat.241 features
Moments	Hu	7 moments
Space-frequency	Log Gabor	4 scales (6 orientations) 241 × 4 = 964 features

4.1. Morphological Descriptors

Morphological features related to frustule's contour and area are computed from the binary masks.

4.1.1. Area

This descriptor is calculated as the sum of pixels in the binary mask ($B \in (0,1)$) of size MxN:

$$Area = \sum_{n=1}^{N} \sum_{m=1}^{M} B(m,n) \quad (1)$$

4.1.2. Eccentricity

These descriptors reflect elongation in relation with the binary mask's center of mass, also called the centroid and defined as:

$$(m_c, n_c) = \left(\frac{1}{Area} \sum_{(m,n) \in Area} m \cdot B(m,n), \frac{1}{Area} \sum_{(m,n) \in Area} n \cdot B(m,n) \right) \quad (2)$$

The first $Eccentricity_1$is defined as a quotient of the maximum and minimum distance between the centroid and binary mask's border (i.e., frustule's contour), also called outer and inner circumference radius.

$$Eccentricity_1 = \frac{Outer - radius}{Inner - radius} \tag{3}$$

Similarly, $Eccentricity_2$ is calculated as the quotient of the semi-axes of the best fitting ellipse for the mask, and $Eccentricity_3$ is the ratio of the inertia moments of the two semi-axes of the best fitting ellipse. The moments are defined in Section 4.4.

4.1.3. Perimeter

This descriptor is the number of pixels that belong to the diatom's outline (i.e., a pixel belongs to the perimeter if it is nonzero and is connected with at least one pixel equal to zero).

$$Perimeter = \sum_{n=1}^{N} \sum_{m=1}^{M} P(n,m) \tag{4}$$

where: $P(m,n) = 1 \;\; if \;\; \exists \;\; B(m \pm 1, n \pm 1) = 1$ and $P(m,n) = 0$ otherwise.

4.1.4. Shape

This descriptor is a measure of the elongation of an object. It is given by:

$$Shape = \frac{4 \cdot \pi \cdot Area}{Perimeter^2} \tag{5}$$

4.1.5. Fullness

This descriptor is the ratio of the mask area to the bounding box area.

4.2. Statistical Descriptors

4.2.1. First Order Statistical: Histogram

These descriptors, listed in Table 3, calculate common statistics in the image histogram $h(i)$ calculated on a 255-bin H. This group of descriptors is sensible to variations of gray pixel levels, but they ignore their local correlation.

Table 3. First-order statistical descriptors.

Mean	$\mu = \sum_{i=0}^{H-1} i \cdot h(i)$
Mode	$i = argmax(h(i))$
Minimum	$min(h(i))$
Maximum	$max(h(i))$
Variance	$\sigma = \sum_{n=0}^{H-1} (i - \mu)^2 \cdot h(i)$
Range	$max(h(i)) - min(h(i))$
Entropy	$\sum_{i=0}^{H-1} h(i) \cdot log(h(i))$
1st *Quartile*	$\mu_{q1} = \sum_{i=3\lceil H/4 \rceil}^{H} i \cdot h(i)$
2nd *Quartile*	$\mu_{q2} = \sum_{i=2\lceil H/4 \rceil}^{3\lceil H/4 \rceil} i \cdot h(i)$
3rd *Quartile*	$\mu_{q3} = \sum_{i=\lceil H/4 \rceil}^{2\lceil H/4 \rceil} i \cdot h(i)$
Interquartile Range	$\mu_{q3} - \mu_{q1}$
Asymmetry	$\frac{1}{\sigma^3} \sum_{n=0}^{H-1} (i - \mu)^3 \cdot h(i)$
Kurtosis	$\frac{1}{\sigma^4} \sum_{n=0}^{H-1} (i - \mu)^4 \cdot h(i)$

Histogram $h(i)$, bins number H, floor operator $\lceil \; \rceil$.

4.2.2. Second Order Statistical: Co-Occurrence Matrix

The co-occurrence matrix $c(m, n)$, which is defined as the distribution of co-occurring pixel values at a given distance (d) and direction (°), can be used for measuring the texture of an image. The distances and direction used in this study are d equal to 1, 3 and 5 pixels and (°) equal to 0°, 45°, 90° and 135° (see Table 2). Feature descriptors extracted from co-occurrence matrices are also called Haralick features (Haralick) [27]. The 19 second order statistical features used in this study are listed in Table 4.

Table 4. Second order statistical descriptors.

Energy	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j)^2$
Variance	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i-\mu)^2 \cdot c(i,j)$
Contrast	$\sum_{n=0}^{H-1} n^2 \left(\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j)\right), \quad \lvert i-j\rvert = n$
Dissimilarity	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} \lvert i-j\rvert \cdot c(i,j)$
Correlation	$\frac{1}{\sigma_x\sigma_y}\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i \cdot j \cdot c(i,j) - \mu_x\mu_y$
Autocorrelation	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i \cdot j \cdot c(i,j)$
Entropy	$T = -\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j) \cdot log(c(i,j))$
Measure of Correlation 1	$\frac{T-HXY1}{max(HX,HY)}$
Measure of Correlation 2	$(1-exp[2 \cdot (HXY2-T)])^{0.5}$
Cluster Shade	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i+j-\mu_x-\mu_y)^3 \cdot c(i,j)$
Cluster Prominence	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1}(i+j-\mu_x-\mu_y)^4 \cdot c(i,j)$
Maximum Probability	$max(c(i,j)), \quad i=[0...H-1], j=[0...H-1]$
Sum Average	$\sum_{i=0}^{2(H-1)} i \cdot c_{x+y}(i)$
Sum Entropy	$\sum_{i=0}^{2(H-1)} c_{x+y}(i) \cdot log(c_{x+y}(i,j))$
Sum Variance	$-\sum_{i=0}^{2(H-1)}(i-SumEntropy)^2 \cdot c_{x+y}(i)$
Difference Entropy	$-\sum_{i=0}^{H-1} c_{x-y}(i) \cdot log(c_{x-y}(i,j))$
Difference Variance	$\sum_{i=0}^{H-1} i^2 \cdot c_{x-y}(i)$
Homogeneity 1	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} \frac{c(i,j)}{1+(i-j)^2}$
Homogeneity 2	$\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} \frac{c(i,j)}{1+\lvert i-j\rvert^2}$

