

## Article

# Machine Learning Methods Modeling Carbohydrate-Enriched Cyanobacteria Biomass Production in Wastewater Treatment Systems

Héctor Rodríguez-Rangel <sup>1,†</sup>, Dulce María Arias <sup>2,†</sup>, Luis Alberto Morales-Rosales <sup>3,\*,†</sup>,  
Victor Gonzalez-Huitron <sup>1,†</sup>, Mario Valenzuela Partida <sup>1,†</sup> and Joan García <sup>4,†</sup>

<sup>1</sup> Tecnológico Nacional de México, Instituto Tecnológico de Culiacán, Juan de Dios Bátiz 310 Pte. Col. Guadalupe, Culiacán C.P. 80014, Sinaloa, Mexico; hector.rr@culiacan.tecnm.mx (H.R.-R.); victor.gh@culiacan.tecnm.mx (V.G.-H.); mariovalpa@gmail.com (M.V.P.)

<sup>2</sup> Instituto de Energías Renovables, Universidad Nacional Autónoma de México (IER-UNAM), Priv. Xochicalco s/n, Col. Centro, Temixco C.P. 62580, Morelos, Mexico; dmarli@ier.unam.mx

<sup>3</sup> Faculty of Civil Engineering, Conacyt-Universidad Michoacana de San Nicolás de Hidalgo, Morelia C.P. 58060, Michoacán, Mexico

<sup>4</sup> GEMMA—Environmental Engineering and Microbiology Research Group, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya-BarcelonaTech, c/ Jordi Girona 1-3, Building D1, E-08034 Barcelona, Spain; joan.garcia@upc.edu

\* Correspondence: lamorales@conacyt.mx

† These authors contributed equally to this work.



**Citation:** Rodríguez-Rangel, H.; Arias, D.M.; Morales-Rosales, L.A.; Gonzalez-Huitron, V.; Valenzuela Partida, M.; García, J. Machine Learning Methods Modeling Carbohydrate-Enriched Cyanobacteria Biomass Production in Wastewater Treatment Systems. *Energies* **2022**, *15*, 2500. <https://doi.org/10.3390/en15072500>

Academic Editors: Štěpán Papáček, Francisco Gabriel Acién Fernández and José M. Fernández-Sevilla

Received: 26 February 2022

Accepted: 24 March 2022

Published: 29 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** One-stage production of carbohydrate-enriched microalgae biomass in wastewater is a promising option to obtain biofuels. Understanding the interaction of water quality parameters such as nutrients, carbon, internal carbohydrates, and microbial composition in the culture is crucial for efficient operation and viable large-scale cultivation. Bioprocess models are an essential tool for studying the simultaneous effect of complex factors on carbohydrate accumulation, optimizing the process, and reducing operational costs. In this sense, we use a dataset obtained from an empirical model that analyzed the accumulation of carbohydrates in a single process (simultaneous growth and accumulation) from real wastewater. In this experiment, there were no ideal conditions (limiting nutrient conditions), but rather these limitations are guaranteed by the operating conditions (hydraulic retention times/nutrient or carbon loads). Thus, the model integrates 18 variables that are affected and not only carbohydrates. The effect of these variables directly influences the accumulation of carbohydrates. Therefore, this paper analyzes artificial intelligence (AI) algorithms to develop a model to forecast biomass production in wastewater treatment systems. Carbohydrates were modeled using five artificial intelligence methods: (1) Artificial Neural Networks (ANNs), (2) Convolutional Neural Networks (CNN), (3) Long Short-Term Memory Network (LSTMs), (4) K-Nearest Neighbors (kNN), and (5) Random Forest (RF)). The AI methods allow learning how several components interact and if their combinations work faster than building the physical experiments over the same period of time. After comparing the five learning models, the CNN-1D model obtained the best results with an MSE (Mean Squared Error) = 0.0028. This result shows that the model adequately approximates the system's dynamics.

**Keywords:** carbohydrate accumulation modeling; deep learning algorithms; microalgae; resource recovery

## 1. Introduction

The increase in global population and rapid industrialization has stimulated high demand and high energy consumption from fossil-fuels [1–3]. The overexploitation of these petroleum-based sources has exhausted that resource, resulting in an imminent energy crisis. Moreover, these fossil fuels' adverse environmental and health effects due to the

associated noxious greenhouse gases have attracted global attention to seek alternative energy sources [4,5]. Hence, the search for climate-neutral fuel technology as part of an energy security strategy in different parts of the world has become a crucial researchable area. To date, biomass from renewable sources (i.e., lignocellulosic waste or microalgae) has been investigated as possible materials for biofuel production.

Cyanobacteria are prokaryotic photosynthetic microalgae that convert and store solar energy and inorganic compounds into chemical energy [6]. In the last years, microalgae cultivation has become an important source to produce a variety of compounds [7,8], and there is much interest in the production and harvest of algal biomass for its conversion potential in different types of biofuels, namely biogas, biodiesel, bioethanol, biobutanol or biohydrogen [9,10]. From these alternatives, the last three are becoming an attractive alternative of biofuels, and carbohydrate is the only substrate required for its production.

In a recent study, a promising alternative to produce carbohydrate-rich biomass in a one-stage process was presented [11]. A mixed culture of N-fixing soil cyanobacteria was cultivated in a semi-continuous reactor fed with domestic wastewater, operated at different nutrients loads and different initial internal carbohydrates. The results showed an increasing accumulation of carbohydrates, (reaching up to 48%) when the culture was submitted to low nutrients loads, while the cultures submitted to high loads showed a decreasing pattern in carbohydrate content. However, it was unclear if other parameters such as influent and effluent nutrients and carbon concentrations, initial internal carbohydrates, and microbial population changes played a significant role in carbohydrate patterns and final accumulation. Hence, understanding the interaction of those parameters is required for viable largescale cultivation.

Robust bioprocess models capable of predicting carbohydrate accumulation considering complex factors and variables influencing a microalgae-based wastewater treatment can significantly aid in establishing optimal cultivation strategies. Artificial intelligence (AI) algorithms have shown excellent performance for tasks such as forecasting and classification [12–17]. AI algorithms can model different behaviors of complex and chaotic set of data. They present rapid development, easy scalability, and repeatability to changes arising in variables such as simplicity and plasticity [18]. These are interesting advantages in comparison to traditional mathematical models based mostly on Monod [19,20] and Droop [6,21,22] formulations, which normally require intense calibration and parameter tuning to be validated in applications.

In the last few years, increasing interest has been paid to AI models to predict interactions in microalgae cultivation systems [23]. Few studies on the field have shown that AI algorithms such as Artificial Neural Network (ANN) and Convolutional Neural Networks (CNN) can be used for the prediction of genome interactions [24,25], microalgae flocculation [26] and biomass quantification [27], providing improvement in the processes by reducing the number of experiments and condition optimization [28]. In this context, AI models can be an alternative to predict complex interactions between wastewater treatment and microalgae growth and the accumulation of internal metabolites such as carbohydrates.

Computational models that can predict biomass growth and carbohydrate accumulation under different growing conditions will help optimize process performance, operating conditions, and scale-up of cultivation systems for commercialization suitability. Some mathematical models based on Droop and Monod models have been applied to forecast pure microalgal cultures growth or lipids production with one variable (carbon source, N or P, and light). Some recent works have also considered multiple variables; for instance, the works of Andreotti et al. [29] and Solimeno et al. [30] built and validated a mathematical model considering TIC, N, light, and temperature on microalgal growth in open and closed reactors. Moreover, Figueroa-Torres et al. [31] developed another multi-parameter kinetic model describing *Chlamydomonas reinhardtii* growth as well as carbohydrate (starch) and lipid accumulation under controlled mixotrophic conditions. However, these models do not consider biotic factors, such as the competition between cyanobacteria and microalgal species and the interaction of other microorganisms (diatoms and grazers rotifers, amoebas,

ciliates, and flagellates) that could affect carbohydrate content. Moreover, variable conditions in wastewater and mixed liquor could make it difficult for those models to capture the adaptation of cyanobacteria to such changing environments.

