CHEMICAL ENGINEERING TRANSACTIONS
VOL. 109, 2024
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Leonardo Tognotti, Rubens Maciel Filho, Viatcheslav Kafarov

ISBN 979-12-81206-09-0; **ISSN** 2283-9216
DOI: 10.3303/CET24109094

Advanced Simulation Model for Studying Biofuel-Producing Microalgae Populations

Bartolomeo Cosenza[a], Michele Miccio[b], Gabriele Pannocchia[a], Ignazio Sammarco[c]

[a] Dipartimento di Ingegneria Civile e Industriale, Università di Pisa, Largo Lucio Lazzarino, 56122 Pisa (PI), Italy, {bartolomeo.cosenza,gabriele.pannocchia}@unipi.it
[b] Dipartimento di Ingegneria Industriale (DIIn), Università degli Studi di Salerno, via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy, michele.miccio@unisa.it
[c] Università degli Studi di Palermo, Italy
ignazio.sammarco@you.unipa.it

Microalgae play a crucial role in various sectors, such as biofuel production or environmental monitoring. The ability to accurately classify and analyze microalgae species from optical images is vital for advancing research and applications. Generally, the data regarding the population of microalgae constitute a valuable input for machine-learning algorithms whose aim is to classify real data derived from optical images of microalgae cells. However, obtaining a diverse dataset of microalgae populations to train machine-learning models can be challenging and resource-intensive. This paper presents a machine-learning algorithm based on a dataset of algal records generated by a simulation model. The simulation model uses a combination of mathematical models, probabilistic distributions, and biological knowledge to create realistic data on microalgae populations. The dataset generated from the developed model serves as a resource for training and validation of the subsequent machine-learning model proposed for the microalgae classification task. The machine learning model, trained on this synthetic data, can subsequently be applied to efficiently classify and analyze optical images of real microalgae populations, leading to improved precision and reliability in the identification of species useful to produce biofuels.

1. Introduction

Microalgal biomass represents a promising resource for third-generation biofuels. Microalgae, in fact, are photosynthetic organisms capable of accumulating large quantities of lipids, which can be extracted and transformed into biodiesel through chemical transformation processes. This approach offers several advantages over producing biodiesel from other sources, such as traditional agricultural crops. Microalgae can be grown on land unsuitable for agriculture and do not compete directly with food resources. Furthermore, they can grow rapidly and with a high rate of efficiency in capturing atmospheric CO2 during photosynthesis, thus contributing to the reduction of greenhouse gas emissions. Beyond their biofuel potential, microalgae are prolific producers of bioactive compounds including proteins, carbohydrates, lipids, and vitamins (Xu et al., 2009). Additionally, these microorganisms serve as a sustainable reservoir of essential pigments such as carotenoids, phycobilins, docosahexaenoic acid, and eicosapentaenoic acid. These pigments exhibit significant applications in industries like cosmetics and pharmaceuticals (Chiu et al., 2017), enhancing the versatility and value of microalgae in various sectors. The characterization of microalgae cultures is essential to guarantee the quality of the biomass, in particular, the species percentages and abundance of some particular genera within the cultures are crucial for assuring good quality of the production process. The control of contamination in microalgae cultures is a major challenge for large-scale production facilities. The continuous monitoring of species within the microalgae culture is important to prevent contamination of the culture with foreign species, which in most cases can be harmful to the entire microalgal population (Di Caprio et al., 2022). Traditionally, microalgae cultures have been characterized through manual analysis of samples, in which microalgae detection and classification are done using optical microscopy. Computer vision techniques are widely used to analyze digital images in many applications such as medical images, spatial images, agriculture, and other microscopic images. Some