H bins number, HX and HY entropy of p_x and p_y.
$\mu_x = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} i \cdot c(i,j); \qquad \mu_y = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} j \cdot c(i,j)$
$c_x(i) = \sum_{j=0}^{H-1} c(i,j); \qquad c_y(j) = \sum_{i=0}^{H-1} c(i,j)$
$\sigma_x = \sqrt{\sum_{i=0}^{H-1} c_x(i)(i-\mu_x)^2}; \qquad \sigma_y = \sqrt{\sum_{j=0}^{H-1} c_y(i)(i-\mu_y)^2}$
$c_{x+y}(k) = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j); \quad i+j=k, k=[0...2(H-1)]$
$c_{x-y}(k) = \sum_{i=0}^{H-1}\sum_{j=0}^{H-1} p(i,j); \quad \lvert i-j\rvert=k, k=[0...H-1]$
$HXY1 = -\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c(i,j) \cdot log(c_x(i) \cdot c_y(j))$
$HXY2 = -\sum_{i=0}^{H-1}\sum_{j=0}^{H-1} c_x(i) \cdot c_y(j) \cdot log(c_x(i) \cdot c_y(j))$

4.3. Local Binary Patterns

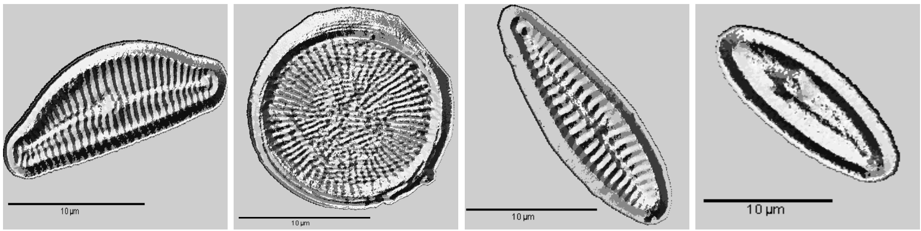
The Local Binary Pattern (LBP) operator is based on the idea that textural properties within homogeneous regions can be represented as patterns [28–31]. These patterns represent micro-features. It analyzes a "texture spectrum", e.g., using a 3 × 3 mask and comparing their values with the central pixel. The pixels with a lower value than the central one are labeled with "0", otherwise with "1". A linear combination is applied where the labeled pixels are multiplied by a fixed weighting function and summed to obtain a label: $LBP(g_c) = \sum_{p=0}^{7} s(g_p - g_c)2^p$, where $\{g_p | p = 0, \ldots, 7\}$ are the neighbors of g_c, and the comparison function is defined as: $s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

The mask may be defined using a circular neighborhood [32], denoted by (P, R), where P is the number of sampling points and R is the radius of the neighborhood. Ojala et al. [32] observed that over 90% of patterns can be described with few LBP patterns, so they introduced a uniformity measure $U(LBP_{P,R}(g_c)) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|$, which corresponds to the number of transitions $(0/1)$ in the labeled LBP.

In this way, the uniform-LBP ($LBP_{P,R}^{uni}$) can be obtained as:

$$LBP_{P,R}^{uni}(g_c) = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R}(g_c)) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \tag{6}$$

After the LPB operator, a labeled image is obtained. Once this process is completed, the pixel-wise information from the labeled image is encoded as a histogram, so that it can be interpreted as a fingerprint of the analyzed diatom area. $LBP_{P,R}^{uni}$ produces $(P+2)$-bin histograms [33] where the statistical descriptors are computed. As far as the authors know, LBP has never been tested before for diatom classification. Figure 5 shows some examples of LBP corresponding to some diatom species.



(**a**) (**b**) (**c**) (**d**)

Figure 5. LBP images corresponding to the following species: (**a**) *Cymbella excisa* var *angusta*; (**b**) *Cyclostephanos dubius*; (**c**) *Gomphonema insigniforme*; (**d**) *Achnanthes subhudsonis*.

4.4. Hu Moments

Image moments provide a shape description both morphologically and statistically [34]. Hu moments are invariant with respect to translation, scale and rotation and can be generated from the central moments. The central moments for an image $g(m, n)$ can be formulated as follows:

$$\mu_{pq} = \sum_m \sum_n (m - m_c)^p \cdot (n - n_c)^q \cdot g(m, n) \tag{7}$$

where $m_c = \frac{M_{10}}{M_{00}}$ and $n_c = \frac{M_{01}}{M_{00}}$ are the components of the centroid and M_{pq} raw moments $M_{pq} = \sum_m \sum_n m^p \cdot n^q \cdot g(m, n)$.

The seven Hu moment invariants are given by:

$$\begin{aligned}
&\phi_1 = \eta_{20} + \eta_{02} \\
&\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
&\phi_3 = (\eta_{30} - 3\eta_{21})^2 + (3\eta_{21} - \eta_{03})^2 \\
&\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
&\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
&\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{8}$$

where $\eta_{rs} = \frac{\mu_{rs}}{\mu00}$ and $\gamma = \frac{r+s}{2} + 1, r+s = 2, 3, \ldots, \text{inf}$.

4.5. Texture Descriptors in the Space-Frequency Domain

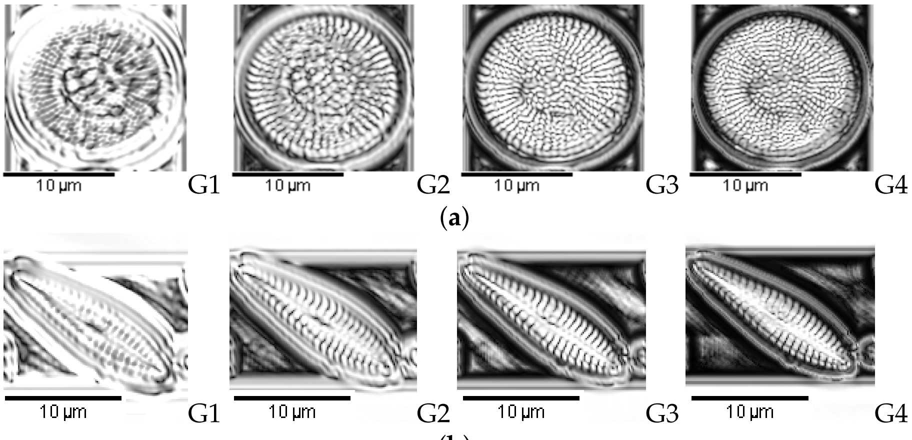
The previously-introduced statistical descriptors are calculated in the space-frequency domain. It is assumed that features, somehow hidden, arise with higher visibility in this domain. Thus, some transformations must be applied to the image to analyze its properties in this domain. In this study, a log-Gaborfilters were used to characterize the texture of diatoms. Then, for every transformed domain or sub-band, a total number of 964 statistical descriptors is calculated; see Section 4.2.

