Early prediction of *Spirulina platensis* biomass yield for biofuel production using machine learning

Phoebe Mae Lim Ching[1*], Andres Philip Mayol[2], Jayne Lois G. San Juan[3,6], Richard H. Y. So[4], Charlle L. Sy[3,6], Emelina Mandia[5], Aristotle T. Ubando[2,6], Alvin B. Culaba[2,6]

[1] Bioengineering Graduate Program, Chemical and Biological Engineering Department, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[2] Mechanical Engineering Department, De La Salle University, 2401 Taft Avenue, 0922 Manila, Philippines
[3] Industrial Engineering Department, De La Salle University, 2401 Taft Avenue, 0922 Manila, Philippines
[4] Industrial Engineering and Decision Analytics Department, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[5] Biology Department, De La Salle University, 2401 Taft Avenue, 0922 Manila, Philippines
[6] Center for Engineering and Sustainable Development Research, De La Salle University, 2401 Taft Avenue, 0922 Manila, Philippines
*corresponding author: pmlching@connect.ust.hk

Abstract

Despite the many advantages of third-generation biofuels, there are still numerous opportunities to improve their production efficiency and streamline their commercialization. The unpredictability of cultivating biomass is a major challenge to consistent, efficient production. In particular, the cultivation of *Spirulina platensis* biomass for biofuel production is affected by various environmental factors such as light, temperature, pH and the nutrient concentration of water. Since controlling these factors is energy intensive, a biomass prediction model would be helpful in anticipating biomass production and in indicating necessary adjustments to the process to improve yield. In this case, earlier is clearly better. This study developed a machine learning-based early prediction model which identifies the earliest time during cultivation that the process parameters optical density and pH can accurately be used to predict biomass yield. In the case study, the early prediction model predicted the final biomass yield (on the 23rd day) by the 8th day of cultivation using ridge regression. Furthermore, an application of this model in pH control led to a 54.1% average improvement in biomass yield. This model may be used to monitor cultivation batches allowing problems (i.e., low yield) to be identified early. It can also be applied in process simulation and optimization to improve biomass yield. In summary, mathematical modelling can make the unpredictable biomass process more predictable, and improve production efficiency.

Keywords: Biofuels; biomass; machine learning

1. Introduction

The increasing consumption of fossil fuels due to population growth and technological advancements have impacted the climate and depleted natural reserves. In light of these alarming issues, the United Nations crafted the sustainable development goals (SDGs). Providing "affordable and clean energy" to everyone is one of the SDGs (Elavarasan et al., 2021). Among the many developments in renewable and alternative energy sources (e.g., solar, hydro, thermal, geothermal, wind, and biomass energy), biofuels have attracted interest because they have a low-carbon footprint and are appropriate for vehicular use. Different feedstocks have been investigated for biofuel utilization such as corn, soybean, oil palm, and crop waste (Chisti, 2007). However, competition with edible plants for arable land is a serious concern. Hence, there is a need to find an alternative feedstock that provides high yield and does not compete with food sources. Biomass has gained interest as a promising source of natural compounds that can be used in a wide array of products and for various applications (Solis et al., 2020). Valuable compounds derived from biomass can be converted into biofuels such as bio-hydrogen, biogas, bioethanol, and biodiesel (Liyanaarachchi et al., 2021). In addition, third-generation biofuels have additional benefits such as carbon sequestration and water treatment capabilities. Compared to other feedstocks such as canola, corn, palm oil, and soybean, the biomass for some third-generation biofuels require much less land area to cultivate, and have high biomass productivity (San Juan et al., 2020). One example is microalgal biomass, for which the production of oil per unit of land is greater than from conventional sources of feedstock (Chisti, 2007). Aside from minimal land requirements, this type of biomass may be cultivated on wastewater and sludge, which can provide critical nutrients that can improve productivity (Caligan et al., 2020). Furthermore, third-generation biofuels have strong potential for scalability and standardization. This is important as up-scaling biofuel production for commercial use is challenging and requires social acceptance (Culaba et al., 2019a). At present, the technology for third-generation biofuels is reaching maturity. High-yield strains of bacteria and microalgae have been identified; and the technology for transesterification has improved significantly (Schade & Meier, 2021). However, these biofuels have

yet to achieve commercial-scale, owing to the lack of standardization for several parts of the biofuel supply chain.

With regards to standardization, the process of cultivating biomass for biofuel production is a major area of improvement. Biomass cultivation can be highly resource-intensive, depending on whether a closed or open system is utilized (Brindhadevi et al., 2021). Closed system cultivation through photo-bioreactors requires significant capital and energy. Open-raceway pond cultivation is inefficient due to highly-variable external factors which determine biomass growth rate and yield (Culaba et al., 2019b). These factors include sunlight, carbon dioxide, temperature, pH, and substrate composition among many others, which are typically left uncontrolled in an open system (Murwanashyaka et al., 2020). Hence, there is a tradeoff between predictability and cost.

Machine learning can be used to predict biomass yield based on relevant process variables (Mowbray et al., 2021). Some notable examples are discussed as follows. Nayak et al. (2018) constructed a non-linear artificial neural network model combined with a genetic algorithm to predict the optimal cultivations settings which enhanced *Scenedesmus* sp. yield. Azari et al. (2020) used a predictive modelling approach to identify the optimal temperature and light intensity for the cultivation of *Chlorella vulgaris*. Banerjee et al. (2020) utilized response surface methodology to identify the optimal pH, temperature, and initial concentrations of acetate and ammonium chloride for the growth of *Chlamydomonas reinhardtii*. Barbosa et al. (2020) applied regression methods to develop a sensor for biomass concentration based on optical density (OD) that aids in the cultivation process. Žitnik et al. (2019) employed decision trees to understand dependencies and interactions in the removal efficiency of *E. coli* from raw blackwater and treated through the cultivation of microalgae *Chlorella vulgaris*. Behera et al. (2019) estimated microalgal productivity and carbon dioxide sequestration potential as influenced by a combination of key climatic variables through analytical modelling.

These studies have identified the factors with the highest contribution to biomass yield, as insights for process control. However, it is notable that the cultivation cycle takes several days and involves inherent nutrient-growth dynamics. During this time, optical density (OD) changes in relation to the growing biomass concentration (Barbosa et al., 2020). In addition, there is a tendency

for the pH to rise as CO_2 is consumed by the microalgae (Richmond & Grobbelaar, 1986). Thus, measuring factors based on their initial conditions, or even during cultivation, may only partially characterize the relationships between the relevant factors and biomass yield.

