Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

CIDACC: *Chlorella vulgaris* image dataset for automated cell counting

Evangelos Pistolas[1], Eleni Kyratzopoulou, Lamprini Malletzidou, Evangelos Nerantzis[2], Chairi Kiourt, Nikolaos Kazakis*

Athena - Research and innovation Center in Information, Communication and Knowledge Technologies, Xanthi 67100, Greece

ARTICLE INFO

Article history:
Received 6 August 2024
Revised 27 August 2024
Accepted 9 September 2024
Available online 14 September 2024

Dataset link: CIDACC: Chlorella vulgaris Image Dataset for Automated Cell Counting (Original data)

Keywords:
Microalgae
Clusters
Machine learning
Object detection
Semantic segmentation

ABSTRACT

This CIDACC dataset was created to determine the cell population of *Chlorella vulgaris* microalga during cultivation. *Chlorella vulgaris* has diverse applications, including use as food supplement, biofuel production, and pollutant removal. High resolution images were collected using a microscope and annotated, focusing on computer vision and machine learning models creation for automatic *Chlorella* cell detection, counting, size and geometry estimation. The dataset comprises 628 images, organized into hierarchical folders for easy access. Detailed segmentation masks and bounding boxes were generated using external tools enhancing the dataset's utility. The dataset's efficacy was demonstrated through preliminary experiments using deep learning architecture such as object detection and localization algorithms, as well as image segmentation algorithms, achieving high precision and accuracy. This dataset is a valuable tool for advancing computer vision applications in microalgae research and other related fields. The dataset is particularly challenging due to its dynamic nature and the complex correlations it presents across various application domains, including cell analysis in medical research. Its intricacies not only push the boundaries of current computer vision algorithms but also

* Corresponding author.
E-mail address: nikkazak@athenarc.gr (N. Kazakis).
Social media: @labrini_wisc (L. Malletzidou)

[1] https://x.com/_epistola.

[2] https://x.com/moeb1us?s=21&t=aWmAKFzS6eiwSP2KQ6oynA.

https://doi.org/10.1016/j.dib.2024.110941

offer significant potential for advancements in diverse fields such as biomedical imaging, environmental monitoring, and biotechnological innovations.

Specifications Table

Subject	*Computer Science / Computer Vision and Pattern Recognition*
Specific subject area	*Chlorella vulgaris* cells counting for computer vision approaches.
Type of data	Images: PNG and annotations (masks: PNG and bounding boxes: TXT)
Data collection	*Chlorella vulgaris* was cultivated in laboratory conditions. At various stages of the cultivation (different cultivation days), 1 ml of the cultivation was extracted with a pipette and diluted appropriately. Then 10 µL of the diluted cultivation were loaded into an improved Neubauer hemocytometer with a cover glass, which was placed under a biological microscope (Olympus CX43RF) for optical observation of the microalga cells. Images of the cells were acquired by means of a digital camera (Olympus EP50) attached to the microscope. A 40× objective lens was used, while uniform Illumination with consistent color temperature was applied to all samples by means of the LED source integrated in the microscope producing daylight conditions so specimens can be viewed with their natural colors. The image is in PNG format with a size of 2592×1944 pixels. Each image was taken from a new sample of *Chlorella vulgaris*.
Data source location	Laboratory: Archaeometry and Physicochemical Measurements Institution: Athena - Research and innovation Center in Information, Communication and Knowledge Technologies City: Xanthi Country: Greece Latitude and longitude: 41.13546263522446, 24.92074822120942
Data accessibility	Repository name: Zenodo Data identification number: zenodo.13219972 Direct URL to data: https://doi.org/10.5281/zenodo.13219972
Related research article	None

1. Value of the Data

- The dataset contains high-quality images and detailed annotations of *Chlorella vulgaris* (*C. vulgaris*) cells, providing a crucial resource for training deep learning models. Focusing on models that can accurately compute the number and the size of cells in an image, thereby determining the concentration and population of microalgae at various stages of cultivation. This ability is essential for optimizing cultivation processes and improving yields in biotechnological and aquaculture applications.
- The dataset is not limited to a single application. Researchers and developers can repurpose the data for various computer vision projects, including but not limited to segmentation, object detection, and classification tasks. This flexibility allows for a broader impact, potentially leading to innovations in other domains where image analysis is crucial.
- Researchers and educators can leverage the dataset to conduct in-depth analyses of *C. vulgaris* clusters. These cell clusters often present challenges in accurately counting the cells, making the dataset valuable for developing and testing new computational methods and algorithms. Educators can also use the data for teaching purposes, demonstrating the complexities and solutions in cell counting and image segmentation.
- Besides researchers, the dataset could be uses by other stakeholders interested in counting *C. vulgaris* cells, such as biologists and microbiologists for studying algal biology and ecology, the aquaculture industry for ensuring proper nutrition and water quality, nutraceutical and

supplement manufacturers for product standardization, wastewater treatment facilities for monitoring bioremediation efficiency, and biofuel producers for optimizing lipid extraction and production processes. In addition, the present dataset could also constitute the basis for clinical centers or microbiologists interested in calculating the population of various similar shape cells other than *C. vulgaris*.