Paper Received: 14 January 2024; Revised: 19 February 2024; Accepted: 15 April 2024
Please cite this article as: Cosenza B., Miccio M., Pannocchia G., Sammarco I., 2024, Advanced Simulation Model for Studying Biofuel-producing Microalgae Populations, Chemical Engineering Transactions, 109, 559-564 DOI:10.3303/CET24109094

developed methods are used for online monitoring, infesting species detection in precision cultivation methods, and other automatic vision applications. Recent techniques of computer vision have been applied to microalgae identification and classification. These techniques are used for detecting, counting, identifying, and classifying algae in images while some are used to measure the density of microalgae in water, and some others to help in the process of recognition images. A commercial cell counter Flowcam (Otálora et al., 2021) is used extensively for microalgae image analysis. This tool can extract a series of descriptive variables like the shape, volume, width, height, and transparency of the cells. These descriptive data are used as input to diverse convolutional Artificial Neural Network (ANN) or Support Vector Machine (SVM) models for identifying the microalgae species present in the culture. Machine learning (ML) is widely employed in algae recognition, providing an effective approach for analyzing and classifying overall data from algae samples. The application of ML in this sector represents an advanced solution to address the challenges related to the morphological diversity and vastness of algal species. Algal recognition via ML is particularly relevant in large-scale production contexts, such as the biofuel or algae-based food supplement production industries. The ability to efficiently monitor the composition of crops and identify any anomalies helps to improve the quality and yield of production. Furthermore, the application of ML in algal recognition can be extended to environmental monitoring systems. The ability to quickly and accurately identify the algal species present in each environment helps in the assessment of water quality and the sustainable management of aquatic ecosystems. An approach for an automatic/semi-automatic classification of microalgae based on semi-supervised and active learning algorithms, using Gaussian mixture models was proposed by Drews et al. (2013). Franco et al. (2019) developed and validated a rapid methodology by use of the artificial neural network to elucidate microalgae species in suspensions. A pipeline for the automatic binary classification of living and dead microalgae was developed by Reimann et al. (2020). Barsanti et al. (2021) presented cutting-edge methods employed in the automated classification of plankton using digital microscopy. The computer-microscope system hardware and the image processing techniques used for the recognition and classification of planktonic organisms (segmentation, shape feature extraction, pigment signature determination, and neural network grouping) are described in their review. Ning et al. (2022) summarized recent advances based on various ML algorithms in microalgae applications, such as microalgae classification, bioenergy generation from microalgae, microalgae environment purification, and microalgae growth monitoring, with a view to future development of machine learning algorithms in the treatment of microalgae. Sultana et al. (2022) provided a Bayesian optimization algorithm based on ML techniques such as ANN and support vector regression as the potential tool for modeling biodiesel production using microalgae oil as feedstock. Wang et al. (2022) analyzed in detail the latest advances in ML-assisted bioenergy technology, including energy utilization of lignocellulosic biomass, microalgae cultivation, biofuels conversion, and application. A machine learning algorithm, a decision tree, was used by Singh et al. (2023) to analyze microalgae-based datasets and predict different optimized combinations of descriptor variables leading to high growth rates and microalgae biomass production. Coşgun et al. (2023) reviewed ML applications in microalgal biofuel production. Another review providing an in-depth discussion on intelligent microalgal wastewater treatment and biorefinery for researchers in the field of microalgae was proposed by Oruganti et al. (2023). Chong et al. (2023) critically examined the transition from earlier metagenomic methods to the contemporary image-based technique for the identification of microalgae. Wu et al. (2023) designed and validated a microalgae biorefinery using ML-assisted modeling of hydrothermal liquefaction. Berenguel et al. (2023) introduced an ANN model for characterizing microalgae cultures, utilizing images captured by Flowcam. Trained on diverse species from six genera, the model incorporates a classification threshold to enhance accuracy by filtering out undesired objects. Otálora et al. (2023, 2021) proposed an artificial intelligence approach for the identification of microalgae cultures. More recently Sayed et al. (2024) used artificial intelligence to identify the best operational factors of microalgae microbial fuel cell (MMFC). The proposed methodology integrates ANN modeling and a forensic-based investigation algorithm (FBI). The procedure that led to the generation of data and the creation of a learning algorithm will now be examined in detail.

2. Material and methods

2.1 Microalgae species

Among microalgae species used in the trials to develop the ANN method, the most representative were *Chlorella vulgaris* (Figure 1), *Scenedesmus almeriensis*, *Nannochloropsis oculata*, *Chroococcus turgidus* (Figure 2), *Oscillatoria okeni.* All of these were freshwater species characterized by its high growth rate and tolerance to wide ranges of growing conditions such as temperature, pH and oxygen demand. Moreover, all the selected species are known to produce a high content of triacylglycerol (TAGs) (Dos Santos et al., 2016). A major resource of images and characteristics of microalgae was the AlgaeBase database. Currently, the database contains 174.018 species and infraspecific names, 23.483 images, 70.139 bibliographic items, and 544.233 distributional records (Guiry and Guiry, 2024).