This study applies machine learning to develop an early prediction model that identifies the shortest and earliest time series of OD and pH respectively as predictors for biomass yield. This allows for problems of low yield to be identified and resolved before the end of the cultivation cycle. The model is trained to represent the process dynamics of cultivation, while observing the preference of monitoring systems for early prediction (with initial process data) and minimal sampling. As a result, the model can accurately predict biomass yield, allowing for the early identification of potential problems such as low-yield batches. Prediction of biomass growth and output can eliminate the costly experimentation that would otherwise be needed to optimize the process. This addresses a major bottleneck in commercialization and scaling up of production systems for energy-producing biomass.

Ridge regression was selected as the modelling approach for developing the early prediction model for two reasons: (1) It is based on linear regression, which complies with the knowledge that there is a linear relationship between OD and biomass concentration (Barbosa et al., 2020); and (2) the error function of ridge regression is penalized with L^2 regularization (eq. 3), which addresses the time-collinearity of its predictors. Using collinear predictors results in a model that is very sensitive to changes in the values of these predictors. As a result, the model is less generalizable and more prone to poor performance when it encounters inputs that are not part of its training set.

The rest of the paper is organized as follows. The cultivation experimental set up and data gathering; the data feature preparation; and the formulation of the early prediction model for biomass cultivation and performance evaluation are detailed in Section 2. Section 3 presents the findings from the prediction model development. Section 3 also demonstrates a potential application of the proposed early prediction model for controlling growth parameters, particularly pH levels. Finally, concluding remarks and recommendations for further research are presented at the end of the paper.

2. Materials and Methods

2.1. Set-up and sampling of biomass cultivation

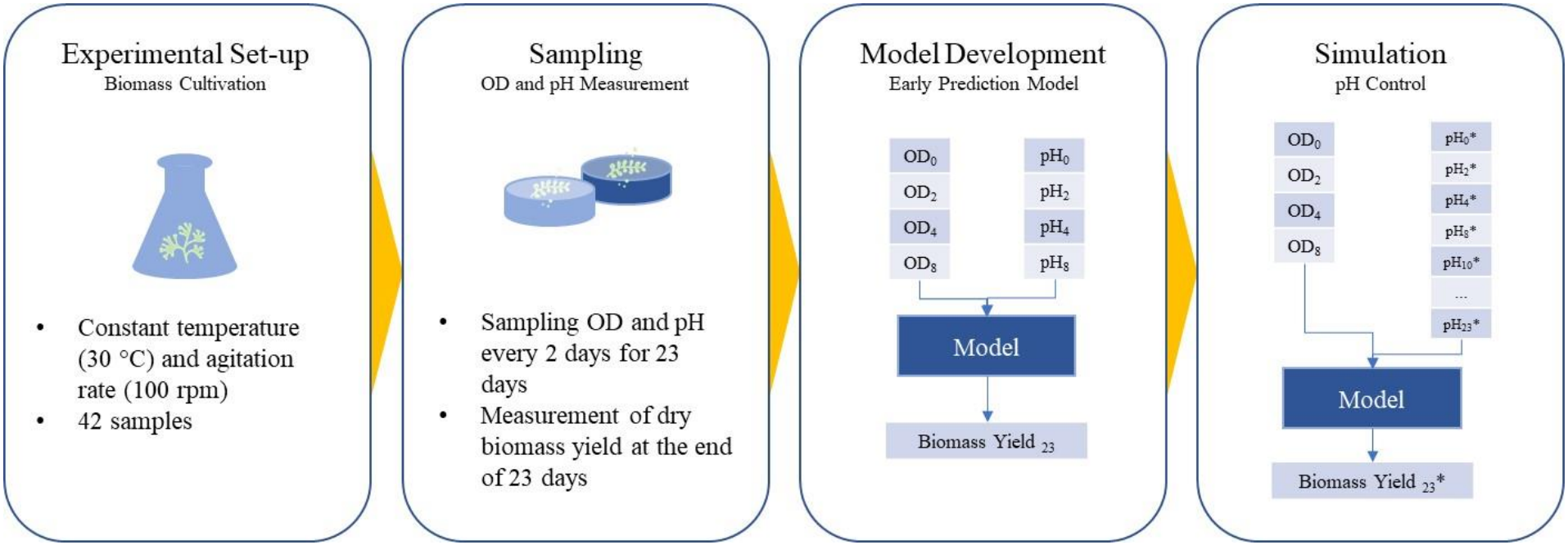
For this study, 42 samples of *S. platensis* were obtained from a freshwater source in Laguna, Philippines. These samples were cultivated in a laboratory set-up using varying cultivation media composed of AZTEC, Zarrouk 1966 and Jourdan. Specifically, all samples used AZTEC media as inoculum. Then, for the initial culture and successive batch feeding, different combinations of the three cultivation media were used. The illumination was also varied between 2,000 lux and 3,500 lux. These factor differences were intended to simulate variations in the environmental conditions for cultivation in a large-scale open-pond cultivation. Specific information on the number of replicates for each condition are given in **Table 1**. The total duration of cultivation was 23 days. During the cultivation period, temperature was maintained at 30 ℃, and a constant agitation rate of 100 rpm was implemented.

Throughout the 23-day cultivation period, optical density (OD) and pH were sampled on alternate days. The OD was measured using a spectrophotometer taken at four wavelengths, 560 nm, 620 nm, 650 nm, and 720 nm. These wavelengths are based on previous experiments to measure the growing parameters of *S. platensis* (Faieta et al., 2021; Zhou et al., 2021), and some related strains (Siedlewicz et al., 2020). At the end of the cultivation period, the sample was filtered and freeze-dried, before measuring the dry weight of the biomass. In summary, a total of 42 samples, consisting of growth parameters and the process outcome were recorded as the dataset to be used in modelling as shown in **Figure 1**. These samples are included as **supplementary information** for this study. The growth parameters used as predictors were 48 measures of OD from different wavelengths and 12 measures of pH sampled on alternate days for 23 days; the process outcome was considered to be biomass yield.