- The images were captured under meticulously controlled conditions across multiple trials and are organized to represent single cells and clusters of cells. This organization ensures that the data is reliable and consistent, making it an excellent benchmark for testing the efficacy of various image analysis techniques.
- Beyond the primary focus on *C. vulgaris*, the dataset holds potential for cross-disciplinary applications. For instance, it can be used in environmental monitoring, water quality assessment, and even in medical research where similar cell counting and segmentation tasks are required. The dataset's adaptability enhances its value, encouraging innovation across different scientific and technological fields.
- Measuring the cells of C. *vulgaris* helps track the growth and productivity of the culture. This is crucial for industrial production, where consistent and high yields are necessary. Cell measurement can reveal information about the health of the culture, such as the presence of infections or nutrient deficiencies. In research, cell counting allows for the evaluation of the effects of various cultivation conditions. In biotechnology, cell number accuracy is essential for producing biofuels, dietary supplements, and other products. In research, cell measurement allows for the evaluation of the effects of various cultivation conditions.

2. Background

C. vulgaris is a widely cultivated microalga due to its high lipid content, making it a potential resource for biofuel production. Furthermore, it is also a valuable dietary supplement for humans and livestock because of its nutrients. *C. vulgaris* is often used in environmental monitoring and ecotoxicological studies due to its sensitivity to pollutants. Apart from industrial applications, it is often cultivated in laboratory photobioreactors. In every case, the biomass estimation of the microalga is critical, to monitor growth verification and cultivation conditions quality [1,2].

To accurately measure the cell number, growth, and other characteristics of *C. vulgaris*, various techniques have been employed, ranging from direct counting techniques to indirect estimations based on optical properties, including hemocytometry, flow cytometry, spectrophotometry, dry weight measurement, and automated cell counters which could fail to count accurately the *C. vulgaris* cells due to their small size. The hemocytometer is a widely used tool for manually counting cells under a microscope [3]. This method is time-consuming and labor-intensive when dealing with large-scale experiments involving many samples and can be challenging due to cell aggregation.

This dataset of *C. vulgaris* cells, featuring detailed segmentation masks and bounding boxes, advances computer vision approaches by enabling precise cell counting, size detection and concentration analysis. It mainly focuses on machine learning models training, fostering cross-disciplinary research, and facilitates innovative solutions in biotechnology, environmental sustainability, and educational applications.

3. Data Description

The dataset is organized hierarchically into multiple folders and subfolders, containing 628 images taken from a microscope and further processed by external tools. There are 2 data classes (distinct and clusters). The images that represent the distinct and the clusters are 464 and 164 respectively (Fig. 1). The data labels that represent distinct *C. vulgaris* cells are 5808 and that of clusters are 183 (Fig. 2).

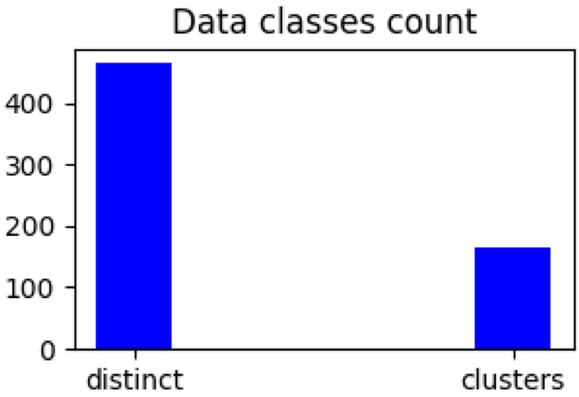


Fig. 1. Data classes count.