Log Gabor Transform

Log-Gabor filters are defined in the frequency domain as Gaussian functions shifted from the origin to avoid the singularity of the log function. In addition, the Gaussian envelope is modulated by a complex exponential with even and odd phases, which is effective for characterizing edges. Here, log Gabor filters proposed by Fischer et al. [35] have been used. The log Gabor descriptor is based on the energy calculated at every scaled level:

$$\begin{aligned} Gabor_l &= \textstyle\sum_{o=1}^{O} \sum_{u=1,v=1}^{U,V} \left| F^{-1} \left(G_{lo} \cdot I \right) \cdot B(u,v) \right|,, \quad l \in \{1,..,L\} \\ G_{lo} &= \exp\left(-\tfrac{1}{2} \left(\tfrac{\rho-\rho_l}{\sigma_\rho} \right)^2 \right) \exp\left(-\tfrac{1}{2} \left(\tfrac{\theta-\theta_{pl}}{\sigma_\theta} \right)^2 \right) \end{aligned} \tag{9}$$

where F^{-1} is the inverse Fourier transform, G_{lo} is the log Gabor filter with L scales and O orientations in log-polar coordinates and (ρ, θ) and $(\sigma_\rho, \sigma_\theta)$ are the angular and radial bandwidths; see [35] for more details. Again, $L = 4$, $O = 6$, and the residual DC-component is discarded. Figure 6 shows the log Gabor filters (four bands: G1, G2, G3 and G4) applied to a diatom example.



(**a**)

(**b**)

Figure 6. Log Gabor filters applied to diatom samples: (**a**) *Cyclostephanos dubius*; (**b**) *Gomphonema insigniforme*.

5. Discriminant Analysis

The handcrafted feature descriptors described previously produce 1460 characteristics. However, not all of the features are discriminating for the problem we are faced with, due to the fact that they are describing properties that are widely spread along all classification groups or because they are redundant (correlated) with respect to other features. In that case, such descriptors provide useless information that will likely impair classification not only in terms of performance and accuracy, but also in terms of speed due to the higher dimensionality [36,37]. Therefore, a feature selection process is required to remove redundant information. To this end, we used the correlation coefficient as the similarity measure between two or more features.

5.1. Correlation

The entire bank of first and second order statistical descriptors is calculated and extracted from the LBP labeled image and from each decomposition band of the log Gabor transformed image. This means that the total number of descriptors becomes four-times larger given a four-level decomposition transform. This approach increases the workload and may cause highly correlated variables.

Correlated variables have been calculated for unsupervised feature selection using the maximal information compression index as the feature similarity measure [38]. The maximal information compression index is defined as the smallest eigenvalue, λ, of the covariance matrix of the set of variables under consideration. λ is zero when the features are linearly dependent and increases as the amount of dependency decrease. A threshold value equal to 95% is used from which features are considered redundant. Classification accuracy, according to calculations, was not significantly affected if varying the threshold from 95–99%.

An overall 81% dimensionality reduction was achieved with this technique. Thus, the 1460 initial descriptors were reduced to 273.

5.2. Sequential Forward Feature Selection

In order to elucidate the most discriminant descriptors, a discriminant analysis was done by means of a Sequential Forward Feature Selection (SFFS), also known as the floating search method [39]. The algorithm selects a subset of features that best predict the objects to be classified. Thus, features are sequentially added to an empty candidate set until the addition of further features does not decrease the misclassification error rate (i.e., the number of misclassified observations divided by the number of observations) of a learning algorithm (quadratic discriminant analysis). Another variant of the method is Backward Selection (SBS), in which features are sequentially removed from a full candidate set until the removal of further features increases the misclassification error.

This methodology gives a list of features ordered by discrimination capacity. In Table 5, such a list is shown for the first 100 most discriminant handcrafted features. The percentage of morphological descriptors is only 4%, although two of them are on the top. For the rest of descriptors, the percentages are as follows: statistical (28%); logGabor (40%) and LBP (30%).

Table 5. The 100 most discriminant handcrafted features listed in decreasing order of importance. LBP stands for Local Binary Patterns, M for Morphological, S for Statistical and Gn for log Gabor band $n = 1 \ldots 4$.

1	M	*Shape*	35	LBP	*Contrast (d = 3, 0°)*	69	G1	*Contrast (d = 1, 90°)*
2	S	*1st Quartile*	36	G4	*Correlation 1 (d = 1, 90°)*	70	S	*ΣAverage (d = 1, 45°)*
3	M	*Asymmetry*	37	G4	*Correlation 1 (d = 1, 135°)*	71	G4	*Entropy*
4	S	*Contrast (d = 1, 45°)*	38	S	*Energy (d = 1, 0°)*	72	LBP	*Cluster Prominence (d = 1, 0°)*
5	S	*Energy*	39	S	*Correlation 1 (d = 3, 45°)*	73	G4	*Cluster Shade (d = 1, 0°)*
6	S	*Contrast (d = 1, 0°)*	40	S	*Homogeneity 1 (d = 1, 45°)*	74	LBP	*Correlation 1 (d = 3, 0°)*
7	G4	*Correlation 1 (d = 3, 90°)*	41	S	*Correlation 1 (d = 3, 90°)*	75	S	*Entropy*
8	S	*Interquartile Range*	42	LBP	*ΔVariance (d = 1, 0°)*	76	G3	*ΔEntropy (d = 1, 45°)*
9	G4	*ΔEntropy (d = 1, 90°)*	43	G3	*Correlation 1 (d = 3, 0°)*	77	LBP	*Correlation (d = 1, 45°)*
10	LBP	*1st Quartile*	44	G2	*Contrast (d = 1, 90°)*	78	LBP	*Contrast (d = 1, 90°)*
11	LBP	*Correlation 1 (d = 3, 135°)*	45	LBP	*Correlation 1 (d = 3, 45°)*	79	G3	*Energy (d = 1, 0°)*
12	LBP	*Max. Probability (d = 1, 45°)*	46	LBP	*Correlation 1 (d = 3, 135°)*	80	G4	*Correlation 1 (d = 5, 0°)*
13	LBP	*Contrast (d = 3, 45°)*	47	LBP	*Max. Probability (d = 1, 0°)*	81	LBP	*Homogeneity 1 (d = 1, 135°)*
14	S	*Variance*	48	S	*Correlation 1 (d = 5, 90°)*	82	G4	*ΔVariance (d = 1, 90°)*
15	LBP	*Correlation 1 (d = 1, 45°)*	49	G4	*ΔVariance (d = 1, 45°)*	83	LBP	*Max. Probability (d = 3, 135°)*
16	S	*Kurtosis*	50	S	*3rd Quartile*	84	G4	*Entropy (d = 1, 0°)*
17	LBP	*Autocorrelation (d = 3, 135°)*	51	S	*Correlation 1 (d = 5, 90°)*	85	G4	*Correlation 1 (d = 1, 90°)*
18	G4	*ΔEntropy (d = 3, 0°)*	52	S	*Dissimilarity (d = 5, 135°)*	86	G3	*ΔVariance (d = 1, 45°)*
19	S	*Autocorrelation (d = 1, 0°)*	53	S	*Max. Probability (d = 1, 0°)*	87	G4	*Correlation (d = 1, 0°)*
20	G4	*Dissimilarity (d = 1, 135°)*	54	S	*Homogeneity 1 (d = 1, 90°)*	88	LBP	*Correlation 1 (d = 3, 90°)*
21	G3	*ΔEntropy (d = 1, 90°)*	55	S	*Correlation (d = 1, 0°)*	89	G3	*Correlation 1 (d = 1, 90°)*
22	LBP	*Contrast (d = 1, 135°)*	56	G4	*Correlation 1 (d = 3, 0°)*	90	G3	*Contrast (d = 1, 90°)*
23	LBP	*Contrast (d = 5, 45°)*	57	S	*Correlation 1 (d = 3, 135°)*	91	G4	*Cluster Prominence (d = 1, 0°)*
24	LBP	*Contrast (d = 3, 90°)*	58	LBP	*Contrast (d = 5, 90°)*	92	G3	*Correlation (d = 1, 90°)*
25	S	*Sum Average (d = 1, 0°)*	59	LBP	*Correlation 1 (d = 5, 135°)*	93	G4	*Contrast (d = 1, 90°)*
26	LBP	*Contrast (d = 5, 0°)*	60	G4	*Correlation 1 (d = 5, 90°)*	94	M	*Area*
27	G4	*Correlation 1 (d = 1, 45°)*	61	LBP	*Correlation 1 (d = 1, 90°)*	95	G3	*Cluster Shade (d = 1, 0°)*
28	G4	*Dissimilarity (d = 1, 45°)*	62	S	*Correlation1 (d = 1, 45°)*	96	G2	*Dissimilarity (d = 1, 135°)*
29	S	*Homogeneity 1 (d = 1, 0°)*	63	LBP	*Homogeneity 1 (d = 3, 135°)*	97	G3	*Correlation 1 (d = 3, 135°)*
30	G4	*Kurtosis*	64	LBP	*Homogeneity 1 (d = 3, 45°)*	98	M	*Eccentricity*
31	G4	*Dissimilarity (d = 1, 45°)*	65	G3	*Correlation (d = 1, 45°)*	99	LBP	*Max. Probability (d = 3, 45°)*
32	G4	*Energy (d = 1, 0°)*	66	G3	*Correlation 1 (d = 1, 135°)*	100	LBP	*Correlation 1 (d = 1, 0°)*
33	LBP	*Homogeneity 1 (d = 1, 0°)*	67	G3	*Kurtosis*			
34	G3	*Dissimilarity (d = 1, 135°)*	68	S	*2nd Quartile*			