142
143 **Table 1**

144 Design of experiments based on different cultivation media and illumination levels.

Cultivation Medium			Illumination (lux)	Replicates
Inoculum (20%)	**Culture (50%)**	**Feeding (30%)**		
AZTEC	Zarrouk	Zarrouk	2,000	3
AZTEC	Zarrouk	Zarrouk	3,500	3
AZTEC	Jourdan	Jourdan	2,000	3
AZTEC	Jourdan	Jourdan	3,500	3
AZTEC	AZTEC	AZTEC	2,000	3
AZTEC	AZTEC	AZTEC	3,500	3
AZTEC	Zarrouk	AZTEC	2,000	3
AZTEC	Zarrouk	AZTEC	3,500	3
AZTEC	Zarrouk	Jourdan	2,000	3
AZTEC	Zarrouk	Jourdan	3,500	3
AZTEC	AZTEC	Jourdan	2,000	3
AZTEC	AZTEC	Jourdan	3,500	3
AZTEC	Jourdan	AZTEC	2,000	3
AZTEC	Jourdan	AZTEC	3,500	3



145

146 **Figure 1.** Process flow for development and validation of the early prediction model.

147 *2.2. Ridge regression model development*

148 The predicted value for biomass yield under ridge regression follows a linear format (eq. 1).
149 The inclusion of a regularization term in the objective function that is used to train the model
150 coefficients makes the ridge regression from ordinary least squares (OLS) linear regression better
151 from other methods. Specifically, the model coefficients are penalized by the L^2 norm with λ, a small
152 constant > 0 (eq. 2). Its function is to shrink the value of the model coefficients, lowering the
153 sensitivity of the model to changes in its input value. This can prevent the model from overfitting its
154 training set (i.e., the data used to calibrate the model coefficients), particularly if the dataset does not
155 capture the complete distribution of each parameter. The difference is mainly in the calculation of
156 model coefficients for ridge regression and OLS linear regression, which are shown in eqs. 3 and 4
157 respectively, where *X* represents the matrix of all samples for all predictors, I_p represents an identity
158 matrix, and *y* represents the vector of all samples of the predicted variable.
159

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \quad \text{eq. 1}$$

$$min_{\beta_o,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_o - {x_i}^T \beta)^2 + \lambda \sum_{j=1}^{M} {\beta_j}^2 \right\} \quad \text{eq. 2}$$

$$\beta_{ridge} = \left(X^T X + \lambda \mathrm{I}_p\right)^{-1} X^T y \quad \text{eq. 3}$$

$$\beta_{OLS} = (X^T X)^{-1} X^T y \quad \text{eq. 4}$$

160
161 *2.3. Other methods for modelling with collinear input variables*

162 Parameter selection refers to the selection of important predictors and the removal of any that
163 are redundant (i.e., collinear to the selected parameters) or unnecessary. For linear regression models,
164 parameter selection can be integrated in the calibration of model coefficients. This is shown in the
165 objective function for training lasso regression (eq. 5). It differs from ridge regression by its use of the
166 L^1 norm in penalizing the model coefficients, as opposed to the L^2 norm. When eq. 5 is optimized, the
167 model coefficients of some predictors are reduced to 0 resulting in the removal of some redundant
168 predictors.
169

$$min_{\beta_o,\beta}\left\{\sum_{i=1}^{N}(y_i-\beta_o-x_i{}^T\beta)^2+\lambda\sum_{j=1}^{M}|\beta_j|\right\} \quad \text{eq. 5}$$

Alternatively, principal components analysis (PCA) can be applied, which has the effect of extracting interaction terms between collinear predictors. In general, the presence an interaction term such as x_1x_2 in eq. 6 can improve the generalizability of the model. Specifically, it represents cases of interactions that may not be represented in the dataset. However, using a model such as eq. 6 in multivariate regression results in there being more predictors than samples, making the problem unsolvable. PCA retains the functionality of interaction terms without the additional terms by redefining all predictors as linear combinations z_k (eq. 7) of the original predictors x_i, such that the number of z_k variables are less than the number of x_i.

$$\hat{y}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2 \quad \text{eq. 6}$$

$$z_k=\sum_{i=1}^{p}\varphi_{ik}x_i \quad \text{eq. 7}$$

The values of z_k are optimized according to the constraints in eqs. 8 and 9. These revised predictors are then used to fit a model. The predicted value is given in eq. 10, where there are m predictors, such that $m < p$.

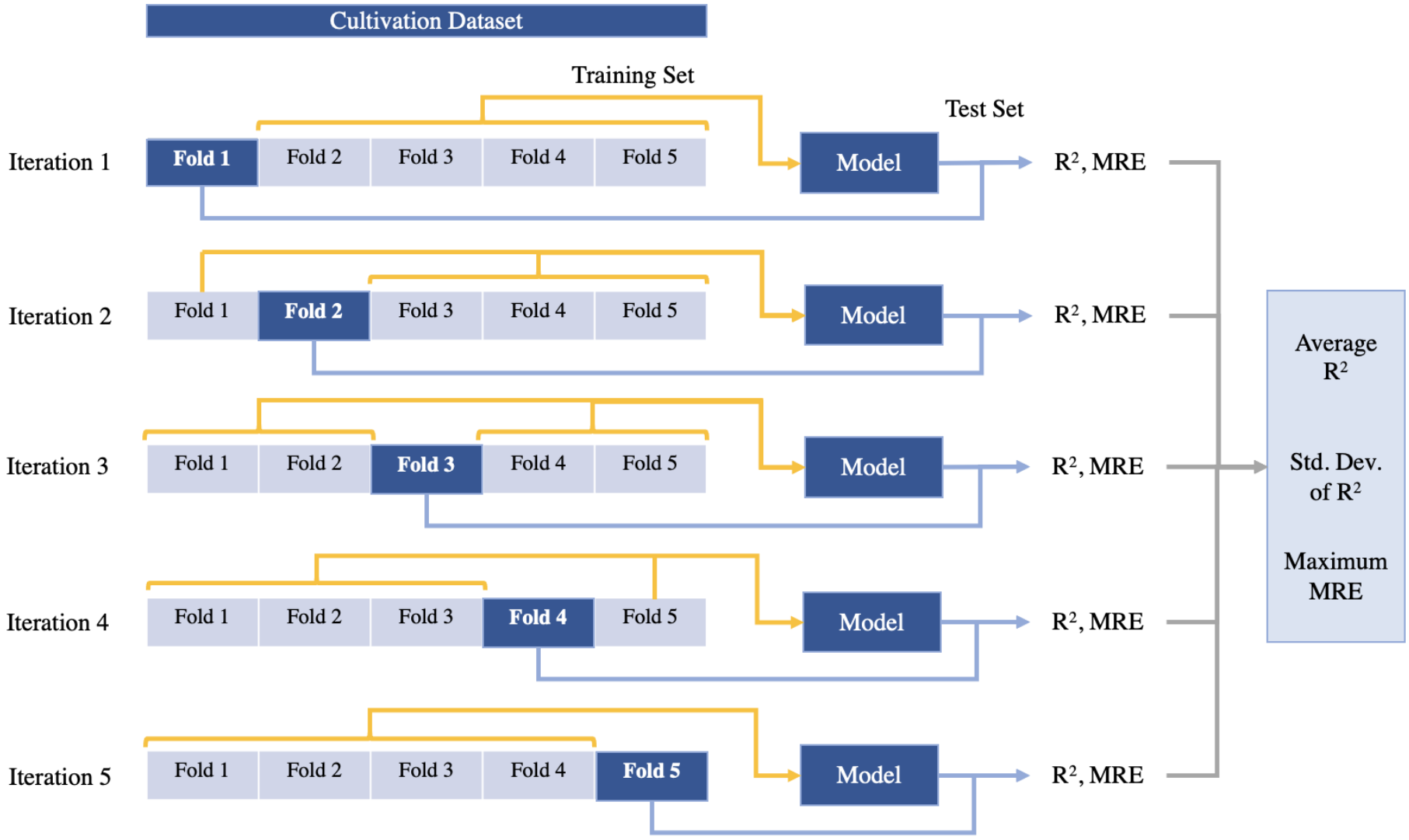
$$z_1>z_2>\cdots>z_m \quad \text{eq. 8}$$