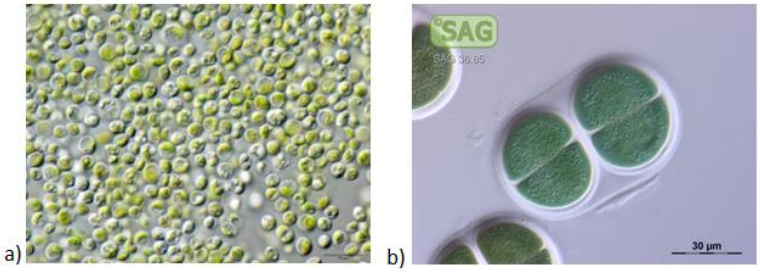

Figure 1. a) Chlorella vulgaris (By ja:User:NEON / User:NEON_ja - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=35450226) b) Chroococcus turgidus (By M. Lorenz: 2016-08-18 - http://sagdb.uni-goettingen.de/detailedList.php?str_number=36.85University of Göttingen: Department Experimental Phycology and Culture Collection of Algae (EPSAG), CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=86078103).

2.2 Synthetic data generation

A synthetic random dataset was generated representing a population of microalgae cells. This synthetic dataset was designed to encompass a broad range of microalgae population scenarios, overcoming the limitations of real-world data scarcity. The model that generated the data was developed considering information from many publications regarding morphology, shape, biochemical, and kinetics of algal growth. A Poisson distribution integrated in Python ver. 3.8 programming language, with the auxilium of some mathematics libraries like Numpy and data analysis framework, was applied for synthetic data generation. The data included a series of records containing descriptive features of the microalgal cells as width (µm), growth rate (doublings/day) and chlorophyll content (mg/L). An initial dataset of 2000 individuals was generated to mimic the distribution of an existing dataset of five species present in the marine and freshwater environment of Palermo. ML algorithms require a substantial amount of data for effective model training. In data science, synthetic data plays a very important role. This data allows scientists to test a new algorithm under controlled conditions. In other words, it is possible to generate data that tests a very specific property or behavior of the ML algorithm. For example, the learning algorithm can be tested on balanced and unbalanced datasets, or its performance can be evaluated by subjecting it to different levels of noise. In this way, it is possible to establish a baseline of the algorithm's performance in various scenarios. Synthetic data also finds application in the medical field. In medical images, analysis, and scientific research datasets are often small and expensive to acquire, and annotations are often sparse (Buslaev et al., 2020). Very extensive image augmentations have become commonly employed to address this need. Data augmentation has emerged as a valuable technique for artificially expanding the training data set by generating modified replicas of existing data (Mikołajczyk and Grochowski, 2018). This involves introducing slight modifications to the dataset or employing deep learning methods to produce novel data points with a generation of synthetic data without using the original dataset. Augmented data is derived from the original dataset with minor alterations. In the context of image augmentation, geometric and color space transformations such as flipping, resizing, cropping, brightness adjustments, and contrast enhancements are applied to amplify the size and diversity of the training set. The main reason for generating synthetic data on microalgae populations, despite the existence of the AlgaeBase database, is that real data may not cover all possible variations or conditions that can occur in nature. Generating synthetic data enables the creation of more diverse scenarios and simulates conditions that may not be present in existing data. Furthermore, synthetic data can be used to strengthen machine learning models, helping them to generalize better across new data and improve performance in classifying microalgae images in real-world settings.

2.3 Machine learning for species detection

Machine learning training and classification models were constructed using Python with Scikit-learn (an open-source machine-learning library for the Python programming language that provides a wide range of tools and algorithms for classification, regression, clustering, and more), Support Vector Machine (SVM) and the Pandas 2.1.4 library. The dataset utilized in this study and elaborated by the ML algorithm consists of 2000 records from the microalgae population generated by the synthetic data generator developed in this work. The chosen machine learning algorithm for this study is the SVM, a powerful supervised learning model commonly used for classification and regression tasks. Unlike neural networks, which leverage layers of interconnected nodes, SVM focuses on an optimal decision boundary approach and excels in scenarios with limited training data and a well-defined problem structure, proving particularly advantageous. The goal is to effectively separate classes in the data. SVM. However, unlike neural networks, SVM can have limitations in handling increasing