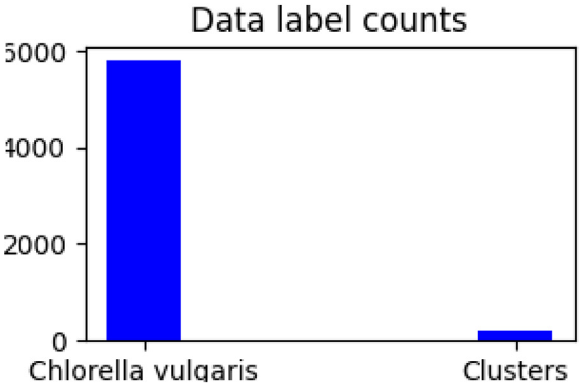


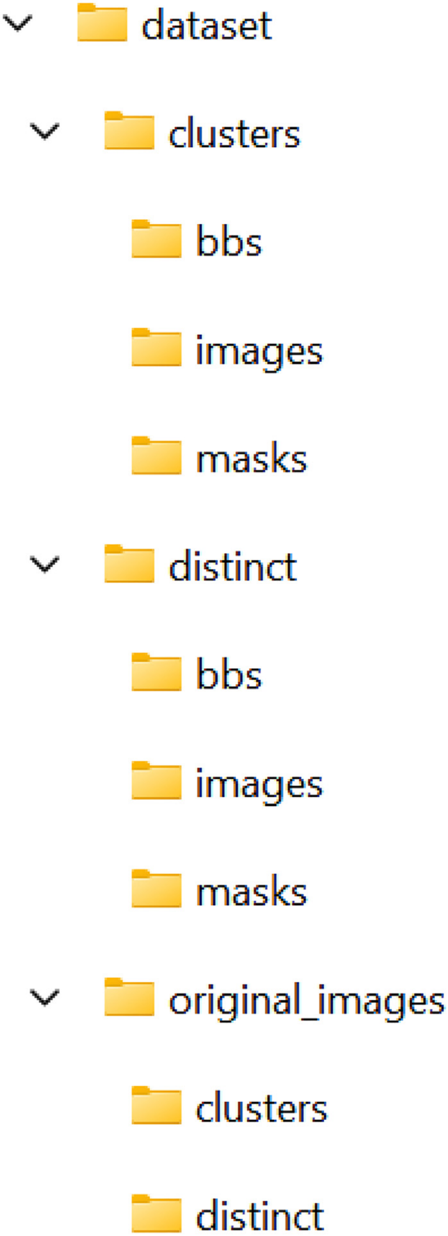
Fig. 2. Data labels counts.

Fig. 3 presents the folder structure. It consists of three root folders: “original_images”, “clusters”, and “distinct”. The “original_images” folder holds the raw microscope images with initial dimensions of 2592×1944 pixels. These images are further subdivided into “clusters” and “distinct” folders, indicating whether they contain single cells (indicative sample shown in Fig. 4) or cell clusters (indicative sample shown in Fig. 5). The “clusters” and “distinct” root folders contain annotated images with reduced dimensions (640×640 pixels). The “clusters” folder includes images showing *C. vulgaris* cells forming clusters, where counting individual cells is not possible. The “distinct” folder contains images of cells that can be counted with high precision.

Each of the folders “clusters” and “distinct” contains three additional subfolders: “images”, which holds the downscaled (to 640×640 pixels) original images in png format; “bbs”, which includes the coordinates of the bounding boxes indicating cell positions in the images in txt format; and “masks”, which contains the cell masks for image segmentation approaches in png format. Bounding boxes were produced using an open-source annotation tool, while the masks were created using Fiji software.

4. Experimental Design, Materials and Methods

A sample of mechanically stirred culture of *C. vulgaris* was placed on a glass slide and covered with a cover slip. The photos were taken with an Olympus CX43RF biological microscope



```
dataset
    clusters
        bbs
        images
        masks
    distinct
        bbs
        images
        masks
    original_images
        clusters
        distinct
```

Fig. 3. Dataset structure including folder names.

equipped with an Olympus EP50 digital camera and a 40× objective lens, while integrated LED source was used for illumination of the samples. These images were then used for data annotation with external tools. Each image was taken from a new sample of *C. vulgaris*. All the data were annotated manually both with the Roboflow and the Fiji software. Since the data were manually annotated there are no annotation errors like those that arise from an automatic process.

Fig. 4. Single cells: Original image, Mask for segmentation and Bounding boxes for detection.

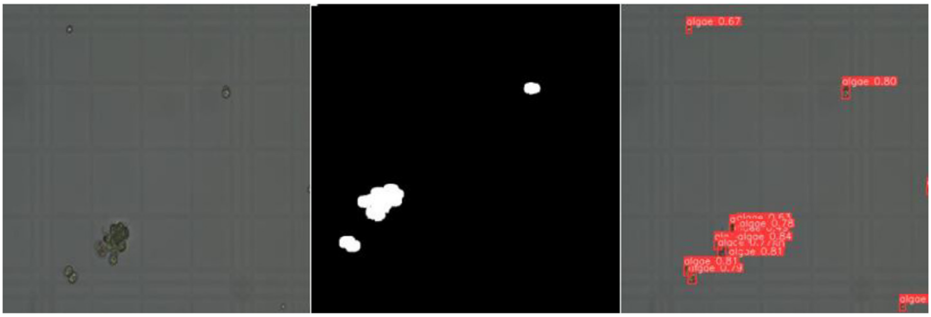


Fig. 5. Clusters cells: Original image, Mask for segmentation and Bounding boxes for detection.