$$\sum_{i=1}^{p}\varphi_i^2=1 \quad \text{eq. 9}$$

$$\hat{y}=\theta_0+\theta_1z_1+\cdots+\theta_mz_m \quad \text{eq. 10}$$

2.3. Evaluating model performance

The dataset is relatively small (42 samples) compared to the number of predictors. This means that the dataset may not represent the complete distribution of the predicted variable. Thus, even if the model performs well on the available data, it may perform poorly on data from the same distribution but not included in the dataset, depending on the model used. Data should be sufficient and the model should be capable of identifying the general behaviour, without overfitting on the available samples. To test if the model is generalizable enough for modelling cultivation, this study employed 5-fold cross validation. This means that the modelling approach was tested on 5 folds (i.e., 5 segments with no overlap) of the dataset. 5-fold cross validation was chosen as opposed to the more common 10-fold cross validation because of the limited number of samples in the dataset, i.e. using 10 folds would result in a test set size of 4-5 samples whereas using 8-9 samples. For each fold, the corresponding training set contained all the samples not included in that specific fold as shown in **Figure 2**. A successful model should have good performance and consistency, which was evaluated in terms of the average and standard deviation of R^2.



201

202 **Figure 2.** Process flow for model evaluation with 5-fold cross validation and the relevant performance measures.

The three performance measures used in this study are average R^2, the standard deviation of R^2, and the maximum residual error (MRE). The metric R^2 was selected because of its comparability for other models by having a typical range from -1 to 1 (eq. 11). Average R^2 shows the general accuracy of the modelling approach. To supplement this, its standard deviation shows how much model performance may change based on the specific components of the training set and test set, and the general variability within the dataset. Finally, the MRE (eq. 12) was selected to give perspective on the maximum potential deviation from the true value, and on performance based on the actual units of measure for biomass yield (g/L).

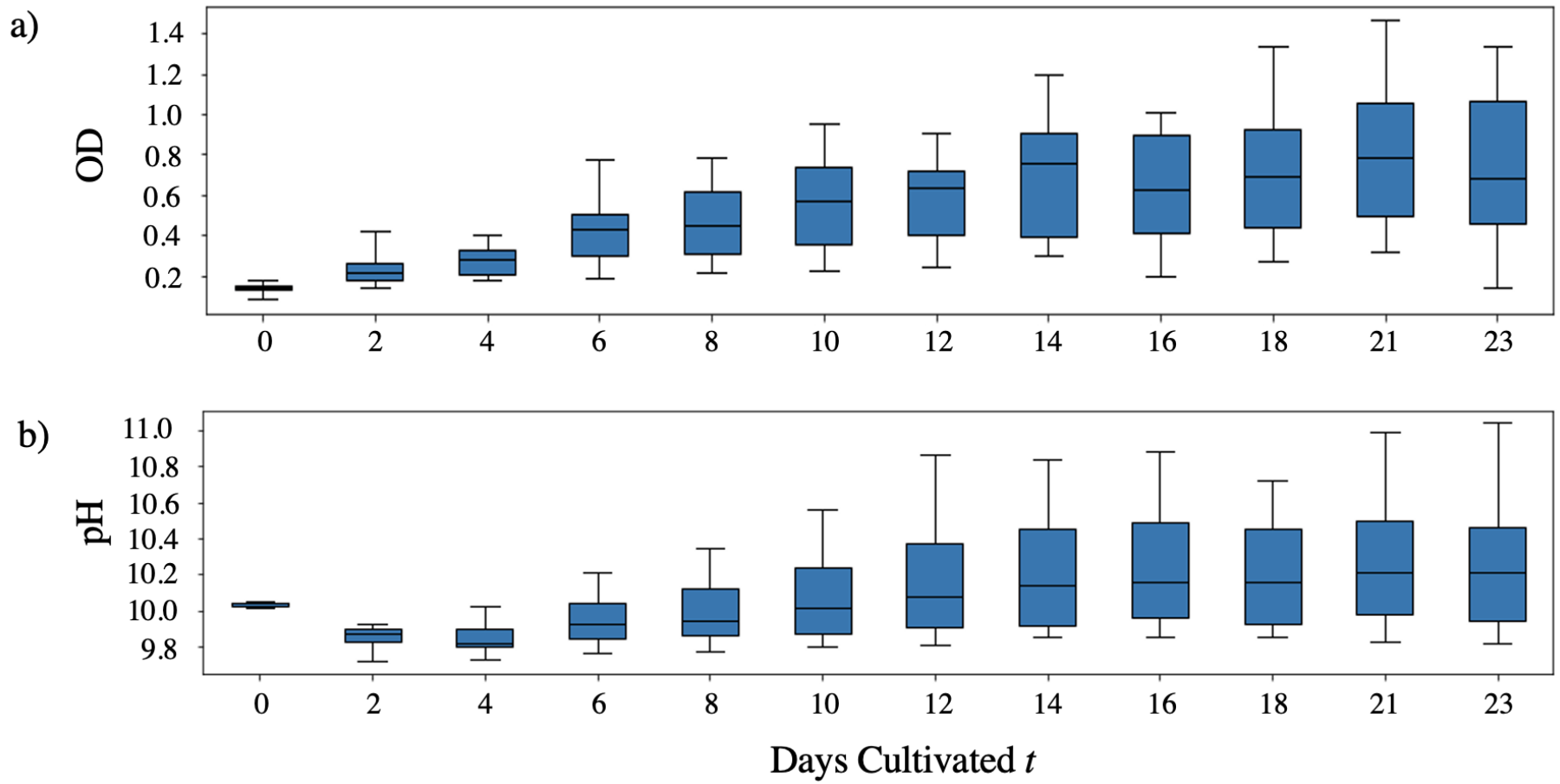
$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad \text{eq. 11}$$

$$MRE = \max_i |y_i - \hat{y}_i| \quad \text{eq. 12}$$

For the initial model, there were 60 potential predictors (48 measures of OD and 12 measures of pH). However, it is understood that these predictors are collinear through (1) the collinearity of these parameters over time, and (2) the collinearity between each OD measure taken on the same day using four different wavelengths. This was addressed by selecting the wavelength which served as the best estimator for biomass yield, based on model accuracy.