Therefore, this work aims to evaluate AI models based on deep learning (ANN, CNN-1D, and Long Short-Term Memory Network (LSTMs)) and machine learning algorithms (K-Nearest Neighbors (kNN) and Random Forest (RF)) for the prediction of carbohydrate-enriched biomass production in semi-continuous reactors used for municipal wastewater treatment. To the authors' knowledge, this is the first time that these types of methods are used to forecast the interaction of input and output microalgae-based wastewater parameters as nutrients (N, P), carbon, and biomass population with total carbohydrate production in microalgae-based wastewater treatment systems. Furthermore, the main contributions of this work can be described as follows:

- This work uses a dataset with 18 variables affecting carbohydrate accumulation in a mixed cyanobacteria consortium treating real wastewater. This work started with an incomplete dataset for three main reasons: (1) the costs involved in water quality analysis; (2) the fact that an automatic device does not monitor the samples, so an analysis has to be performed on a daily or weekly basis; and (3) water quality analysis generates waste in most of the cases, which generates the need to perform a reconstruction of information.
- A trade-off analysis of five learning models of artificial intelligence was carried out using the database of cyanobacterial biomass production in wastewater treatment systems. The five learning models considered were as follows: (1) Artificial Neural Networks (ANNs), (2) Convolutional Neural Networks (CNN), (3) Long Short-Term Memory Network (LSTMs), (4) K-Nearest Neighbors (kNN), and (5) Random Forest (RF). The AI methods allow learning how the 18 variables of the database interact in order to predict biomass production.
- The best result to predict the biomass production was the CNN-1D model with an MSE (Mean Squared Error) = 0.0028. If a model has an MSE close to zero, it can be concluded that the learning model adequately follows the system's dynamics.
- By using the forecast model (CNN-1D), several simulations can be generated to evaluate the conditions of experiments that predict the accumulation of carbohydrates, which could reduce sampling analysis, reagents cost, human work, and waste generation.

## 2. Materials and Methods

### 2.1. Experimental Data

This subsection will present a brief description of the data obtention, providing a general description of experimental procedures.

#### Microalgae Inoculum and Culture

Experimental data were obtained from the study of Arias et al. [11]; this study aimed to evaluate the one-stage operation to produce carbohydrate-enriched cyanobacterial biomass while treating domestic wastewater. The experiments consisted of four closed photoreactors operated at a semi-continuous regime hypothesizing that nutrient and carbon loads controlled by hydraulic retention times could increase biomass carbohydrate production in a medium-term study. Briefly, dry soil crusts were used as inoculum and cultivated in four lab-scale photobioreactors (PBRs) (3L). PBRs were operated in a semi-continuous mode using municipal wastewater as feeding following the operation shown in Table 1. All PBRs were continuously maintained in alternate light/dark phases of 12 h, while temperature and pH were continuously controlled at 27 ( $\pm 2$ ) °C and 7.5, respectively. Mixed liquor removal and feeding were performed each day at the end of the dark phase by the automatic addition/withdrawal accomplished by peristaltic pumps.

**Table 1.** Operational characteristics of four photobioreactors treating municipal wastewater under semi-continuous regime used in the experimental setup.

Experiment	HRT/SRT (d)	Dilution Rate Influent <sup>a</sup>	Volume Removed (L) <sup>b</sup>	Volume Added (L) <sup>c</sup>
Set 1	A1	10	1:1	0.25
	A2	10	2:1	0.25
Set 2	A3	8	1	0.31
	A4	6	1	0.42

<sup>a</sup> Dilution rate of wastewater with distilled water. <sup>b</sup> Volume of mixed liquor removed. <sup>c</sup> Volume of wastewater added to the photobioreactor.

Water quality and biomass production parameters were measured to determine the PBR performance for a period of 24–31 days. Total organic carbon (TOC), total inorganic carbon (TIC), and total nitrogen (TN); and total phosphorus (TP) and total inorganic phosphorus (TIP) were analyzed in triplicate from the influent and effluent (equivalent to the mixed liquor of the culture) twice a week at the end of the dark phase. Biomass was measured as total suspended solids (TSS) and volatile suspended solids (VSS) in the mixed liquor three days per week. These parameters were analyzed using the procedures described in the Standard Methods [32]. Quantitative analyses of microalgae and cyanobacteria were performed by microscopic area cell counting ( $\text{cell}\cdot\text{mL}^{-1}$ ) three times a week according to the methodology of [33], and total carbohydrate (intracellular and exopolysaccharides) contents were measured twice per week in all PBRs at the end of the dark phase [34]. The experiments were performed during approximately 30 days of operation, obtaining the data shown in Table 2.

**Table 2.** Averages and standard deviations of water quality and biomass parameters in the influent, effluent, and mixed liquor used as model inputs.

Parameter	Units	Experiment				
		A1	A2	A3	A4	
		Average Value				
Mixed liquor	Biomass production	g/L·d	0.04 ± 0.01	0.04 ± 0.01	0.06 ± 0.01	0.05 ± 0.01
	Carbohydrates	%	15.51 ± 11.8	23.78 ± 13.82	18.20 ± 5.08	16.03 ± 6.07
	Cyanobacteria population	Log cell/mL	11.9 ± 0.01	11.91 ± 0.02	11.92 ± 0.02	11.95 ± 0.04
	Diatom population	Log cell/mL	9.14 ± 0.39	9.24 ± 0.28	9.45 ± 0.22	9.66 ± 0.28
	Green algae population	Log cell/mL	8.50 ± 0.37	8.43 ± 0.42	8.48 ± 0.52	8.76 ± 0.46
	Protozoa population	Log cell/mL	6.93 ± 3.06	7.84 ± 0.38	7.99 ± 0.19	8.15 ± 0.16
Influent	TIC	mg/L	71.46 ± 12.25	71.46 ± 12.25	84.25 ± 6.00	84.25 ± 6.00
	TOC	mg/L	161.82 ± 52.5	161.83 ± 52.50	117.56 ± 25.46	117.56 ± 25.46
	TIN	mg/L	67.5 ± 12.83	67.5 ± 13.84	65.12 ± 6.87	65.12 ± 6.87
	TON	mg/L	35.18 ± 17.32	35.19 ± 17.33	19.69 ± 8.91	19.69 ± 8.91
	TIP	mg/L	8.59 ± 3.06	8.59 ± 3.07	5.22 ± 1.41	5.22 ± 1.41
	TOP	mg/L	5.68 ± 3.08	5.69 ± 3.09	2.50 ± 1.24	2.5 ± 1.24
Effluent	TIC	mg/L	4.48 ± 1.48	3.58 ± 3.16	0.93 ± 3.01	2.94 ± 3.78
	TOC	mg/L	22.52 ± 10.32	24.06 ± 16.39	0.54 ± 1.39	2.53 ± 3.47
	TIN	mg/L	2.10 ± 3.41	1.16 ± 1.66	21.92 ± 11.42	16.17 ± 10.44
	TON	mg/L	3.40 ± 1.69	0	0.05 ± 0.15	2.33 ± 3.6
	TIP	mg/L	8.59 ± 3.06	0.37 ± 0.83	3.72 ± 1.61	3.73 ± 1.78
	TOP	mg/L	0.73 ± 0.82	2.41 ± 0.95	0.98 ± 1.88	0.79 ± 1.23

## 2.2. Machine Learning Approach Description

This section describes the data preparation and the configuration of the machine learning models used to carry out the task of predicting the carbohydrate's percentage. The code and the experimental data can be found in <https://github.com/mariovp/ml-microalgae> (accessed on 26 February 2022) [35].