6. Classification

A classifier is a function that maps the features extracted from descriptors to an output probability. This output is the probability that the input features belong to a given class. There are many approaches to develop these functions; it is out of the scope of this paper to explain these methods in detail. For this, the reader is referred to [40]. The purpose here is to find a suitable set of discriminant features and the classifier that best maps these features to the correct taxon. Thus, we have compared an extensive range of classifiers like nearest-neighbor, *k*-means and SVM, Decision Trees (DT) by means of random forest and bagging trees, the quadratic Bayes normal classifier and the Fisher classifier. The methods used here are all supervised learning methods. The best results were obtained with SVM and bagging decision trees. SVM achieved an accuracy up to 95.38% and bagging DT up to 98.11%; therefore, we show here the results of the latest method.

To train and test the algorithms, a 10-fold cross-validation has been applied, that is the full set of images was split into a training and a test set, where nine folds was used for training and the remaining one in each iteration for testing. Both the training and test processes are based on a random selection of samples; therefore, the ordering of the taxa does not affect the accuracy. Experiments with leave-one-out were done for 20 classes, and the results were very similar, with just 0.04% less accuracy than using 10 fcv; however, since the process takes too long for more classes, we decided to report the widely-used 10 fcv for the 80 classes.

Bagging Trees

A Decision Tree (DT) is a method in which classification is performed through a tree graph. The input feeds an initialization node (root node) from which a given test sample is tested at each stage (internal node) of the classification tree, all the way down through the leaves or internal node to the end of a tree branch or terminal node. The 'path' followed by the sample depends on the conditions associated with each internal node. These conditions are established during training rather manually or automatically. To select automatically the optimal conditions, DT algorithms consist of testing all potential variables and selecting the variable that maximizes a given criterion.

The bagging tree classifier is used to improve robustness and classification accuracy. Bagging improves variance by averaging/majority selection of the outcome from multiple fully-grown trees on variants of the training set. It uses bootstrap with replacement to generate multiple training sets. All trees are fully-grown binary trees (unpruned), and at each node in the tree, one searches over all features for splitting a node, that is to find the feature that best splits the data at that node [41]. In this study, a bagging tree with 200 learners and 30 splits provided the best results. All of the experiments were done using the classification learner apps of MATLAB R2016a.

7. Results

The results are organized into several experiments to show concrete aspects of descriptors and classifiers. To this end, the list of handcrafted feature descriptors (1460 in total) after feature reduction with 95% correlation (273 features) and the selected classifier (bagging DT) are used. Notice that the morphological, statistical and moment-based features are invariant to rotation and mirroring, but not the LBP and log Gabor features used in this study. Therefore, in total, only 255 features out of the 1460 are invariant to rotation and mirroring. This is reduced to 42 features after the correlation process, which means that only 15% of features are invariant to the data augmentation performed.

While data augmentation may introduce some bias in the experiments, mainly related to the invariant features, our aim in adding the augmentation was to compare classic methods with deep learning (see [24]), for which the same augmentation is done. Moreover, as mentioned above, most features (85%) are not invariant to rotation and mirroring, and therefore, data represented with these features can be considered without bias.

7.1. Experiment 1: Comparing Descriptor Types

All handcrafted feature descriptors mentioned above were tested here according to their category (see Table 2). The selected classifier was bagging DT.From the plot in Figure 7, most descriptors lead to similar classification results when they operate individually, with an overall accuracy around 95% for bagging DT. Furthermore, two of them are above the average (95%); these are LBP and statistical, though the moments are very low. This supports morphological and textural features. Thus, it corroborates that macro-features' analysis obtained from morphological descriptors and textural patterns already provides competitive accuracy compared to local micro-feature analysis done by spatio-frequential descriptors. Experiment 1, shown in Figure 7, has been carried out with 100 samples per class to avoid any bias of the data. A similar conclusion was obtained with 300 samples per class.

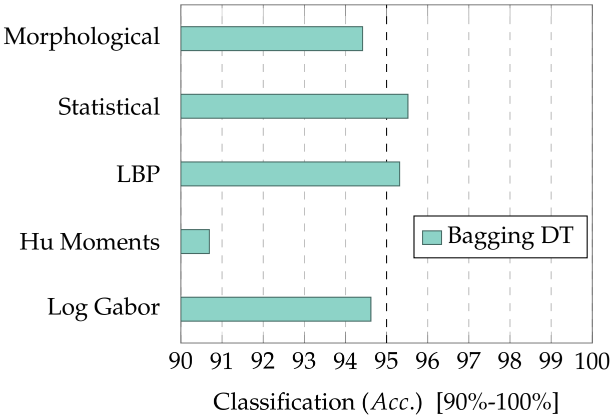


Figure 7. Comparison of different categories of descriptors under the bagging DT classifier.