2.4. Development of early prediction model

The feasibility of developing an early prediction model was confirmed by evaluating the ridge regression coefficients. Based on previous studies, concurrent measures of OD and biomass concentration have a linear relationship. However, high model coefficients in earlier values of OD would indicate that the information needed to estimate biomass concentration is already present in earlier stages of cultivation. This hypothesis was based on the tendency of OD and pH to plateau as cultivation progresses as observed in **Figure 3**.



225

226 **Figure 3.** Series plot of (a) OD and (b) pH throughout the cultivation cycle.

To develop the early prediction model, sensitivity analysis was conducted by varying the 'earliest prediction day', such that input variables from the start of cultivation (Day 0) to the earliest prediction day are used as predictors. Although having less input variables and using earlier data points may negatively affect the model accuracy, there is potential to offset this through the reduction in variables that are collinear along *t.* The objective of this approach is to identify the earliest point in time that an accurate prediction of biomass yield can be made.

2.5. Application of early prediction model in pH control

Studies on the characteristics of *S. platensis* have observed that the culture exhibits signs of deterioration close to pH 11, and have identified an optimum at pH 10.5 (Richmond & Grobbelaar, 1986). The potential impact of controlling pH at 10.5 was evaluated using the early prediction model, extended to include pH from the earliest prediction day onwards as controllable variables. This demonstrates the potential of using the early prediction model to simulate solutions; in this case, it specifically determines the impact of enhancing pH control to improve biomass yield.

First, the Wilcoxon signed-rank test was used to compare the observed and predicted biomass yield. If the early prediction model could fairly represent the cultivation process, then the test should show that there are no significant differences between the measured and predicted values. Then, the Wilcoxon signed-rank test was used to compare the values of biomass yield without pH control and the values with simulated pH control. If the benefits to be gained from pH control are substantial, there should be a significant difference between the measured and controlled values, and between the predicted and controlled values.

3. Results and discussion

3.1. Biomass cultivation model based on ridge regression

The dataset consisted of the measurements of four wavelengths used in OD measurement, which could be used as predictors for biomass yield. However, it is understood that concurrent measures of OD taken at different wavelengths are collinear. As such, the first means of reducing collinearity was to select a single wavelength for measuring OD values as predictors of biomass yield. The comparison of models using different wavelengths for OD as predictors is shown in **Table 2**.

Based on the performance of the model using the wavelengths 560 nm, 620 nm, 650 nm and 720 nm, good accuracy was achieved between the range of 620 nm and 650 nm. As the OD measured at 620 nm yielded the highest accuracy, this was selected as the predictor to be used in further analysis.

Table 2

Selection of optimal wavelength for modelling biomass cultivation and predicting biomass yield based on model accuracy based on 5-fold cross validation.

Wavelength for measuring OD	**R^2**		**MRE**
	Average	*Standard Deviation*	*(in g/L)*
Baseline	0.733	0.091	0.145
560 nm	0.674	0.134	0.214
620 nm	0.754	0.076	0.144
650 nm	0.747	0.084	0.184
720 nm	0.623	0.174	0.245

This study proposes ridge regression as the appropriate way of modelling the cultivation process given that (1) the predictors are in a time series, (2) the predictors are collinear, and (3) and the number of samples is limited. A comparison of the different methods for addressing collinearity is given in **Table 3**. First, it can be observed that applying any method for addressing collinearity, such as the proposed L^2 regularization, parameter selection with L^1 regularization, or interaction terms, makes a significant improvement over models with no methods to address collinearity (i.e., OLS). As hypothesized based on the context of application, ridge regression, or linear regression with L^2 regularization, had the highest accuracy compared to models that used other methods to address collinearity.

Table 3

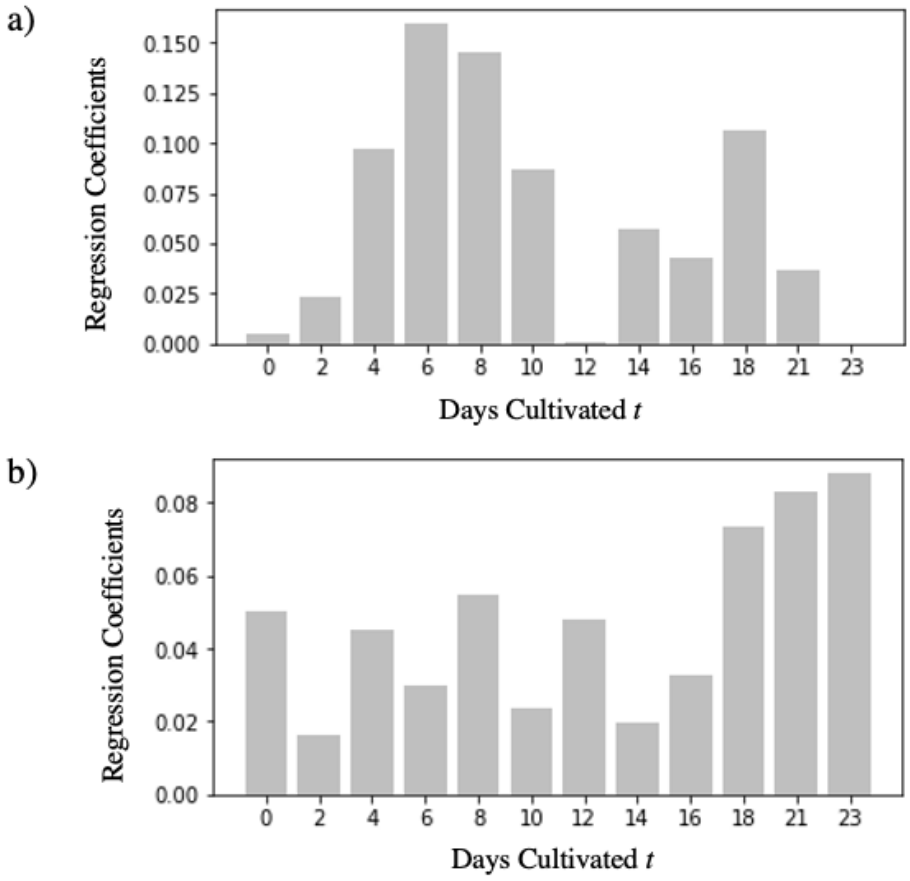
Comparison of methods for addressing collinearity in modelling cultivation and predicting biomass yield based on 5-fold cross validation.