### 2.2.1. Dataset Preparation

Dataset preparation comprises two main stages: (1) database reconstruction and (2) database building, which are described as follows:

**Dataset Reconstruction.** Not all variables were recorded with the same periodicity (i.e., the variable carbohydrates was only measured twice per week during the 30-day experiment) during carbohydrate-enriched biomass cultivation. A device does not automatically monitor these methodological techniques, and they have to be performed by an analyst. This is why not all measurements are taken. Some water quality methods (i.e., COD) even generate hazardous residues, making continuous sampling unsustainable in the long term. Therefore, the database contained some missing values in the dataset. As shown in Table 3, some of the values in the table are *null*. Therefore, it was necessary to perform a reconstruction process of the database.

Data reconstruction techniques were used to prepare the data for working with Machine Learning models. Polynomial interpolation reconstructed the missing values in Carbohydrates, Biomass, Cyanobacteria, Green Algae, Diatom, and Protozoa in the range from the first Carbohydrate registry to the last one because they change sequentially with time.

**Table 3.** A database example with missing values.

Database	
$X_1$	$x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}$
$X_2$	$x_{2,1}, null, x_{2,3}, \dots, x_{2,n}$
$X_3$	$x_{3,1}, x_{3,2}, null, \dots, x_{3,n}$
...	...
$X_{n-2}$	$x_{n-2,1}, x_{n-2,2}, null, \dots, x_{n-2,n}$
$X_{n-1}$	$x_{n-1,1}, null, x_{n-1,3}, \dots, x_{n-1,n}$
$X_n$	$x_{n,1}, x_{n,2}, null, \dots, x_{n,n}$

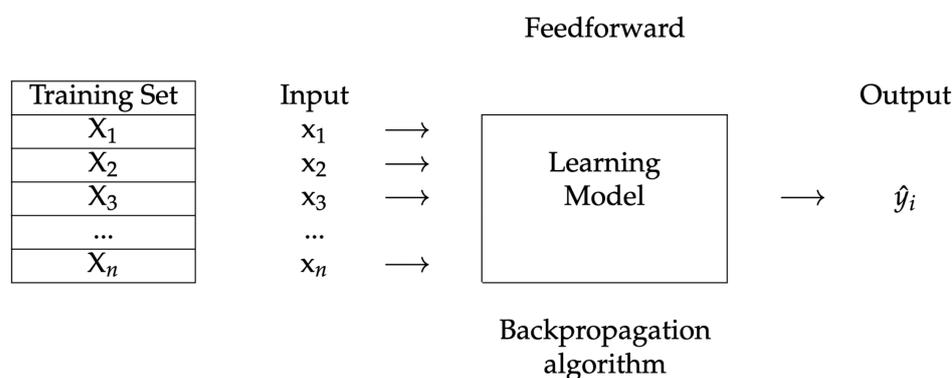
**Database building.** After the data reconstruction process, each database record (18 entries, see parameter in Table 2) was taken as the input data. By modeling by supervised machine learning, dividing the total number of samples into two sets is necessary (training and validation sets). For this experiment, there were 108 samples, where 81 samples (75%) were taken for the training set and 27 samples (25%) for the validation set. Then, both sets were transformed using min–max normalization to place all the values in a range of real numbers between 0 and 1, because the variables were in very different scales and could introduce a bias towards features at larger scales. Table 4 shows a database split into a training set and validation set. Inside the database, each sample has a vector  $X_i$  that corresponds to the 18 entries ( $x_1, x_2, x_3, \dots, x_{17}, x_{18}$ ) used as an input of the learning model, and  $y_i$  corresponds to the desired output used for the learning model’s parameter adjustment.

**Table 4.** Database split into 75% of samples for training and 25% of samples for validation.

	Sample	Input	Output
Training data 75%	$X_1$	$x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}$	$y_1$
	$X_2$	$x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}$	$y_2$
	$X_3$	$x_{3,1}, x_{3,2}, x_{3,3}, \dots, x_{3,n}$	$y_3$
	...	...	...
Validation data 25%	...	...	...
	$X_{n-2}$	$x_{n-2,1}, x_{n-2,2}, x_{n-2,3}, \dots, x_{n-2,n}$	$y_{n-2}$
	$X_{n-1}$	$x_{n-1,1}, x_{n-1,2}, x_{n-1,3}, \dots, x_{n-1,n}$	$y_{n-1}$
	$X_n$	$x_{n,1}, x_{n,2}, x_{n,3}, \dots, x_{n,n}$	$y_n$

### 2.2.2. Machine Learning Setup

Forecasting a variable using machine learning is performed by adjusting the model's parameters from a training set. As shown in Figure 1, the training set contains a set of samples ( $X$ ) that are evaluated at a certain number of epochs. During each epoch (evaluation of the entire training set), the model is scored (*feedforward*) by obtaining its error from the outputs contained in the database ( $e = \hat{y}_i - y_i$ ). Moreover, this error is propagated backward (*backpropagation algorithm*) to improve the model's fitness in the next iteration (epoch).



**Figure 1.** Illustration about the machine learning model's setup. Stars form a training set used for the tuning process.

Once the training process is completed, the model undergoes a validation stage. This is based on the model trained in the previous stage, and all the validation set samples are evaluated. It should be noted that these samples are never provided prior to the model. The error obtained is used to validate whether the model is suitable for implementation or if it is necessary to repeat the process.

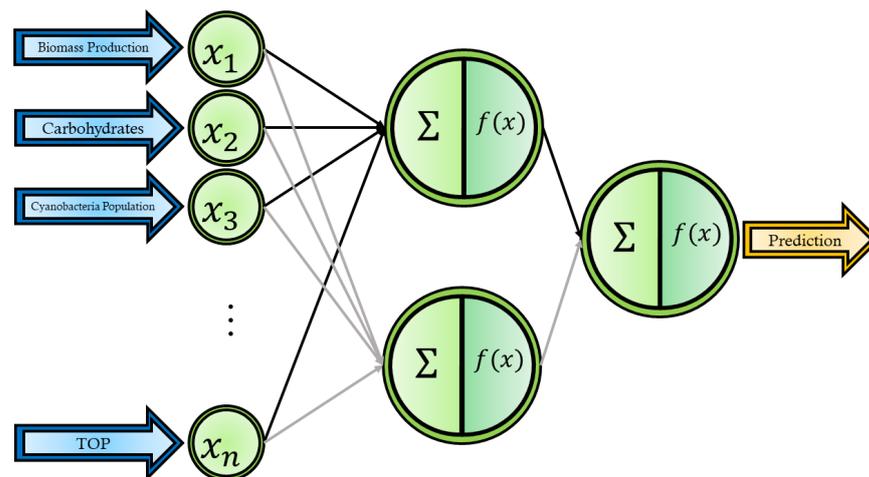
The applications that the generated model can have include the forecast of the modeled variable (biomass production), reconstruction of missing information, and experimentation simulation. For the experiment's simulation, the model allows a virtual recreation of the experiment with the desired conditions, without the need to perform it physically.

### 2.2.3. Machine Learning Model Design

The experimental design consisted of five machine learning models to predict the carbohydrate percentage, taking the other variables as predictors. Model description and their configurations are described in the following sections.

#### Artificial Neural Network

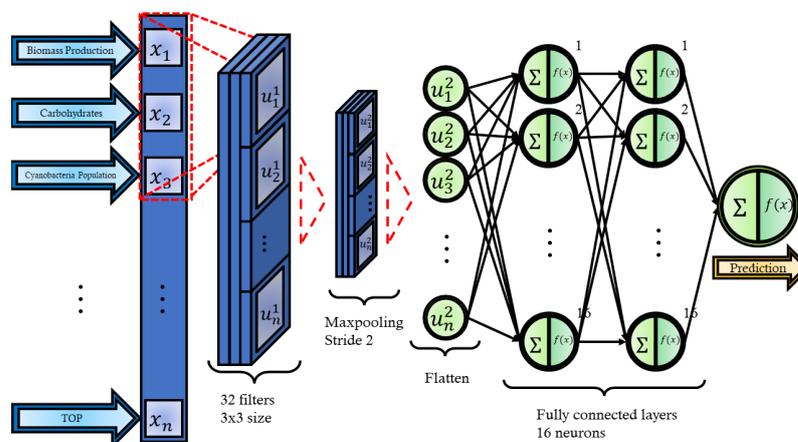
Artificial Neural Networks (ANNs) are computational models based on biological brains. One of the most popular ANNs is the Multilayer Perceptron or MLP, which consists of one input layer, one or more hidden layers, and an output layer [36], as represented in Figure 2. The model experiment began with the Artificial Neural Network, composed of dense layers with ReLu activation function and an output layer with Sigmoid activation. It is a simple architecture by Deep Learning standards. Still, it can deliver great results with the available cyanobacteria data, where a larger model would need more data to be adequately trained.