complexities and learning intricate representations of data. The choice between SVM and neural networks depends on the specific needs of the problem and the nature of the data available, with SVM often preferred in situations with limited data and well-defined structures. In the context of this study, SVM was employed for the classification of microalgae species based on a previously cited list of features. To evaluate the model's performance, the dataset was split into training and testing sets. The training dataset, comprising 80% of the data, was used to train the SVM model, while the remaining 20% served as a test set for model evaluation (Table 1). The SVM model was fitted to the training data using the kernel function, radial basis function (RBF, a kernel function that helps capture complex relationships between data, making it possible to separate classes more flexibly in machine learning models), and other relevant parameters.

Table 1. Synthetic data generated and ratio between training data and validation data.

	Training Subset	**Validating Subset**	**Total**
Sample number	1400	600	2000
Sampling ratio	0.7	0.3	1

The performance of the SVM model was assessed using various metrics, including accuracy, precision, and recall. These metrics provide insights into the model's ability to correctly classify instances and its overall predictive accuracy. The computations were executed on a PC with 32GB of RAM, 1 Tb SSD Hard disk and a GPU with 8GB of RAM, to ensure computational efficiency.

3. Results and Discussion

Some results deriving from the data generator and the learning algorithm will now be shown and commented. Figure 2 shows the size distribution of microalgal cells (one of the descriptive characteristics of microalgal cells) of the synthetic data generated for Chlorella vulgaris specie. Cell size may be particularly relevant when considering aspects such as the distribution of microalgae in an environment and their ability to compete for environmental resources.

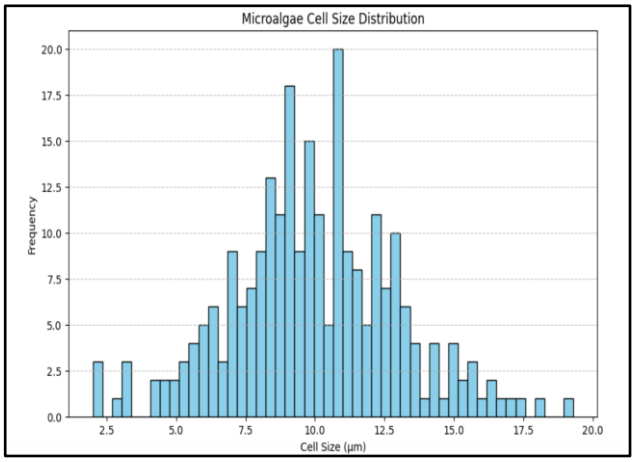

Figure 2. Microalgal size distribution of generated synthetic data for Chlorella vulgaris specie, using normal distribution model of Numpy library.

The developed machine-learning algorithm, SVM, based on the synthetic dataset generated by the simulation model, demonstrated promising outcomes in the detection and classification of microalgae species.
The SVM algorithm proved to be an effective choice, achieving high accuracy, precision, and recall, demonstrating robust performance in accurately classifying microalgae instances. In particular, the ML algorithm performance can be evaluated using a “confusion matrix”, automatically generated by the program.
In Figure 3 the confusion matrix predicted for 5 species is shown. This matrix is a table used in evaluating the performance of a classification model and represents the number of instances of each class that were correctly or incorrectly predicted by the model. On the abscissa, indicated as ‘Predicted label’, the species identified by the algorithm is reported. The ordinate shows the ‘True label’, the real species that the algorithm should identify. Higher values on the main diagonal indicate correct identifications, highlighting accurate performance, as in the case of the Scenedesmus species (cell at row 1, column 1, the value is 120). Lower numbers outside the diagonal indicate incorrect identifications, but the error rate is still low overall. For example, the value 0 in the cell at row 4, column 3 indicates that Chlorella has never been confused with Nannochloropsis species, but the same Chlorella has been confused 28 times with Chroococcus species (cell at row 5, column 3).

