Table 1: Number of images used per specie

Specie	Number of images
G. Bulloides	2, 848
N. Dutertrei	2, 452
G. Sacculifer	2, 400
G. Ruber	2, 912
N. Incompta	2, 784
N. Pachyderma	2, 432

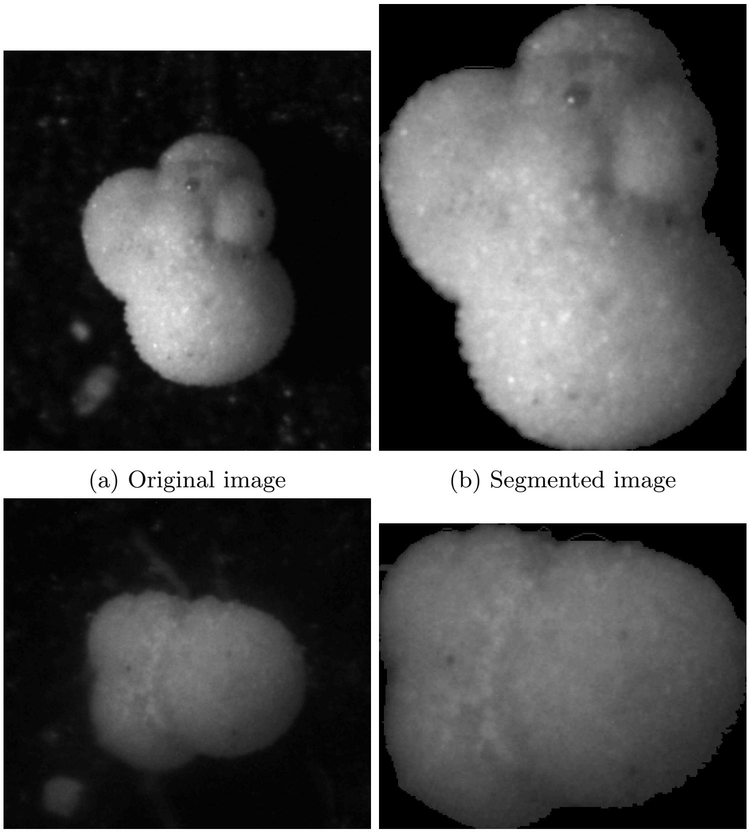
There were three main stages for the foraminifera identification:

- **Data collection**: Sample images were obtained from two main sources, the NC STATE University [2] and Centro de Estudios Científicos y Tecnológicos No. 2 ”Miguel Bernard” (C. E. C. y T. 2). A total of $15,828$ images were used. In Table 1 detailed information is provided.
- **Segmentation and preprocessing**: Once the images were collected a segmentation process was carried out to remove noise and imperfections from the image. Figure 1 shows a comparison between original images and preprocessed. This process was automated with Python and OpenCV [1].
- **Data exploration and specie identification**: The images were analysed to find useful information or features that could lead to identify which specie the foraminifera belongs to. When doing the analysis it was discovered most foraminifera species have common features that can make the neural network model more complex, which is what is needed to really obtain features that differentiate one specie from another and omit the common ones.

  Despite that, a simple neural network with 6 hidden layers was tried, but the results were not the expected as the accuracy of the model was about 16% after the first epoch, and small improvements in subsequent epochs.

  Supervised machine learning was used because every image on the dataset was labelled. Despite that, it was also decided to try with unsupervised machine learning, however when using the elbow method to identify the optimal number of clusters, what was discovered previously was confirmed. In Figure 2 a graph obtained from the elbow method, shows that, indeed, there are quite a few common features between species.

  From that it is concluded that using a simple neural network model has not good results, hence, hyper parameters should be modified or/and images should have better preprocessing.



(a) Original image

(b) Segmented image

(c) Original image

(d) Segmented image

Figure 1: Original images compared to processed images

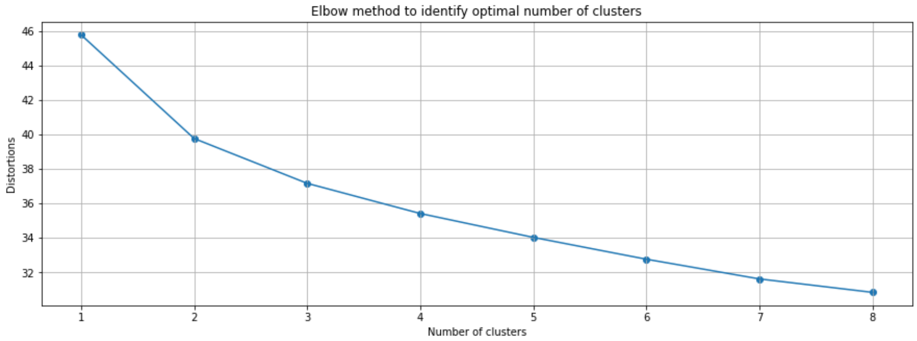


Figure 2: Elbow graph

References

[1] G. Bradski. “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (2000).

[2] R Mitra et al. “Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance”. In: *Marine Micropaleontology* 147 (2019), pp. 16–24.