**Figure 2.** Artificial Neural Network scheme applied for the regression task.

### Convolutional Neural Network

A Convolutional Neural Network (CNN) comprises two parts: a feature extractor and a classifier or regressor [37]. A simple one-dimensional convolutional neural network was implemented due to data size constraints. The feature extractor consists of a 1D convolutional layer and a 1D max-pooling layer. Then, the features are converted from a matrix to an array in the flattened layer in order to be inputted in the MLP regressor. The regressor section has two layers with 16 neurons each and an output layer with one neuron, as represented in Figure 3.



**Figure 3.** The Convolutional Neural Network scheme proposed.

### Long Short-Term Memory Network

Long Short-Term Memory Network (LSTMs) is a type of neural network with the particular ability to remember previous important information. While basic ANNs have neurons, LSTMs have memory cells. Cells have an internal state to store previous information for managing the state; each cell has a set of internal gates through which the input information passes so the cell can forget unimportant information, learn new information, and affect the output [38]. In Figure 4, a brief graphical representation is presented. The complexity of the cells in the LSTM network compared to simple neurons allows it to use fewer results and obtains good results. The LSTM network used in the experiment had two LSTM layers with three and five cells and a ReLu activation function for the cell's output. After those layers, a layer with a neuron is used as output.

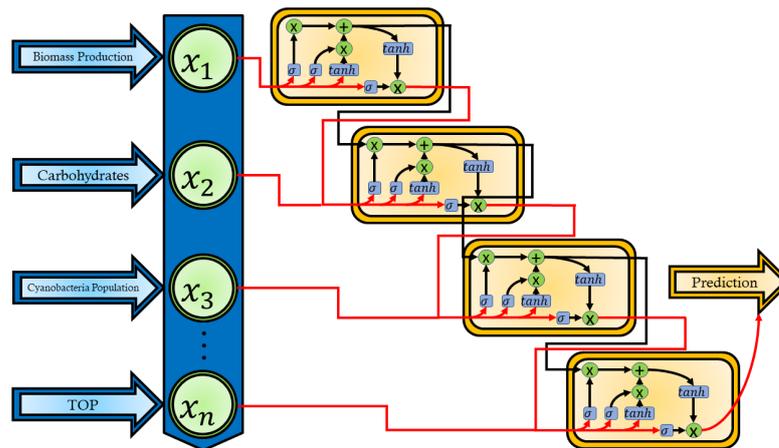


Figure 4. LSTM cell and connections for regression tasks.

### K-Nearest Neighbors

K-Nearest Neighbors (kNN) is a supervised learning algorithm that takes an unlabeled object and compares its features to its nearest neighbors to assign it a class or a numerical value. The algorithm has two important parameters. The first one is  $k$ , which is the number of neighbors to which the object is compared. The second one is the distance metric used to calculate similarity between features [39], as shown in Figure 5. The kNN model was optimized using a grid search with parameter  $k$  in a range of 2–21, the optimum found for this problem was  $k = 6$ .

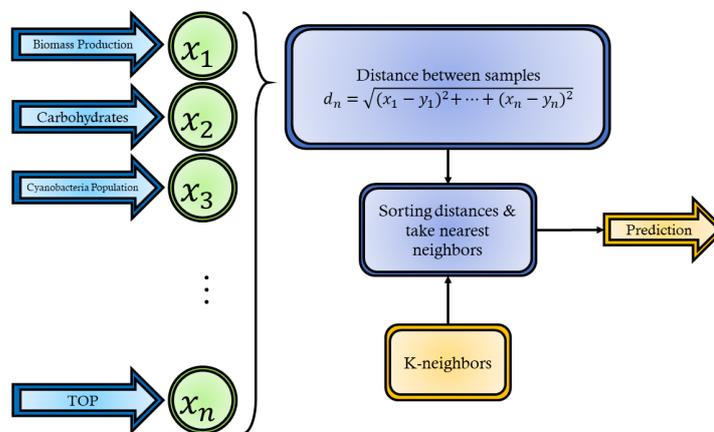
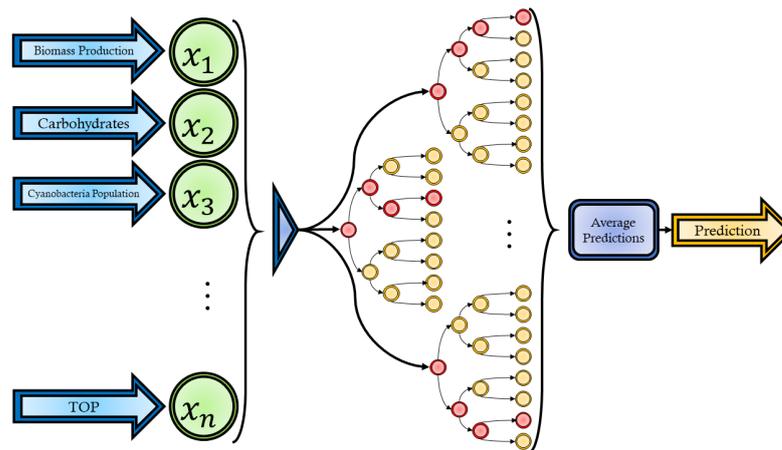


Figure 5. kNN flow representation for regression tasks.

### Random Forest

Random Forest (RF) is an ensemble machine learning method. It consists of many decision trees trained with random data points drawn from the complete data set. To make a classification, all decision trees in the ensemble provide an answer, and the most voted answer is given as output; when performing regression, the output results from averaging all answers. A flowchart in Figure 6 is presented to illustrate this ensemble. The plurality and low correlation of answers between many decision trees provide this method with good results in tabular data [37]. The random forest used in the experiment is composed of 150 estimators or decision trees to make the prediction.



**Figure 6.** Flowchart for the Random Forest method.

### 2.3. Evaluation Metric and Optimization

The model's predictions was based on three metrics: (1) Mean Squared Error (MSE), (2) Root Mean Squared Error (RMSE), and (3) Coefficient of Determination ( $R^2$ ). Each one is described as follows.

The MSE ensures that our trained models have no outlier predictions with huge errors since MSE places larger weight on these errors due to the squaring part of the function (see Equation (1)).

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (1)$$

The RMSE provides a relatively high weight to large errors since the errors are squared before they are averaged (see Equation (2)). RMSE is most useful when large errors are particularly undesirable; hence, we include this metric to detect if the predictions are near the real values. In this metric, the lower the value, the better the model's performance.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2} \quad (2)$$

The  $R^2$  score is known as the coefficient of determination (see Equation (3)).  $R^2$  indicates how good a model fits a given dataset, showing how close the regression line is to the actual data values. The  $R^2$  value lies between 0 and 1, where 0 indicates that this model does not fit the given data, and 1 implies that the model perfectly fits the dataset provided (see Equation (3)).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (3)$$

The sum of squares of residuals  $SS_{res}$ , also called the residual sum of squares, is defines as follows.

$$SS_{res} = \sum_i e_i^2 \quad (4)$$

The total sum of squares  $SS_{tot}$  is proportional to the variance of the data, which are defined as follows:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad (5)$$

where the dataset has  $n$  values marked  $y_1, \dots, y_n$  (collectively known as  $y_i$  or as a vector  $y = [y_1, \dots, y_n]^T$ ), and each value is associated with a fitted (or modeled or predicted) value  $f_1, \dots, f_n$  known as  $f_i$  as a vector  $f$ . The residuals are defined as  $e_i = y_i - f_i$  (forming a vector  $e$ ), and  $\bar{y}$  is the mean of the observed data.