Modelling Approach		**R^2**		**MRE**
Approach for Collinearity	*Model*	*Average*	*Standard Deviation*	*(in g / L)*
None	OLS linear regression	0.238	0.466	0.282
L^2 regularization	Ridge regression	0.754	0.076	0.144
Parameter selection (via L^1 regularization)	Lasso regression	0.446	0.202	0.308
Interaction terms (via PCA)	OLS linear regression with PCA	0.506	0.325	0.232
L^2 regularization and Interaction terms	Ridge regression with PCA	0.754	0.076	0.144
Parameter selection and Interaction terms	Lasso regression with PCA	0.484	0.106	0.266

In addition, the study considered integrating interaction terms with ridge regression and lasso regression, or linear regression with L^1 regularization. Using interaction terms improved the performance of lasso regression. However, it did not improve the performance of ridge regression. This could be because the number of interaction terms that could be generated using PCA was limited by the number of predictors and samples; thus, applying PCA did not make any further improvement in this case.

3.2. Prediction of final biomass yield by the 8^{th} day of cultivation

The absolute values of the coefficients of the ridge regression model as shown in **Figure 4** support the possibility of early prediction, as the highest coefficients for OD are from Days 6 and 8. There is a reduction in the absolute values of the OD coefficients after the 10^{th} day, which can be attributed to the tendency of growth to plateau, as shown in **Figure 3**. There is also an increase beginning at the 18^{th} day which supports the insight from past studies, that concurrent measures of OD and biomass concentration are related.



289

290 **Figure 4.** Absolute values of ridge regression coefficients for (a) OD and (b) pH for each day of

291 cultivation *t.*

292

293 The earliest day on which yield could be predicted was defined as the range of days from 0 to

294 *t* that can be used to predict biomass yield. Sensitivity analysis based on the earliest prediction day

295 showed steady improvement in average R^2 and MRE from the start (Day 0) to the 8^{th} day of

296 cultivation as shown in **Figures 5a** and **5c**. A meaningful reduction in the standard deviation of R^2 can

297 be observed from the 2^{nd} to 8^{th} day as shown in **Figure 5b.** The consistency of performance

298 beginning on the 2^{nd} day is significant as this is the point when R^2 values become positive. As the

299 improvement stagnated after the 8^{th} day, and even marginally worsened between the 12^{th} and 14^{th} days,

300 it was concluded that ridge regression could make an early prediction of final biomass yield at the 8^{th}

301 day of cultivation.

302

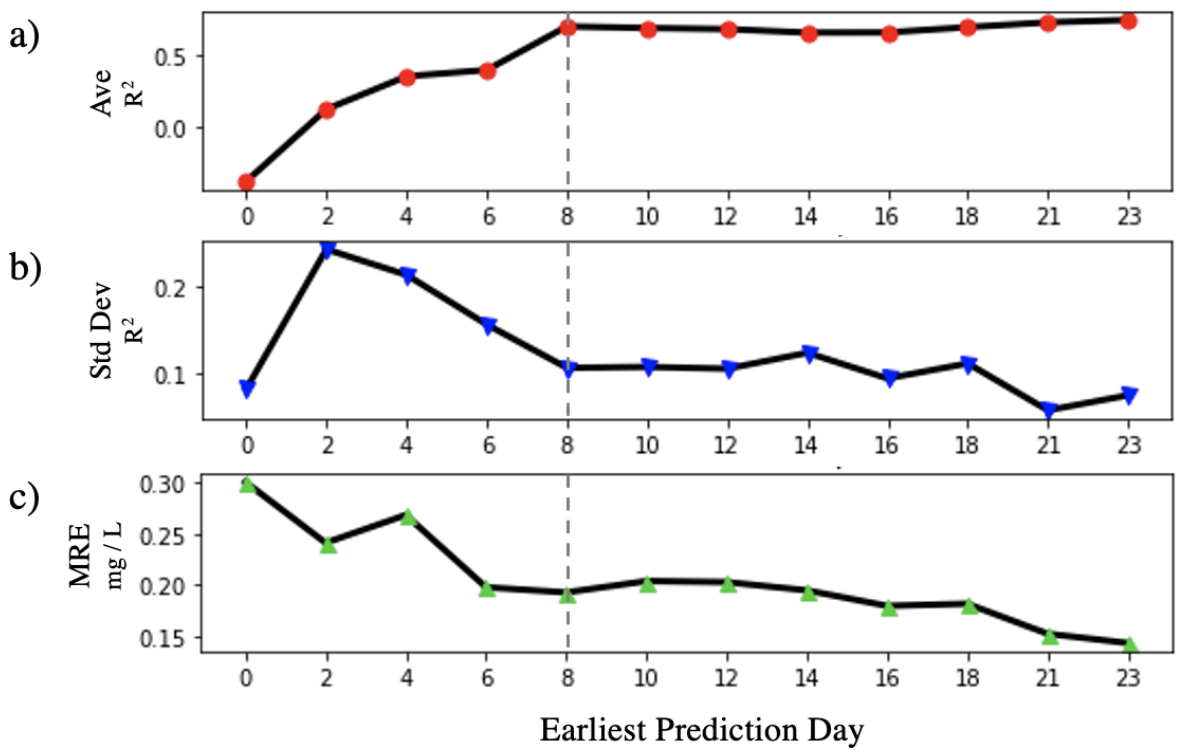


Figure 5. Change in (a) average R^2, (b) standard deviation of R^2 and (c) MRE based on the earliest prediction day (i.e., Days from 0 to current day included as predictors).