7.2. Experiment 2: Combinations of Descriptor Types

The discrimination capacity of the classifiers may be enhanced by the combination of different types of descriptors. In Figure 8, morphological and statistical descriptors provide high accuracy rates. These descriptors bring together the two main handcrafted discriminating features: shape and texture. Thus, morphological and statistical descriptors constitute a baseline for comparison. The results show that statistical descriptors combined with morphological descriptors provided together an improvement of 3%. The combination of statistical with texture (LBP) improves the performance of the morphological descriptors. This proves the relevance and importance of textural features versus morphological properties. The combination of statistical with the space-frequency descriptors provided an overall accuracy improvement of 2%, while the combination of all features (textural + morphological + frequency + statistical) is the most accurate with an improvement of 4%. Note that the moments were discarded due to the low individual performance they gave and also the low relevance obtained when calculating the SFFS (see Section 5.2).

Thus, the combination of morphological + textural + space-frequency + statistical descriptors provided the lowest error rate, with an overall accuracy (*Acc.*) equal to 98.11% using a bagging DT with 10 fcv. The combination of textural and space-frequency (LBP + Log Gabor), that is using the data represented without any augmentation bias, provided an overall accuracy of 97.63% using bagging DT with 10 fcv. This is highlighted in Figure 8.

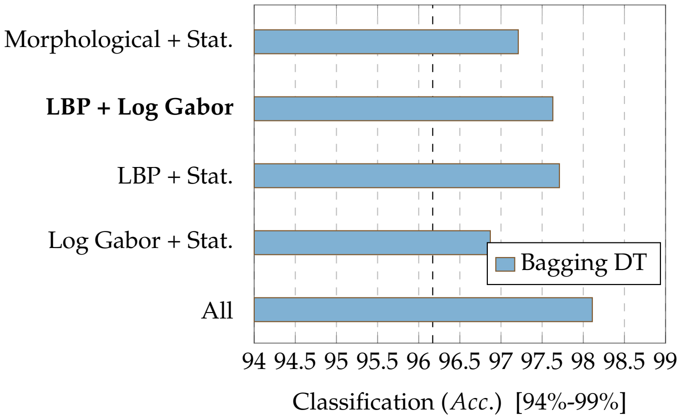


Figure 8. Accuracy performance combining descriptor types and using 300 samples per class.

7.3. Experiment 3: Dataset Dimension

Some questions arise at this point about the database dimensionality. How does this affect the performance? Is the classifier able to learn more with a higher number of samples? Is the classifier able to get the same performance with a lower number of data, that is without the use of augmented data? Thus, we tested with several subsets of 20 diatom types and 40 diatom types with 300 samples per diatom; a subset with the 35 classes that had a minimum of 100 samples per diatom; as well as with 80 diatom types with 2000 samples per diatom versus the previous dataset with 300 samples per diatom. To extend the dataset, a data augmentation with rotations every 2° was performed.

Figure 9 shows the results when classifying with the bagging DT and 10 fcv. It represents the minimum and maximum *Acc.* value obtained in the different trials. The decrease in the number of classes with 300 samples per class increased its accuracy, though not significantly. Additionally, the increase in the number of samples lowered the error. The use of 100 samples per class classifying 35 classes with no data augmentation decreased accuracy to 96.25% versus 98.11% with 300 samples per class. A similar decrease happens when classifying 20 classes, where an accuracy of 97.56% (100 samples/class without data augmentation) is obtained versus 98.82% (300 samples/class with data augmentation).

Results are usually illustrated with confusion matrices. Each column of a confusion matrix represents the instances in a predicted class while each row represents the instances in an actual class. Since there are many classes in this study, heat maps have been used to display the % of correct and incorrect classifications. A heat map displays the confusion matrix as an image whose color intensities reflect the magnitude of its values. In this case, green values indicate the % of correct classifications and pink-red values indicate the % of instances incorrectly classified. These percentages, as well as the true positive rate (green column) and false negative rate (pink-red column) for each class are shown when no more than 20 classes are classified, as well as the true positive rate (green column) and false negative rate (pink-red column).

The confusion matrix heat map for the classification of the 80 diatom classes with the bagging tree is shown in Figure 10. Figure 11 shows the confusion matrix heat map for the classification of the diatom classes without data augmentation. Figure 11 shows 20 classes chosen randomly for visualization purpose. The confusion matrix heat map of several trials done with the 20 classes is shown in Figure 12. The main diagonal of Figure 12 shows how some classes obtained 100% correct classification.

Figure 13 shows those diatoms species that produce the major misclassification errors, i.e., *Nitzschia costei*, *Gomphonema angustatum*, *Gomphonema micropumilum*, *Gomphonema minusculum* and *Gomphonema minutum*. Automation methods may be a tool to help both the expert and the non-expert, but in difficult cases, it should always be the expert who makes the final decision. We believe that it is the diatomist who must consider if an automatic classification system is acceptable, assuming a certain level of accuracy to classify different taxa. A reject option is also possible, so that only difficult cases are presented to the diatomist, while the easy ones are classified automatically.

Notice that this methodology may be applied to other taxa not included in this study. Moreover, it is currently being applied to another 20 different taxa not shown in this study, and similar results are being obtained.

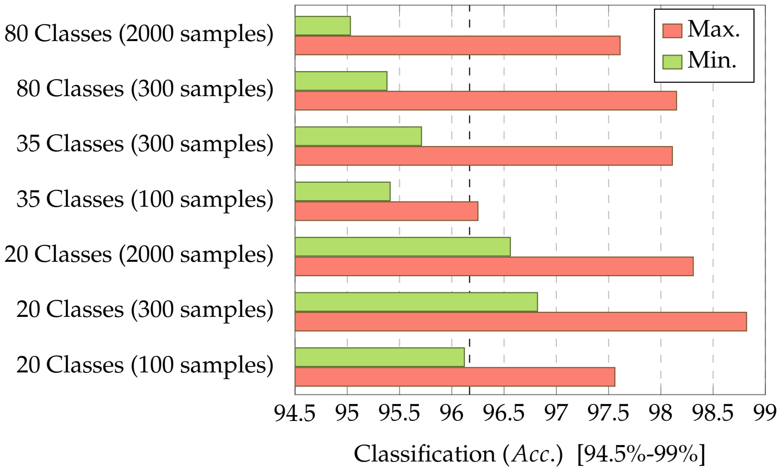


Figure 9. Accuracy performance with different numbers of classes using the bagging DT classifier. No data augmentation is applied when using 100 samples per class.