One of the models in the experiment was optimized using a method called Grid Search. It consists of defining a limited set of values for parameters and iterating across all

combinations possible to find the best solution within that search space. It was applied to the K-Nearest Neighbors model because the implementation used had only two parameters with a small range of values.

### 3. Results

#### 3.1. Experiments Configuration

A total of 108 registries were obtained from the sliding window process. The registers were split into two sets, 75% for a training set and 25% for the testing set, resulting in 81 samples for model training and 27 for model validation. Both groups were preprocessed by both groups using a *min-max* normalization to place all values in a range of real numbers between 0 and 1. Normalization was performed because the variables were in very different scales and could introduce a bias towards features in larger scales.

The configurations shown in Table 5 were used for different algorithms used to model the accumulation of carbohydrates. Parameter optimization was performed based on experience. Scikit-Learn library [40], for instance, was used for the algorithms implementation of KNN and RF. Tensor Flow library [41] was used to implement the ANN, CNN, and LSTM algorithms. The dataset and source code of the experimentation are published at <https://github.com/mariovp/ml-microalgae> (accessed on 26 February 2022). The meaning or interpretation of the configuration parameters can be found in reference [40,41].

**Table 5.** Learning model architecture.

Model	Configurations	Hyperparameters
ANN	Input layer = 2, relu (input shape = 18) Hidden layer = 20, relu Hidden layer = 20, relu Output layer = 1, sigmoid Optimizer = adam Loss = MSE Metrics = accuracy	Epochs = 1500 Batch size = 16 Verbose = 0 Shuffle = 1 EarlyStop Patience = 20 Monitor loss
CNN	Conv1D (filters = 32, kernel size = 3, activation = relu) MaxPooling (pool Size = 2) Flatten() Dense(16, activation = relu) Dense (16, activation = relu) Dense (1, activation = sigmoid) Optimizer = adam Loss = mse Metrics = Accuracy	Epochs = 400 Batch size = 16 Verbose = 0 Shuffle = 1 EarlyStop Patience = 20 Monitor loss
KNN	weights = uniform n_neighbors = 5 algorithm = 'auto' leaf_size = 30 metric = 'minkowski' metric_params = None n_jobs = None	
LSTM	LAYERS LSTM (3, activation = relu) LSTM (5, activation relu, return_sequence = true) Timedistributed (dense(1)) Optimizer = adam Loss = MSE Metrics = accuracy	Epochs = 500 Batch size = 16 Verbose = 0 Shuffle = 1 EarlyStop Patience = 20 Monitor loss

Table 5. Cont.

Model	Configurations	Hyperparameters
RF	criterion = 'squared_error' max_depth = None min_samples_split = 2 min_samples_leaf = 1 min_weight_fraction_leaf = 0.0 min_weight_fraction_leaf = 0.0 max_leaf_nodes = None min_impurity_decrease = 0.0 bootstrap = True random_state = None verbose = 0 ccp_alpha = 0.0 max_samples = None	

### 3.2. Experiments Results

Table 6 shows the three loss functions (metrics) to understand the differences between the model's predictions and the ground truth: (1) Mean Squared Error (MSE), (2) Root Mean Squared Error (RMSE), and (3) Coefficient of Determination ( $R^2$ ).

Table 6. Model performance comparison.

Models	MSE	RMSE	$R^2$
ANN	0.0043	0.0655	0.8403
CNN 1D	0.0028	0.0529	0.8966
LSTM	0.0036	0.0600	0.8646
kNN	0.0085	0.0921	0.6831
Random Forest	0.0046	0.0678	0.8286

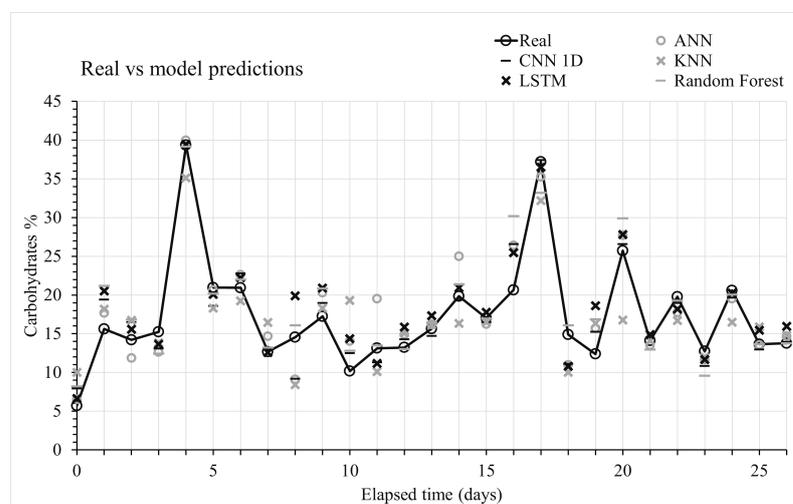
Most machine learning models presented in Table 6 were able to provide an accurate prediction about carbohydrate production. Nevertheless, the CNN 1D model obtained the most accurate predictions of all models with a value of MSE = 0.0028, which is the smallest of the models, as shown in Table 6. The CNN 1D model produced a very close estimation compared to the real value with a few exceptions, see Figure 7. Hence, this model was chosen to predict carbohydrate production. Furthermore, Figure 8 illustrates the errors obtained by each model in a bar graph.

The LSTM model comes second with an MSE of 0.0036; test results are also depicted in Figure 7. The ANN model obtained an MSE of 0.0043; observed in Figure 7, the test results for carbohydrate production were overestimated. This behavior is similar for Random Forest with an MSE = 0.0046 and the kNN with an MSE = 0.0085 (see Figure 7). The model with the worst performance in the experiment was kNN with  $R^2 = 0.6831$ .

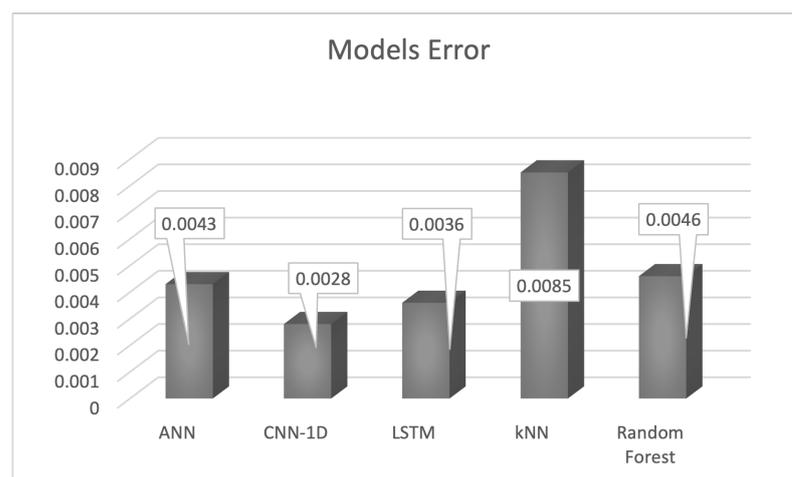
The  $R^2$  score shows a correlation between the forecast model and the interest variable, which in this case involves the carbohydrates. CNN-1D shows a model that is more correlated to the real variable compared to the other three proposed methods according to the  $R^2$  column see (Table 6). Despite being less complex than neural network models (ANN and LSTM), the Random Forest was able to obtain excellent performance with a lot less effort in the model design phase (see Figure 7).

The results demonstrated in Figure 7 exhibit that ML models could learn the growth behavior of carbohydrates. Notably, using the best model CNN D1, with an MSE of 0.0028 and a coefficient of determination of 89%, it is possible to simulate many different scenarios reaching a defined objective without additional experimental work, implying money savings when not spending on inputs or personnel. In addition, it can reduce the contamination caused by physical tests. The process's optimization could be crucial for the cost-effective application of microalgae/cyanobacteria-based wastewater treatment for

carbohydrate production at a large scale. In this manner, biomass can be submitted to a fermentation process that could further convert it into valuable biofuels.