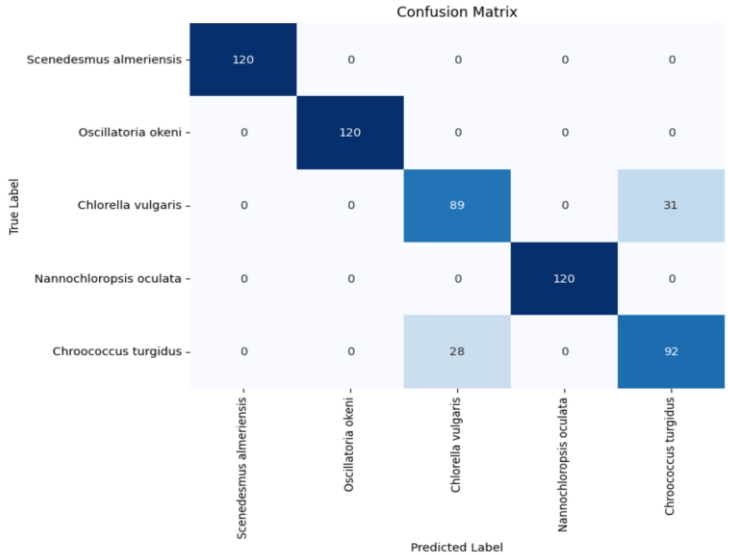

Figure 3. Confusion matrix for 5 species predicted.

These results suggest that the proposed approach, utilizing a simulation model for synthetic data generation, has the potential to revolutionize the training process for machine-learning algorithms in the field of microalgae classification. This application can help augment training data, often incomplete or completely lacking, for machine learning algorithms. The ability to bypass data scarcity issues and efficiently train models on a realistic yet controlled dataset opens new avenues for advancing research and applications in biofuel production, environmental monitoring, obtaining axenic culture, preventing foreign or unwanted species or strains within the culture, and beyond. In Table 2 accuracy and cross-validated accuracy values are reported for the algorithm.

Table 2. Accuracy of model results in species identification

	SVM algorithm
Accuracy	90.16%
Cross-validated Accuracy	89.85%

The inclusion of environmental factors such as light intensity, nutrient availability, and temperature, not yet tested in the current model, could further enhance the model's adaptability to diverse conditions. As known, these factors influence the morphology and growth conditions of microalgae, causing microscopic images to be heterogeneous even within the same species.

4. Conclusions

The authors presented a method to generate a synthetic dataset of 5 *genera* of environmental microalgae generally found in biofuel production facilities and explored a feasible way to classify microalgae using a set of characteristics typical of cellular structures to train a machine learning (ML) model. At this early stage, the dataset includes the diversity of microalgae, covering properties such as cell size, growth rate, and chlorophyll content. The results showed that these microalgae species could be identified with an accuracy greater than 90% using a machine learning algorithm SVM developed by the authors. It is important to highlight that, while real data is crucial for the initial training of algorithms, synthetic data plays a valuable role. The generation of synthetic data allows the user to expand and diversify the training set, thus helping to improve the generalization ability of learning-based algorithms. Judicious integration of synthetic data can therefore significantly enrich the robustness and flexibility of algorithms, promoting greater resilience in limited or difficult-to-access data scenarios. Overall, the proposed approach opens new avenues to advance research and applications in biofuel production, environmental monitoring, and related fields. More comprehensive methods, abundant data, and a richer set of descriptive futures would lead to better classification performance.

Acknowledgments

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 1561 of 11.10.2022 of Ministero dell'Università e della Ricerca (MUR); funded by the

European Union – NextGenerationEU: Award Number: Project code PE0000021, Concession Decree No. 1561 of 11.10.2022 adopted by Ministero dell'Università e della Ricerca (MUR), CUP - I53C22001450006, Project title "Network 4 Energy Sustainable Transition – NEST".

References

Barsanti L., Birindelli L., Gualtieri P., 2021, Water monitoring by means of digital microscopy identification and classification of microalgae. *Environmental Science: Processes & Impacts*, *23*(10), 1443-1457.

Berenguel P. O., Guzmán J. L., Acién G., Berenguel M., Reul A., 2023, An artificial intelligence approach for identification of microalgae cultures.

Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A.A., 2020, Albumentations: Fast and Flexible Image Augmentations*. Information*, 11, 125.