3.3. pH control to improve biomass yield

In this section, the study considers the potential impact of controlling pH at a recommended setpoint in improving biomass yield. It compares three values of biomass yield as shown in **Figure 6**. The measured values represent the biomass yield obtained from the experiment. The predicted values of biomass yield are obtained from a ridge regression model trained using measures of OD from Days 0 to 8, and pH from Days 0 to 23 as predictors. The measured and predicted values of biomass yield were compared using a Wilcoxon signed-rank test, which confirmed that the measured and predicted values came from the same population. This means that the cultivation model used to predict biomass yield is accurate.

Using the pH readings recorded throughout the cultivation cycle as inputs for predicting biomass yield was intended to compare the case where pH is left uncontrolled (as in the measured and predicted values), and the case where pH is controlled at 10.5. The improved values, as shown in **Figure 6,** represent the case were pH is controlled at 10.5 from Days 0 to 23. The simulation showed

that controlling pH resulted in an average improvement of 54.1% in biomass yield as observed in **Figure 6**. The improvement was likewise validated with a Wilcoxon signed-rank test, which confirmed that there are significant differences between the measured and improved values of biomass yield, and between the predicted and improved values of biomass yield.

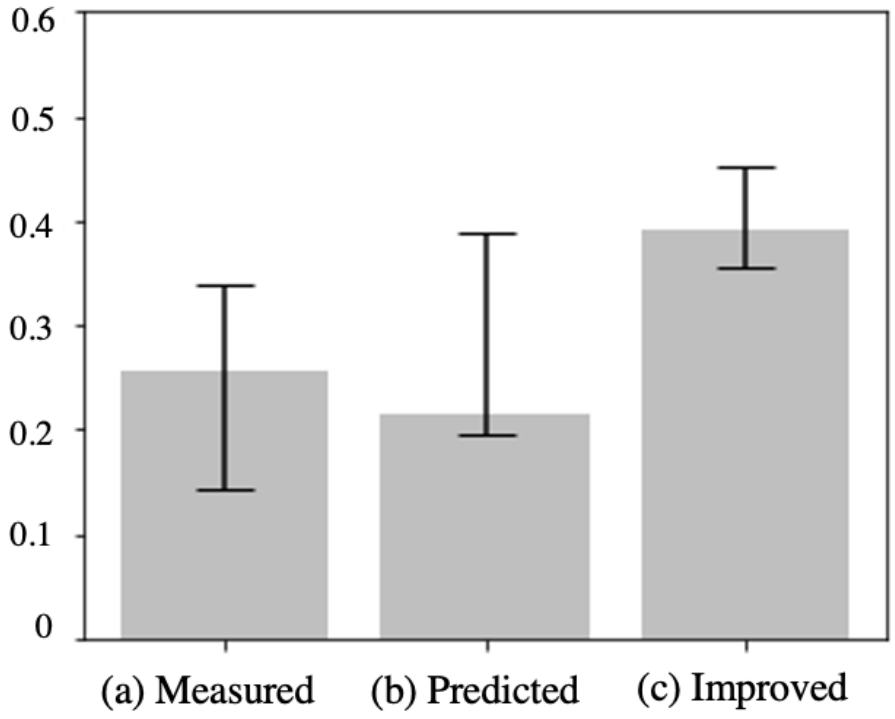


Figure 6. Median values of biomass yield that is (a) measured, (b) predicted by the early prediction model (i.e., ridge regression model using OD from Days 0 to 8, and pH from Days 0 to 23 as inputs), and (c) simulated and improved with pH control (i.e., ridge regression model wherein the pH from Days 8 to 23 is controlled). Inter-quartile range is reflected in the error bars. Measured and predicted values are statistically similar, and improved values are significantly larger than measured and predicted values ($p < 0.05$).

3.4. Potential commercial applications of the model

An early prediction model may be utilized in the design and management of biomass cultivation systems, and may encourage improved and sustainable commercialization of third-generation biofuels and other value-added products. Specifically, a powerful prediction mode is useful in enhancing biomass growth rate and product synthesis, which remains to be one of the most challenging and crucial issues in this industry.

The ability to accurately predict biomass yield early in the cultivation period allows cultivation systems to be designed cost-effectively with a balanced accurate expectation of yield. Thus, the proposed model may be integrated into a decision support system. The earlier that sub-optimal yields and other issues can be identified, the earlier interventions can be implemented. This avoids wastage of resources (i.e., labor, capital, water, land, nutrients, etc.) on a problematic cultivation batch. Furthermore, selections between cultivation media, technologies (e.g., photobioreactors), and environmental conditions (e.g. light, temperature, CO_2, etc.), as well as design of new technologies, strain improvement strategies, and other emerging opportunities may be accelerated. In addition to techno-economic considerations, any environmental consequences and impacts can be mitigated and controlled.

4. Conclusions

Managing biomass cultivation remains challenging because of its heavy dependence on environmental factors and the inherent uncertainty of working with any living system. The process spans several days and weeks with variations in process parameters, making it difficult to ensure final biomass yield just from initial conditions. Predictive modelling enables the use of historical data to predict future unknown outcomes. Early prediction allows for risks and issues to be identified early in the process; thus allowing solutions to be implemented to ensure that sufficient biomass is produced to support biofuel production. This study developed an early prediction model, enabling the prediction of biomass yield at the end or 23^{rd} day of cultivation by the 8^{th} day of cultivation. This prediction was made using a ridge regression model, and based on measures of OD and pH taken during the

cultivation process. The advantage of ridge regression over the other methods tested in this study was attributed to L^2 regularization, which was observed to be suitable for modelling with time-collinear predictors in this context. Furthermore, unlike existing prediction modelling techniques which predict based only on initial conditions determined from empirical optimization models, the approach proposed in this study captures inherent process dynamics through a rolling inclusion of parameters as cultivation progresses. The early prediction model may be integrated with control and optimization systems to achieve better biomass yield. To demonstrate its effectiveness in analytics for improving biomass yield, the early prediction model was applied in simulating pH control at a recommended setpoint of 10.5. The solution resulted in an average improvement of 54.1%, with significantly reduced variability. A decision support system is also a natural extension of this model to determine appropriate interventions to implement during cultivation. The decision support can be further extended to correct cultivation conditions for various growth scenarios. With this, resources, such as material, land, labor, and capital, are utilized efficiently, while waste, pollution, and other negative impacts are minimized.

For large-scale applications, the study recommends the use of larger datasets to improve accuracy, as the standard deviation of R^2 from cross validation indicated that further improvement is possible. Moreover, future studies may explore the use of other cultivation process parameters in early prediction modelling to improve biomass yield and thereby energy production. In this way, one of the major roadblocks, namely inconsistent and insufficient biomass supply, to the transition to bioenergy can be addressed.