**Figure 7.** Comparison of the five artificial intelligence models vs. the real value of % carbohydrates.



**Figure 8.** Bar graph illustrating the aptitude obtained by each learning model used in this work.

In the last years, the most used methods have been based on kinetic models predicting algal growth or metabolites (lipids or carbohydrates) accumulation [31,42–47]. However, these studies focused on controlled studies with pure cultures, which involve an expensive cost of biomass or metabolites production. We remark that the comparisons presented in Table 7 are not 100% fair since the related works aim at different objectives by using different datasets with varying size. In this sense, Table 7 shows a comparison of several works focused on microalgae growth that has proposed kinetic models or dynamic equations to model growth. Only the study of Supriyanto et al. [48] employed ANN to predict microalgal concentration using mixed cultures but with controlled conditions and artificial growth mediums. Conversely, this study proposed a comparison among five specialized machine learning models to forecast carbohydrate accumulation by considering uncontrolled conditions provided by the wastewater influent and also the variations in the microbial populations of a mixed microalgal consortium. In fact, this work estimated carbohydrate content by considering 18 parameters—this helps to model real conditions experimented within the lab and in real scenarios. In addition, each of these parameters can be modified and observed in terms of how machine learning model prediction affects growth. It should be observed that this model and previous studies focused on laboratory scale works with constant physicochemical parameters such as pH, light, and temperature.

In this context, future work can be directed towards restructuring the model to include variations of such parameters. Hence, more scalable scenarios can be proposed and optimized for large-scale outdoors carbohydrate-rich biomass production.

As is well known, the production of biofuels derived from carbohydrates (e.g., bioethanol) obtained from microalgae (on a large scale) has been limited by the low amount of carbohydrates within the algal biomass that is processed. Most of the previous studies (and models) presented in Table 7 focus on performing two processes to obtain microalgae to accumulate carbohydrates. The first process focuses on the production of algal biomass; once they achieve the highest biomass growth, the biomass is passed onto the next process (carbohydrate accumulation), where the environmental conditions of the crop are modified so that a nutrient (nitrogen or phosphorus) is limited, and stress is achieved in the cell and carbohydrates accumulate. It should be noted that the accumulation of carbohydrates, in most cases, is carried out in batch processes. Even all previous models have been in batch processes in synthetic culture media.

An experimental model considered in this study was used to develop a forecast biomass production in wastewater treatment systems (see the experiment details in Table 2 and the database construction explained in Section 2.2.1). The empirical model is a process that seeks to model the accumulation of carbohydrates in a single process (simultaneous growth and accumulation) from real wastewater. In this process, there are no ideal conditions (limiting nutrient conditions); rather, these limitations are guaranteed by the operating conditions (hydraulic retention times/nutrient or carbon loads). Thus, this model integrates all variables that are affected and not only carbohydrates. The effect of these variables directly influences the accumulation of carbohydrates. Studies such as these, where understanding and predicting these interactions are aimed at, are very important since they are the first step to achieve the simultaneous accumulation of carbohydrates in microalgae in wastewater and, in the future, to achieve the appropriate biomass to undergo the production of biofuels on a larger scale.

Therefore, to reduce the number of high-resolution simulations, machine learning models were implemented with a set of simulations normally performed. The machine learning algorithms can learn how produce the most optimal solutions, which is what the experiment requires. In addition, the evaluation can be fast even when using multiple combinations of input parameters to forecast the expected output. This fact provides a wide breadth of knowledge on how several components interact and if their combinations work faster than building the physical experiments over the same period of time. In this sense, modeling carbohydrate-enriched cyanobacterial biomass production in wastewater treatment systems by using machine learning can result in the identification of the value of factors required to maximize the output of biomass production. This impacts the time and cost associated with several experiments to determine the best combination between the 18 factors.

Artificial intelligence has a study area that focuses on using data and algorithms to imitate the manner humans learn, gradually improving its accuracy; this is named machine learning. These algorithms have been used to solve fault detection, object classification, control, diagnostics, and forecasting problems. This work proposed using five artificial intelligence to determine the best candidate to predict carbohydrate production in wastewater. Normally, for artificial intelligence algorithms, information or samples of the problem must be collected and preprocessed to eliminate noise or apply reconstruction techniques for data reconstruction. For instance, in the problem presented in this paper, the samples generated required more than 90 days, where the costs of materials and human resources are high, introducing delays if one experiment is not well designed or does not offer an expected result. Therefore, if a simulation is carried out, an expected output (forecast) about the future behavior will be obtained by considering the input parameters so that the real experiment can confirm the assumptions obtained by artificial intelligence algorithms. In this sense, the forecast should only be tested in a real scenario with the best results resulting in saving resources and time.

**Table 7.** Comparison of state-of-the art works and our proposal.

Culture	Model Inputs	Model Type	Model Outputs	Ref.
<i>Isochrysis Galbana</i>	<ul style="list-style-type: none"> <li>• Biomass</li> <li>• Lipids</li> <li>• NaNO<sub>3</sub></li> </ul>	Baranyi–Roberts and logistic equation, and Luedeking–Piret model	Lipid production	[42]
<i>Dunaliella viridis</i>	<ul style="list-style-type: none"> <li>• Functional biomass</li> <li>• Carbohydrates</li> <li>• Lipids</li> <li>• Chlorophyll a</li> <li>• Extracelullar nitrogen</li> <li>• Intracelullar nitrogen</li> </ul>	Kinetic model	Lipids, carbohydrates, and biomass	[43]
<i>Chlamydomonas reinhardtii</i>	<ul style="list-style-type: none"> <li>• Nitrogen</li> <li>• Acetate in biomass starch</li> <li>• Lipid formation</li> </ul>	Kinetic model	Biomass, starch, and lipid	[31]
<i>Chlorella vulgaris</i> SAG 211–12	<ul style="list-style-type: none"> <li>• NaCl</li> <li>• Glucose</li> <li>• Glycerol</li> </ul>	ow-order polynomial models	Growth, lipid and Starch	[44]
<i>Chlorella sorokiniana</i> FACHB-275	<ul style="list-style-type: none"> <li>• Glucose</li> <li>• Nitrogen</li> <li>• Phosphorus</li> </ul>	Kinetic model based on Monod and Luedeking–Piret expressions	Biomass, carbohydrate and lipid	[45]
<i>Coelastrella</i> sp. 3–4, <i>Scenedesmus</i> sp. B2-2 and <i>Scenedesmus obliquus</i> RISE (UTEX 417)	<ul style="list-style-type: none"> <li>• Lipid</li> <li>• Biomass</li> <li>• Nitrogen</li> <li>• Carbohydrates</li> </ul>	Kinetic model based on Droop’s mathematical model	Biomass growth	[46]
<i>Pseudochlorococcum</i> sp.	<ul style="list-style-type: none"> <li>• Algal biomass concentration, excluding neutral lipids</li> <li>• Neutral lipid concentration</li> <li>• Chlorophyll a</li> <li>• Extracellular nitrogen concentration</li> </ul>	Kinetic model based on Droop’s mathematical model	Microalgae growth and neutral lipids	[47]

Table 7. Cont.

Culture	Model Inputs	Model Type	Model Outputs	Ref.
Mixed culture	<ul style="list-style-type: none"> <li>• Initial microalgae Concentration (dry basis)</li> <li>• Harvesting period</li> <li>• Hydraulic retention time</li> <li>• Sodium acetate addition</li> <li>• Average solar irradiance</li> <li>• Average water temperature</li> <li>• Average pH</li> <li>• Nitrate concentration</li> </ul>	ANN	Microalgae concentration	[48]
Mixed culture	<ul style="list-style-type: none"> <li>• Mixed liquor</li> <li>• Biomass production</li> <li>• Carbohydrates</li> <li>• Cyanobacteria population</li> <li>• Diatom population</li> <li>• Green algae population</li> <li>• Protozoa population Influent</li> <li>• TIC</li> <li>• TOC</li> <li>• TIN</li> </ul>	ANNs, CNN, LSTMs, KNN, RF	Carbohydrate's content	This study

We remark that the generated model allows performing simulations very close to reality. They are performed instantaneously, and if they are combined with optimization algorithms (i.e., evolutionary computation and Bayesian optimization), better results in real experiments can be confirmed; this could be a research area for future work. Thus, they can help generate a procedure that maximizes the desired variable without high consumptions of time, financial, and/or human resources.