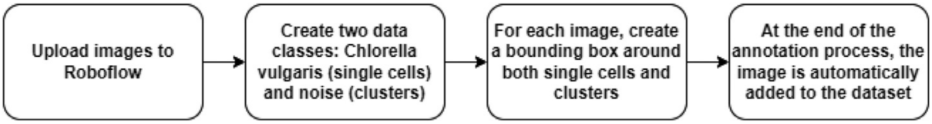

Upload images to Roboflow
Create two data classes: Chlorella vulgaris (single cells) and noise (clusters)
For each image, create a bounding box around both single cells and clusters
At the end of the annotation process, the image is automatically added to the dataset

Fig. 6. Roboflow data processing flow.

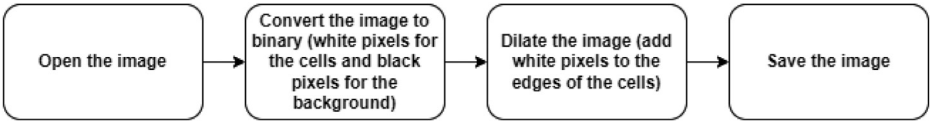

Open the image
Convert the image to binary (white pixels for the cells and black pixels for the background)
Dilate the image (add white pixels to the edges of the cells)
Save the image

Fig. 7. Fiji data processing flow.

The annotated images with bounding boxes were created using Roboflow following this procedure (Fig. 6):

1. Upload a set of images taken from the microscope.
2. Create two classes to match the cells, the *C. vulgaris* (single cells) and the noise (cell clusters).
3. For each image, create a bounding box around both single cells and clusters.
4. Export the bounding boxes to txt files. One file for each image with the same name.

The masks for the images were created using the desktop software Fiji. In Fiji, the process is as follows (Fig. 7):

1. Open an image.

2. Make it binary using the corresponding feature which converts an image to black (background) and white (cells/clusters).
3. The converted image is dilated which is a process that adds pixels to the edges of the white objects (cells and clusters). As a second step, the masks are manually processed to remove any noise of white pixels that do not correspond to cells.
4. Save the image.

To determine the effectiveness and applicability of the dataset, we tested it in two experiments. The first experiment used YOLOv8 for cell detection and localization, achieving a precision of about 0.75 and an mAP50 of about 0.605. The second experiment focused on image segmentation using a UNET architecture (4×4), which achieved a loss (training loss) of about 0.013. Both experiments highlighted the importance and impact of the dataset.

Limitations

'Not applicable'.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

CIDACC: Chlorella vulgaris Image Dataset for Automated Cell Counting (Original data) (Zenodo).

CRediT Author Statement

Evangelos Pistolas: Conceptualization, Methodology, Software, Data curation, Writing – original draft; **Eleni Kyratzopoulou:** Methodology, Investigation, Writing – original draft; **Lamprini Malletzidou:** Investigation; **Evangelos Nerantzis:** Investigation; **Chairi Kiourt:** Conceptualization, Methodology, Supervision, Writing – review & editing; **Nikolaos Kazakis:** Conceptualization, Supervision, Writing – review & editing, Resources, Project administration, Funding acquisition.

Acknowledgements

This research is financially supported under the Research Project "ALGEBRA - Innovative waste bioremediation practice for the removal of toxic compounds with the use of microalgae in the context of a circular economy", which is funded by the program "NATURAL ENVIRONMENT & INNOVATIVE ACTIONS 2022", PRIORITY AXIS 3: RESEARCH & IMPLEMENTATION, total budget: 199.501,23 €, Green Fund, Athena - Research & Innovation Center in Information Communication & Knowledge Technologies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Liu, F. Chen, Biology and industrial applications of chlorella: advances and prospects. In: Posten, C., Feng Chen, S. Microalgae Biotechnology. Advances in Biochemical Engineering/Biotechnology, vol 153. Springer (2014).

[2] L. Gómez-Luna, L. Tormos-Cedeño, Y. Ortega-Díaz, Culture and applications of Chlorella vulgaris: main trends and potential on agriculture, Chem. Technol. 42 (1) (2022) 70–93.

[3] P. Buescher, & R. Dringen, Determination of cell numbers with a Neubauer improved hemocytometer. In: Radeke H. H. Cell Culture Techniques. Methods in Molecular Biology, vol 1879. Humana Press, New York (2019).