Declarations

The authors acknowledge the support of the University Research Coordination Office of De La Salle University (DLSU); the Office of the Vice Chancellor for Research and Innovation of DLSU; and Department of Science and Technology of the Philippines through the Engineering Research and Development for Technology grant. This work is also partially funded by the Innovative Technology Commission through SST_182_20GP. The data used in this study is included as Supplementary

Information. The authors also thank Mr. Yudong He and Mr. Jessie James Malit for their technical advice.

Conflicts of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Azari, A., Tavakoli, H., Barkdoll, B. D., & Haddad, O. B. (2020). Predictive model of algal biofuel production based on experimental data. *Algal Research, 47*, 101843. https://doi.org/10.1016/j.algal.2020.101843

Banerjee, S., Ray, A., & Das, D. (2020). Optimization of chlamydomonas reinhardtii cultivation with simultaneous CO2 sequestration and biofuels production in a biorefinery framework. *Science of The Total Environment*, 143080. https://doi.org/10.1016/j.scitotenv.2020.143080

Barbosa, R. C., Soares, J., & Martins, M. A. (2020). Low-cost and versatile sensor based on multi-wavelengths for real-time estimation of microalgal biomass concentration in open and closed cultivation systems. *Computers and Electronics in Agriculture, 176*, 105641. https://doi.org/10.1016/j.compag.2020.105641

Behera, B., Aly, N., & Balasubramanian, P. (2019). Biophysical model and techno-economic assessment of carbon sequestration by microalgal ponds in Indian coal based power plants. *Journal of Cleaner Production, 221*, 587-597. https://doi.org/10.1016/j.jclepro.2019.02.263

Brindhadevi, K., Mathimani, T., Rene, E. R., Shanmugam, S., Chi, N. T. L., & Pugazhendhi, A. (2021). Impact of cultivation conditions on the biomass and lipid in microalgae with an emphasis on biodiesel. *Fuel, 284*, 119058.

Caligan, C. J., Garcia, M. M., Mitra, J. L., Mayol, A. P., San Juan, J. L. G., & Culaba, A. B. (2020). Multi-objective optimization of water exchanges between a wastewater treatment facility and algal biofuel production plant. *IOP Conference Series: Earth and Environmental Science, 463*, 012050. https://doi.org/10.1088/1755-1315/463/1/012050

Culaba, A. B., Juan, J. L. G., Ching, P. M. L., Mayol, A. P., Sybingco, E., & Ubando, A. (2019a). Optimal synthesis of algal biorefineries for Biofuel production based on techno-economic and environmental efficiency. *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. https://doi.org/10.1109/hnicem48295.2019.9072730

Culaba, A. B., Ching, P. M. L., San Juan, J. L., Philip Mayol, A., Sybingco, E., & Ubando, A. T. (2019b). A dynamic sustainability assessment of algal biorefineries for Biofuel production. *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. https://doi.org/10.1109/hnicem48295.2019.9072721

Elavarasan, R. M., Pugazhendhi, R., Jamal, T., Dyduch, J., Arif, M. T., Kumar, N. M., ... & Nadarajah, M. (2021). Envisioning the UN Sustainable Development Goals (SDGs) through the lens of energy sustainability (SDG 7) in the post-COVID-19 world. *Applied Energy, 292*, 116665.

Faieta, M., Neri, L., Di Michele, A., Di Mattia, C. D., & Pittia, P. (2021). High hydrostatic pressure treatment of Arthrospira (Spirulina) platensis extracts and the baroprotective effect of sugars on phycobiliproteins. *Innovative Food Science & Emerging Technologies*, 102693.

Liyanaarachchi, V. C., Premaratne, M., Ariyadasa, T. U., Nimarshana, P. H. V., & Malik, A. (2021). Two-stage cultivation of microalgae for production of high-value compounds and biofuels: A review. *Algal Research, 57*, 102353.

Mowbray, M., Savage, T., Wu, C., Song, Z., Cho, B. A., Del Rio-Chanona, E. A., & Zhang, D. (2021). Machine learning for biochemical engineering: A review. *Biochemical Engineering Journal*, 108054.

Nayak, M., Dhanarajan, G., Dineshkumar, R., & Sen, R. (2018). Artificial intelligence driven process optimization for cleaner production of biomass with Co-valorization of wastewater and flue gas in an algal biorefinery. *Journal of Cleaner Production, 201*, 1092-1100. https://doi.org/10.1016/j.jclepro.2018.08.048

Richmond, A., & Grobbelaar, J. U. (1986). Factors affecting the output rate of Spirulina platensis with reference to mass cultivation. *Biomass, 10*(4), 253-264.

San Juan, J. L. G., Mayol, A. P., Sybingco, E., Ubando, A. T., Culaba, A. B., Chen, W. H., & Chang, J. S. (2020). A scheduling and planning algorithm for microalgal cultivation and harvesting for biofuel production. *IOP Conference Series: Earth and Environmental Science, 463*, 012010. https://doi.org/10.1088/1755-1315/463/1/012010

Schade, S., & Meier, T. (2021). Techno-economic assessment of microalgae cultivation in a tubular photobioreactor for food in a humid continental climate. *Clean Technologies and Environmental Policy*, 1-18.

Siedlewicz, G., Żak, A., Sharma, L., Kosakowska, A., & Pazdro, K. (2020). Effects of oxytetracycline on growth and chlorophyll a fluorescence in green algae (Chlorella vulgaris), diatom (Phaeodactylum tricornutum) and cyanobacteria (Microcystis aeruginosa and Nodularia spumigena). *Oceanologia, 62*(2), 214-225.

Solis, C. A., Mayol, A. P., San Juan, J. G., Ubando, A. T., & Culaba, A. B. (2020). Multi-objective optimal synthesis of algal biorefineries toward a sustainable circular bioeconomy. *IOP Conference Series: Earth and Environmental Science, 463*, 012051. https://doi.org/10.1088/1755-1315/463/1/012051

Zhou, T., Cao, L., Zhang, Q., Liu, Y., Xiang, S., Liu, T., & Ruan, R. (2021). Effect of chlortetracycline on the growth and intracellular components of Spirulina platensis and its biodegradation pathway. *Journal of Hazardous Materials, 413*, 125310.

Žitnik, M., Šunta, U., Godič Torkar, K., Krivograd Klemenčič, A., Atanasova, N., & Griessler Bulc, T. (2019). The study of interactions and removal efficiency of escherichia coli in raw blackwater treated by microalgae chlorella vulgaris. *Journal of Cleaner Production, 238*, 117865. https://doi.org/10.1016/j.jclepro.2019.117865