#### 4. Conclusions

This work deals with modeling carbohydrate accumulation in microalgae cultivated in domestic wastewater. Traditionally, related works are carried out by an expert in mathematical modeling, generating differential equations to describe microalgae growth or subproducts accumulation. Hence, it is needed to solve different challenges with time and cost restrictions, such as calibration and validation of complex dynamic models.

This work explored five supervised learning methods (ANN, CNN-1D, LSTM, kNN, and Random Forest) to model carbohydrates accumulation considering the interaction of nutrients, carbon, and biomass growth and population. The models based on deep learning (LSTM, ANN, and CNN-1D) showed a better performance in estimating carbohydrate accumulation present in the cultures. The CNN-1D model showed better performance than the models evaluated. These deep learning models allow us to describe the cultures' behavior when considering a larger number of descriptors.

The best result obtained among the methods explored has a mean square error of 0.002, with a variance of 5%. It produces a certainty of 89% ( $R^2$ ) with respect to correlation when simulating different scenarios without performing the experiments physically. The advantages of using machine learning models include reducing the number of experiments and process optimization and saving the amount of reagents used for analysis and lab work.

As future work, the models' results could be used as an input to optimize other AI methods. For instance, the inputs based on heuristics and meta-heuristics stories such as genetic algorithms, particle swarm, ant colony, and neuro-fuzzy could find a combination of elements that maximize the production of carbohydrates in microalgae.

**Author Contributions:** Conceptualization, H.R.-R., D.M.A. and L.A.M.-R.; data curation, D.M.A.; formal analysis, H.R.-R., D.M.A. and L.A.M.-R.; funding acquisition, D.M.A.; investigation, V.G.-H.; methodology, H.R.-R., D.M.A., L.A.M.-R., M.V.P. and J.G.; software, V.G.-H. and M.V.P.; supervision, J.G.; validation, H.R.-R., L.A.M.-R., V.G.-H. and J.G.; visualization, V.G.-H. and M.V.P.; writing—original draft, H.R.-R., D.M.A. and L.A.M.-R.; writing—review and editing, V.G.-H., M.V.P. and J.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was financially supported by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México (UNAM), México, under the PAPIIT Project (No. IA102821). In addition, this work was supported by the Mexican National Council for Science and Technology (CONACYT) by Research Project 613.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** We published the dataset and source code of the experimentation at <https://github.com/mariovp/ml-microalgae> (accessed on 26 February 2022).

**Conflicts of Interest:** The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors declare no conflict of interest.

#### References

1. Okoye, P.; Abdullah, A.; Hameed, B. A review on recent developments and progress in the kinetics and deactivation of catalytic acetylation of glycerol—A byproduct of biodiesel. *Renew. Sustain. Energy Rev.* **2017**, *74*, 387–401. [[CrossRef](#)]
2. Wang, C.; Luo, D.; Zhang, X.; Huang, R.; Cao, Y.; Liu, G.; Zhang, Y.; Wang, H. Biochar-based slow-release of fertilizers for sustainable agriculture: A mini review. *Environ. Sci. Ecotechnol.* **2022**, *10*, 100167. [[CrossRef](#)]

3. Wang, C.; Wang, H. Carboxyl functionalized Cinnamomum camphora for removal of heavy metals from synthetic wastewater-contribution to sustainability in agroforestry. *J. Clean. Prod.* **2018**, *184*, 921–928. [[CrossRef](#)]
4. Aitken, D.; Bulboa, C.; Godoy-Faundez, A.; Turrion-Gomez, J.L.; Antizar-Ladislao, B. Life cycle assessment of macroalgae cultivation and processing for biofuel production. *J. Clean. Prod.* **2014**, *75*, 45–56. [[CrossRef](#)]
5. Wang, C.; Sun, R.; Huang, R. Highly dispersed iron-doped biochar derived from sawdust for Fenton-like degradation of toxic dyes. *J. Clean. Prod.* **2021**, *297*, 126681. [[CrossRef](#)]
6. Huang, Y.; Luo, L.; Xu, K.; Wang, X.C. Characteristics of external carbon uptake by microalgae growth and associated effects on algal biomass composition. *Bioresour. Technol.* **2019**, *292*, 121887. [[CrossRef](#)] [[PubMed](#)]
7. Arias, D.M.; García, J.; Uggetti, E. Production of polymers by cyanobacteria grown in wastewater: Current status, challenges and future perspectives. *New Biotechnol.* **2020**, *55*, 46–57. [[CrossRef](#)]
8. Rizwan, M.; Mujtaba, G.; Memon, S.A.; Lee, K.; Rashid, N. Exploring the potential of microalgae for new biotechnology applications and beyond: A review. *Renew. Sustain. Energy Rev.* **2018**, *92*, 394–404. [[CrossRef](#)]
9. de Farias Silva, C.E.; Bertuccio, A. Bioethanol from microalgae and cyanobacteria: A review and technological outlook. *Process Biochem.* **2016**, *51*, 1833–1842. [[CrossRef](#)]
10. Vargas, S.R.; dos Santos, P.V.; Zaiat, M.; do Carmo Calijuri, M. Optimization of biomass and hydrogen production by *Anabaena* sp. (UTEX 1448) in nitrogen-deprived cultures. *Biomass Bioenergy* **2018**, *111*, 70–76. [[CrossRef](#)]
11. Arias, D.M.; Uggetti, E.; García, J. Assessing the potential of soil cyanobacteria for simultaneous wastewater treatment and carbohydrate-enriched biomass production. *Algal Res.* **2020**, *51*, 102042. [[CrossRef](#)]
12. Tóth, P.; Garami, A.; Csordás, B. Image-based deep neural network prediction of the heat output of a step-grate biomass boiler. *Appl. Energy* **2017**, *200*, 155–169. [[CrossRef](#)]
13. Rodriguez, H.; Flores, J.J.; Morales, L.A.; Lara, C.; Guerra, A.; Manjarrez, G. Forecasting from incomplete and chaotic wind speed data. *Soft Comput.* **2019**, *23*, 10119–10127. [[CrossRef](#)]
14. Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y.M.; Park, J.p.; Kim, J.H.; Cho, K.H. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *J. Environ. Sci.* **2015**, *32*, 90–101. [[CrossRef](#)] [[PubMed](#)]
15. Rodriguez, H.; Puig, V.; Flores, J.J.; Lopez, R. Flow meter data validation and reconstruction using neural networks: Application to the Barcelona water network. In Proceedings of the 2016 European Control Conference (ECC), Aalborg, Denmark, 29 June–1 July 2016; pp. 1746–1751. [[CrossRef](#)]
16. Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; De Marinis, G. Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. *Water* **2017**, *9*, 105. [[CrossRef](#)]
17. Gonzalez-Huitron, V.; León-Borges, J.A.; Rodriguez-Mata, A.E.; Amabilis-Sosa, L.E.; Ramírez-Pereda, B.; Rodriguez, H. Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. *Comput. Electron. Agric.* **2021**, *181*, 105951. [[CrossRef](#)]
18. Ali, I.; Greifeneder, F.; Stamenkovic, J.; Neumann, M.; Notarnicola, C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sens.* **2015**, *7*, 16398–16421. [[CrossRef](#)]
19. Farza, M.; Rodriguez-Mata, A.; Robles-Magdaleno, J.; M'Saad, M. A new filtered high gain observer design for the estimation of the components concentrations in a photobioreactor in microalgae culture. *IFAC-PapersOnLine* **2019**, *52*, 904–909. [[CrossRef](#)]
20. Zentou, H.; Zainal Abidin, Z.; Yunus, R.; Awang Biak, D.R.; Zouanti, M.; Hassani, A. Modelling of Molasses Fermentation for Bioethanol Production: A Comparative Investigation of Monod and Andrews Models Accuracy Assessment. *Biomolecules* **2019**, *9*, 308. [[CrossRef](#)]
21. Bekirogullari, M.; Pittman, J.K.; Theodoropoulos, C. Multi-factor kinetic modelling of microalgal biomass cultivation for optimised lipid production. *Bioresour. Technol.* **2018**, *269*, 417–425. [[CrossRef](#)]
22. Jerono, P.; Schaum, A.; Meurer, T. Observer Design for the Droop Model with Biased Measurement: Application to *Haematococcus Pluvialis*. In Proceedings of the 2018 IEEE Conference on Decision and Control (CDC), Miami Beach, FL, USA, 17–19 December 2018; pp. 6295–6300. [[CrossRef](#)]
23. Teng, S.Y.; Yew, G.Y.; Sukačová, K.; Show, P.L.; Máša, V.; Chang, J.S. Microalgae with artificial intelligence: A digitalized perspective on genetics, systems and products. *Biotechnol. Adv.* **2020**, *44*, 107631. [[CrossRef](#)] [[PubMed](#)]
24. Dutta, S.; Madan, S.; Parikh, H.; Sundar, D. An ensemble micro neural network approach for elucidating interactions between zinc finger proteins and their target DNA. *BMC Genom.* **2016**, *17*, 1033. [[CrossRef](#)] [[PubMed](#)]
25. Linder, J.; Bogard, N.; Rosenberg, A.B.; Seelig, G. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. *Cell Syst.* **2020**, *11*, 49–62. [[CrossRef](#)] [[PubMed](#)]
26. Zenooz, A.M.; Ashtiani, F.Z.; Ranjbar, R.; Nikbakht, F.; Bolouri, O. Comparison of different artificial neural network architectures in modeling of *Chlorella* sp. flocculation. *Prep. Biochem. Biotechnol.* **2017**, *47*, 570–577. [[CrossRef](#)]
27. Ruiz-Santaquiteria, J.; Bueno, G.; Deniz, O.; Vallez, N.; Cristobal, G. Semantic versus instance segmentation in microscopic algae detection. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103271. [[CrossRef](#)]
28. Treloar, N.J.; Fedorec, A.J.H.; Ingalls, B.; Barnes, C.P. Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLoS Comput. Biol.* **2020**, *16*, e1007783. [[CrossRef](#)]
29. Andreotti, V.; Solimeno, A.; Rossi, S.; Ficara, E.; Marazzi, F.; Mezzanotte, V.; García, J. Bioremediation of aquaculture wastewater with the microalgae *Tetraselmis suecica*: Semi-continuous experiments, simulation and photo-respirometric tests. *Sci. Total Environ.* **2020**, *738*, 139859. [[CrossRef](#)]