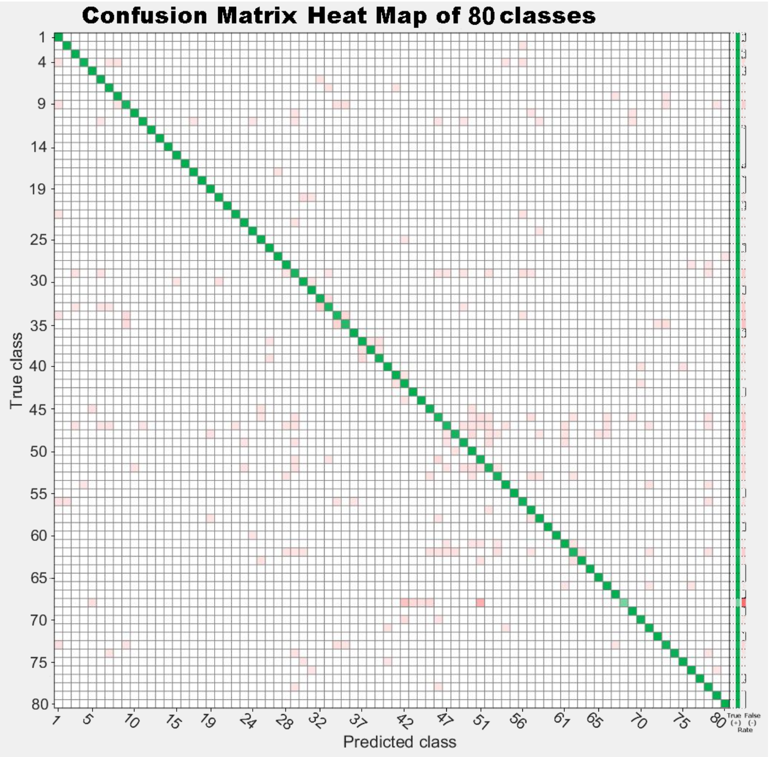


Figure 10. Confusion matrix classifying 80 classes with the bagging tree classifier 10-fold cross-validation (10 fcv).

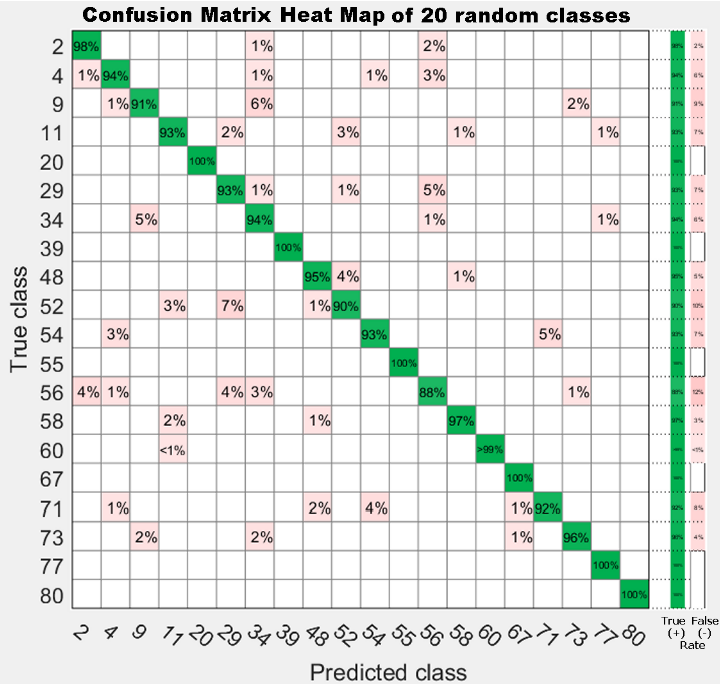


Figure 11. Confusion matrix classifying 20 classes out of the 35 classes with 100 samples per class with no data augmentation and using the bagging tree classifier 10 fcv.