Chiu H. F., Liao J. Y., Lu Y. Y., Han Y. C., Shen Y. C., Venkatakrishnan K., Golovinskaia O., Wang C. K., 2017, Anti-proliferative, anti-inflammatory and pro-apoptotic effects of Dunaliella salina on human KB oral carcinoma cells, *Journal of Food Biochemistry*, *41*(3), e12349.

Chong J.W.R., Khoo K.S., Chew K.W., Vo D.V.N., Balakrishnan D., Banat F., Munawaroh H.S.H., Iwamoto K., Show P.L., 2023, Microalgae identification: Future of image processing and digital algorithm. *Bioresour. Technol*. 369, 128418

Coşgun A., Günay M. E., Yildirim R., 2023, Machine learning for algal biofuels: A critical review and perspective for future. *Green Chemistry*.

Di Caprio F., Proietti Tocca G., Stoller M., Pagnanelli F., Altimari P., 2022, Control of bacterial contamination in microalgae cultures integrated with wastewater treatment by applying feast and famine conditions, *Journal of Environmental Chemical Engineering*, Volume 10, Issue 5,

Dos Santos R. R., Kunigami C. N., Aranda D. A. G., Teixeira C. M. L. L., 2016, Assessment of triacylglycerol content in Chlorella vulgaris cultivated in a two-stage process. *Biomass and Bioenergy*, *92*, 55-60.

Franco B. M., Navas L. M., Gómez C., Sepúlveda C., Acién, F. G., 2019, Monoalgal and mixed algal cultures discrimination by using an artificial neural network. *Algal Research*, *38*, 101419.

Guiry M.D., Guiry G.M., 2024. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. https://www.algaebase.org; searched on January 10, 2024.

Mikołajczyk A., Grochowski M., 2028, Data augmentation for improving deep learning in image classification problem," 2018 International Interdisciplinary PhD Workshop (IIPhDW), Świnoujście, Poland, pp. 117-122.

Ning H., Li R., Zhou T., 2022, Machine learning for microalgae detection and utilization. *Frontiers in Marine Science*, *9*, 947394.

Oruganti, R. K., Biji, A. P., Lanuyanger, T., Show, P. L., Sriariyanun, M., Upadhyayula, V. K., ... & Bhattacharyya, D. (2023). Artificial intelligence and machine learning tools for high-performance microalgal wastewater treatment and algal biorefinery: A critical review. *Science of The Total Environment*, *876*, 162797.

Otálora P., Guzmán J. L., Acién F. G., Berenguel M., Reul A., 2023, An artificial intelligence approach for identification of microalgae cultures. *New Biotechnology*, *77*, 58-67.

Otálora P., Guzmán J. L., Acién F. G., Berenguel M., Reul A., 2021, Microalgae classification based on machine learning techniques. *Algal Research*, *55*, 102256.

Reimann R., Zeng B., Jakopec M., Burdukiewicz M., Petrick I., Schierack P., Rödiger S., 2020, Classification of dead and living microalgae Chlorella vulgaris by bioimage informatics and machine learning. *Algal research*, 48, 101908.

Sayed E. T., Rezk H., Abdelkareem M. A., Olabi, A. G., 2024, Artificial neural network based modelling and optimization of microalgae microbial fuel cell. *International Journal of Hydrogen Energy*, *52*, 1015-1025.

Singh, V., Verma, M., Chivate, M. S., & Mishra, V. (2023). Machine learning-based optimisation of microalgae biomass production by using wastewater. *Journal of Environmental Chemical Engineering*, *11*(6), 111387.

Sultana N., Hossain S. Z., Abusaad M., Alanbar N., Senan Y., Razzak S. A., 2022, Prediction of biodiesel production from microalgal oil using Bayesian optimization algorithm-based machine learning approaches. *Fuel,* 309, 122184.

Wang Z., Peng X., Xia A., Shah A. A., Huang Y., Zhu X., Liao Q, 2022, The role of machine learning to boost the bioenergy and biofuels conversion. *Bioresource Technology*, *343*, 126099.

Wu W., Huang C. M., Tsai Y. H., 2023, Design and validation of a microalgae biorefinery using machine learning-assisted modeling of hydrothermal liquefaction. *Algal Research*, 74, 103230.

Xu L., Weathers P. J., Xiong X. R., Liu C. Z., 2009, Microalgal bioreactors: challenges and opportunities. *Engineering in life sciences*, *9*(3), 178-189.