30. Solimeno, A.; Samsó, R.; Uggetti, E.; Sialve, B.; Steyer, J.P.; Gabarró, A.; García, J. New mechanistic model to simulate microalgae growth. *Algal Res.* **2015**, *12*, 350–358. [CrossRef]
31. Figueroa-Torres, G.M.; Pittman, J.K.; Theodoropoulos, C. Kinetic modelling of starch and lipid formation during mixotrophic, nutrient-limited microalgal growth. *Bioresour. Technol.* **2017**, *241*, 868–878. [CrossRef]
32. Federation, W.E.; Association, A. *Standard Methods for the Examination of Water and Wastewater*; American Public Health Association (APHA): Washington, DC, USA, 2005; Volume 21.
33. Lin, S. Algal Culturing Techniques. *J. Phycol.* **2005**, *41*, 906–908. [CrossRef]
34. Lanham, A.B.; Ricardo, A.R.; Coma, M.; Fradinho, J.; Carvalheira, M.; Oehmen, A.; Carvalho, G.; Reis, M.A.M. Optimisation of glycogen quantification in mixed microbial cultures. *Bioresour. Technol.* **2012**, *118*, 518–525. [CrossRef] [PubMed]
35. Valenzuela, M. Machine Learning Microalgae. 2021. Available online: <https://cvalenzuelab.com/> (accessed on 26 February 2022).
36. Patterson, J.; Gibson, A. *Deep Learning: A Practitioner's Approach*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
37. Guidici, D.; Clark, M.L. One-Dimensional Convolutional Neural Network Land-Cover Classification of Multi-Seasonal Hyperspectral Imagery in the San Francisco Bay Area, California. *Remote Sens.* **2017**, *9*, 629. [CrossRef]
38. Xiaoyun, Q.; Xiaoning, K.; Chao, Z.; Shuai, J.; Xiuda, M. Short-term prediction of wind power based on deep Long Short-Term Memory. In Proceedings of the 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Xi'an, China, 25–28 October 2016; pp. 1148–1152. [CrossRef]
39. Al-Qahtani, F.H.; Crone, S.F. Multivariate k-nearest neighbour regression for time series data—A novel algorithm for forecasting UK electricity demand. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–8. [CrossRef]
40. Kramer, O. *Scikit-Learn*; Springer: New York, NY, USA, 2016; pp. 45–53.
41. Dillon, J.V.; Langmore, I.; Tran, D.; Brevdo, E.; Vasudevan, S.; Moore, D.; Patton, B.; Alemi, A.; Hoffman, M.; Saurous, R.A. Tensorflow distributions. *arXiv* **2017**, arXiv:1711.10604.
42. He, Y.; Chen, L.; Zhou, Y.; Chen, H.; Zhou, X.; Cai, F.; Huang, J.; Wang, M.; Chen, B.; Guo, Z. Analysis and model delineation of marine microalgae growth and lipid accumulation in flat-plate photobioreactor. *Biochem. Eng. J.* **2016**, *111*, 108–116. [CrossRef]
43. Wang, D.; Lai, Y.C.; Karam, A.L.; de los Reyes, F.L.; Ducoste, J.J. Dynamic Modeling of Microalgae Growth and Lipid Production under Transient Light and Nitrogen Conditions. *Environ. Sci. Technol.* **2019**, *53*, 11560–11568. [CrossRef]
44. Kaplan, E.; Sayar, N.A.; Kazan, D.; Sayar, A.A. Assessment of different carbon and salinity level on growth kinetics, lipid, and starch composition of *Chlorella vulgaris* SAG 211-12. *Int. J. Green Energy* **2020**, *17*, 290–300. [CrossRef]
45. Murwanashyaka, T.; Shen, L.; Yang, Z.; Chang, J.S.; Manirafasha, E.; Ndikubwimana, T.; Chen, C.; Lu, Y. Kinetic modelling of heterotrophic microalgae culture in wastewater: Storage molecule generation and pollutants mitigation. *Biochem. Eng. J.* **2020**, *157*, 107523. [CrossRef]
46. Gojkovic, Z.; Lu, Y.; Ferro, L.; Toffolo, A.; Funk, C. Modeling biomass production during progressive nitrogen starvation by North Swedish green microalgae. *Algal Res.* **2020**, *47*, 101835. [CrossRef]
47. Packer, A.; Li, Y.; Andersen, T.; Hu, Q.; Kuang, Y.; Sommerfeld, M. Growth and neutral lipid synthesis in green microalgae: A mathematical model. *Bioresour. Technol.* **2011**, *102*, 111–117. [CrossRef]
48. Supriyanto.; Noguchi, R.; Ahamed, T.; Rani, D.S.; Sakurai, K.; Nasution, M.A.; Wibawa, D.S.; Demura, M.; Watanabe, M.M. Artificial neural networks model for estimating growth of polyculture microalgae in an open raceway pond. *Biosyst. Eng.* **2019**, *177*, 122–129. [CrossRef]