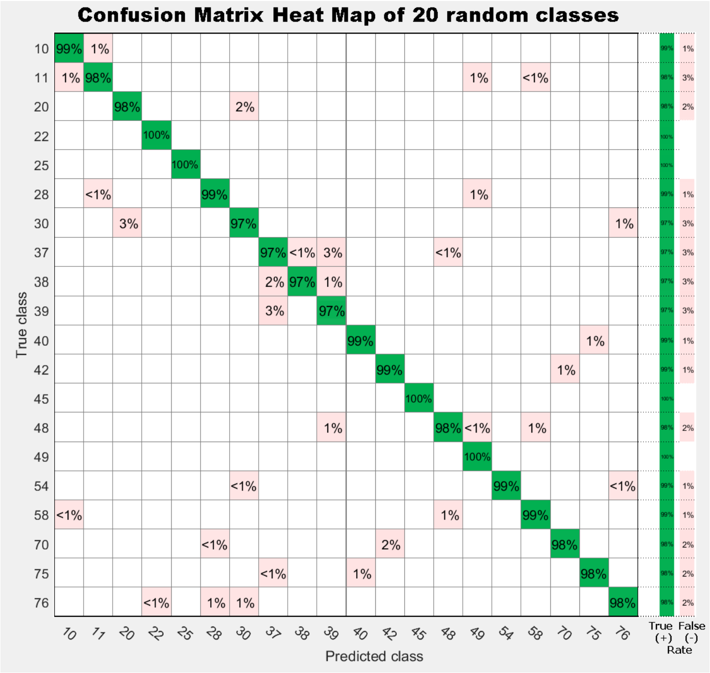


Figure 12. *Cont.*

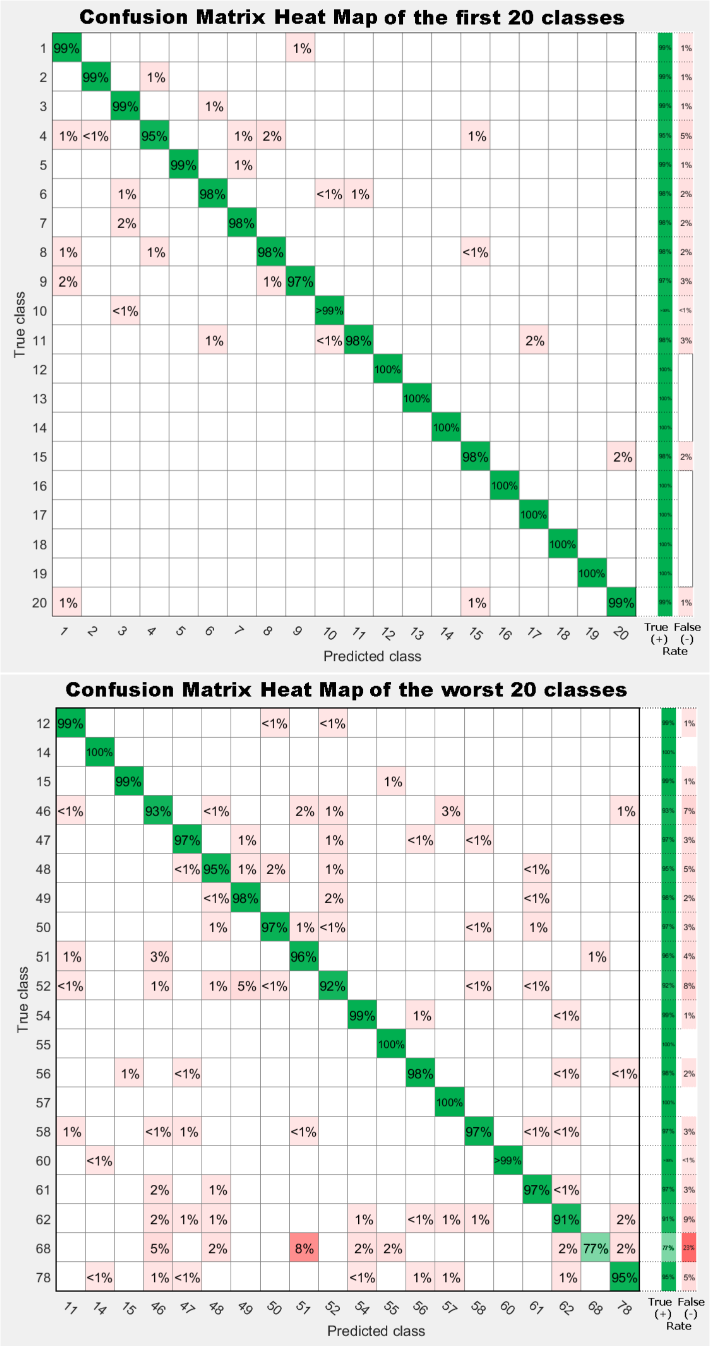


Figure 12. Confusion matrix map classifying 20 classes with the bagging tree classifier 10 fcv.

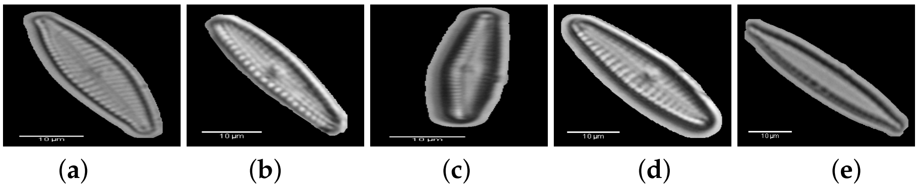


Figure 13. Diatoms species that produce the major misclassification errors: (**a**) *Gomphonema angustatum*; (**b**) *Gomphonema micropumilum*; (**c**) *Gomphonema minusculum*; (**d**) *Gomphonema minutum*; and (**e**) *Nitzschia costei*.

8. Conclusions

The main contributions of this work are: (a) big database collection: a big database of diatoms composed of 80 diatom types with an average of 100 distinct diatom valves per type manually cropped has been collected; this database is being extended to 100 types; and (b) the study of the main handcrafted features to classify diatoms, proposing an efficient method for classification.

Thus, the state of the art in morphological, statistical and texture descriptors has been exhaustively tested for classifying 80 diatom types. Some of them, like texture based on LBP and log Gabor descriptors, had not been evaluated before in this field. The combination of these features provided an improvement up to 98.11% accuracy compared to previous related works. We concluded that the combination of different descriptors categories, such as morphological and statistical descriptors, together with space-frequency representations, specifically log Gabor and texture by means of LBP, provided the best classification accuracy rates.

Along this research, we also come across several other challenging issues, such as the classification of the same diatom type during its life cycle and from different views. A diatom type may present different appearances according to its view with respect to the z-view, and consequently, its morphological and also statistical descriptors can drastically vary. This causes the morphological descriptors to not always analyze the most discriminating features. This could be solved by having more data representing all possible cases or alternatively by using a hierarchical tree classification. Thus, it could be recommended to simplify the classification using morphological refinement in two classes: centric diatoms and pennate diatoms. Afterwards, depending on the taxon, the search would focus on some other details, like texture, striae density (number of stria per micron), lineolae density (pixels per micron), etc. It should be noted that some diatom taxa are almost the same, even for experts, which proves the difficulty of this task.

Precise segmentation is a critical point for the whole classification process. This is a limitation for handcrafted feature approaches. Effectively, segmentation should be done accurately, although our purpose in the current study focuses on comparing descriptor's discriminant capacity. Therefore, we do not pursue perfect binary masks, but suitable enough to be equally shared by all descriptors.

In order to handle the above-mentioned difficulties, the authors suggest to explore new classification techniques based on deep learning (Convolutional Neural Networks (CNN)) able to learn further from larger datasets and without segmentation. In this study, it has been proven that the accuracy in traditional methods does not always improve with augmented data (it is also counterproductive due to orientation-invariant features and possible overfitting). An accuracy up to 97.56% was obtained classifying 20 classes with no data augmentation (100 samples/class) versus 98.82% with data augmentation (300 samples/class).

Acknowledgments: The authors acknowledge financial support of the Spanish Government under the Aqualitas-retos project (Ref. CTM2014-51907-C2-2-R-MINECO), http://aqualitas-retos.es/en/.

Author Contributions: Gloria Bueno designed and performed the experiments. She wrote the paper, and she is the corresponding author. Oscar Deniz designed and conceived of the experiments. He also revised the manuscript and supervised the research. Anibal Pedraza prepared the original and augmented the database. Jesús Ruiz-Santaquitaria performed part of the data acquisition and some graphics of the figures. Jesús Salido prepared the database to be available for dissemination. Gabriel Cristóbal supervised the research. María Borrego-Ramos performed part of the data acquisition. Saúl Blanco performed part of the data acquisition, and he made the annotations of the entire original database.

Conflicts of Interest: There is no conflict of interest.

References

1. Ector, L.; Rimet, F. Using bioindicators to assess rivers in Europe: An overview. In *Modelling Community Structure Infreshwater Ecosystems*; Lek, S., Scardi, M., Verdonschot, P.F.M., Descy, J.-P., Park, Y.-S., Eds.; Springer: Berlin, Germany, 2005; Chapter 1, pp. 7–19.
2. Wua, N.; Dong, X.; Liu, Y.; Wang, C.; Baattrup-Pedersen, A.; Riis, T. Using river microalgae as indicators for freshwater biomonitoring: Review of published research and future directions. *Ecol. Indic.* **2017**, *81*, 124–131.

3. Blanco, S.; Becares, E.; Cauchie, H.; Hoffmann, L.; Ector, L. Comparison of biotic indices for water quality diagnosis in the Duero Basin (Spain). *Arch. Hydrobiol. Suppl. Large Rivers* **2007**, *17*, 267–286.
4. Round, F.E.; Crawford, R.M.; Mann, D.G. *Diatoms: Biology and Morphology of the Genera*; Cambridge University Press: Cambridge, UK, 1990.
5. Mann, D. The species concept in diatoms. *Phycologia* **1999**, *38*, 437–495.
6. John, D. Use of Algae for Monitoring Rivers III. *J. Appl. Phycol.* **1999**, *11*, 596–597.
7. Hicks, Y.A.; Marshall, D.; Rosin, P.; Martin, R.R.; Mann, D.; Droop, S. A model of diatom shape and texture for analysis, synthesis and identification. *Mach. Vis. Appl.* **2006**, *17*, 297–307.
8. Smol, J.; Stoermer, E. *The Diatoms: Applications for the Environmental and Earth Sciences*; Cambridge University Press: Cambridge, UK, 2010.
9. European Standard, EN 14407: 2004. *Water Quality—Guidance Standard for the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters*; Technical Report; European Commission: Brussels, Belgium, 2004.
10. Wayne, R. *Light and Video Microscopy*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 2014.
11. Desikachary, T. Electron microscope studies on diatoms. *J. Microsc.* **1956**, *76*, 9–36.
12. Pappas, J.; Kociolek, P.; Stoermer, E. Quantitative morphometric methods in diatom research. *Nova Hedwig. Beih.* **2014**, *143*, 281–306.
13. Kloster, M.; Kauer, G.; Beszteri, B. SHERPA: An image segmentation and outline feature extraction tool for diatoms and other objects. *BMC Bioinform.* **2014**, *15*, 1.
14. Cairns, J.; Dickson, K.; Pryfogle, P.; Almeida, S.; Case, S.; Fournier, J.; Fuji, H. Determining the accuracy of coherent optical identification of diatoms. *J. Am. Water Resour. Assoc.* **1979**, *15*, 1770–1775.
15. Culverhouse, P.; Simpson, R.G.; Ellis, R.; Lindley, J.; Williams, R.; Parisini, T.; Reguera, B.; Bravo, I.; Zoppoli, R.; Earnshaw, G.; et al. Automatic classification of field-collected dinoflagellates by artificial neural network. *Mar. Ecol. Prog. Ser.* **1996**, *139*, 281–287.
16. Pech-Pacheco, J.; Alvarez-Borrego, J. Optical-digital system applied to the identification of five phytoplankton species. *Mar. Biol.* **1998**, *132*, 357–365.
17. Pech-Pacheco, J.; Cristobal, G.; Alvarez-Borrego, J.; Cohen, L. Automatic system for phytoplanktonic algae identification. *Limnetica* **2001**, *20*, 143–158.
18. Du Buf, H.; Bayer, M. Series in Machine Perception and Artificial Intelligence. In *Automatic Diatom Identification*; World Scientific Publishing Co.: Singapore, 2002.
19. Pappas, J.; Stoermer, E. Legendre shape descriptors and shape group determination of specimens in the Cymbella cistula species complex. *Phycologia* **2003**, *42*, 90–97.
20. Du Buf, H.; Bayer, M.; Droop, S.; Head, R.; Juggins, S.; Fischer, S.; Bunke, H.; Wilkinson, M.; Roerdink, J.; Pech-Pacheco, J.; et al. Diatom identification: A double challenge called ADIAC. In Proceedings of the International Conference on Image Analysis and Processing, Venice, Italy, 27–29 September 1999; pp. 734–739.
21. Dimitrovski, I.; Kocev, D.; Loskovska, S.; Dzeroski, S. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecol. Inform.* **2012**, *7*, 19–29.
22. Kuang, Y. *Deep Neural Network for Deep Sea Plankton Classification*; Technical Report; Stanford University: Stanford, CA, USA, 2015.
23. Dai, J.; Yu, Z.; Zheng, H.; Zheng, B.; Wang, N. A Hybrid Convolutional Neural Network for Plankton Classification. In *Lecture Notes in Computer Science—Computer Vision, ACCV 2016 Workshops*; Chen, C.-S., Lu, J., Ma, K.-K., Eds.; Springer International Publisher: Amsterdam, The Netherlands, 2017; Volume 10118, pp. 102–114.
24. Pedraza, A.; Deniz, O.; Bueno, G.; Cristobal, G.; Borrego-Ramos, M.; Blanco, S. Automated Diatom Classification (Part B): A deep learning approach. *Appl. Sci.* **2017**, *7*, 460.
25. Lai, Q.T.; Lee, K.C.; Tang, A.H.; Wong, K.K.; So, H.K.; Tsia, K.K. High-throughput time-stretch imaging flow cytometry for multi-class classification of phytoplankton. *Opt. Express* **2016**, *24*, 28170–28184.
26. Blanco, S.; Bécares, E.; Hernández, N.; Ector, L. Evaluación de la calidad del agua en los ríos de la cuenca del Duero mediante índices diatomológicos. *Publ. Téc. CEDEX Ing. Civ.* **2008**, *148*, 139–143.
27. Haralick, R.; Shanmugam, K.; Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621.
28. Wang, L.; He, D. Texture classification using texture spectrum. *Pattern Recognit.* **1990**, *23*, 905–910.

29. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th International Conference on Pattern Recognition-Conference A: Computer Vision Image Processing (IAPR), Jerusalem, Israel, 9–13 October 1994; Volume 1, pp. 582–585.
30. Nava, R.; Cristobal, G.; Escalante-Ramirez, B. A comprehensive study of texture analysis based on local binary patterns. *Proc. SPIE* **2012**, *8436*, 84360E–84372E.
31. Sahu, H.; Bhanodia, P. An Analysis of Texture Classification: Local Binary Patterns. *J. Glob. Res. Comput. Sci.* **2013**, *4*, 17–20.
32. Ojala, T.; Pietikäinen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987.
33. Nava, R.; Escalante-Ramírez, B.; Cristóbal, G. Texture Image Retrieval Based on Log-Gabor Features. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Alvarez, L., Mejail, M., Gomez, L., Jacobo, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7441, pp. 414–421.
34. Hu, M.K. Visual Pattern Recognition by Moment Invariants. *IRE Trans. Inf. Theory* **1962**, *IT-8*, 179–187.
35. Fischer, S.; Sroubek, F.; Perrinet, L.; Redondo, R.; Cristóbal, G. Self Invertible Gabor Wavelets. *Int. J. Comput. Vis.* **2007**, *75*, 231–246.
36. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
37. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley: New York, NY, USA, 2001.
38. Mitra, P.; Murthy, C.; Pal, S.K. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 301–312.
39. Pudil, P.; Novovicova, J.; Kittler, J. Floating search methods in feature selection. *Pattern Recognit. Lett.* **1994**, *15*, 1119–1125.
40. Alpaydin, E. *Introduction to Machine Learning*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2010.
41. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.