processes

MDPI

Article

Data-Driven Ensemble Machine Learning for Multi-Horizon Microalgal Bioprocess Forecasting

Bartolomeo Cosenza [1,2], Riccardo Minardi [3], Luca Usai [3], Riccardo Allodi [3], Alessandro Concas [4,5], Daniele Sofia [6], Giancarlo Cravotto [1], Giovanni Denaro [7], Antonio Messineo [8], Maurizio Volpe [8], Antonio Picone [8], Robinson Soto-Ramirez [9], Catalina Valencia Peroni [10] and Giovanni Antonio Lutzu [3,*]

1 Department of Drug Science and Technology, University of Turin, Via P. Giuria 9, 10125 Turin, Italy; bartolomeo.cosenza@unito.it (B.C.); giancarlo.cravotto@unito.it (G.C.)
2 Department of Civil and Industrial Engineering, University of Pisa, Largo Lucio Lazzarino, 56122 Pisa, Italy
3 Teregroup Srl, Via David Livingstone 37, 41123 Modena, Italy; riccardo.minardi@teregroup.net (R.M.); luca.usai@teregroup.net (L.U.); riccardo.allodi@teregroup.net (R.A.)
4 Department of Mechanical, Chemical and Materials Engineering, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy; alessandro.concas@unica.it
5 Interdepartmental Center of Environmental Science and Engineering (CINSA), University of Cagliari, Via San Giorgio 12, 09124 Cagliari, Italy
6 Department of Computer, Modeling, Electronics and Systems Engineering, University of Calabria, Via P. Bucci Edificio 42, 87036 Rende, Italy; daniele.sofia@unical.it
7 Department of Earth and Marine Science (DiSTeM), University of Palermo, Via Archirafi 36, 90123 Palermo, Italy; giovanni.denaro@unipa.it
8 Department of Engineering and Architecture, University of Enna, Kore, Cittadella Universitaria, 94100 Enna, Italy; antonio.messineo@unikore.it (A.M.); maurizio.volpe@unikore.it (M.V.); antonio.picone@unikore.it (A.P.)
9 Departamento de Ingenieria de Procesos y Bioproductos, Facultad de Ingenieria, Universidad del Bio-Bio, Avda Collao 1202, Concepción 4051381, Chile; rdsoto@ubiobio.cl
10 Departamento de Procesos y Energia, Facultad de Minas Universidad Nacional de Colombia-Sede Medellin, Medellín 050034, Colombia; cavalenciapa@unal.edu.co
* Correspondence: gianni.lutzu@teregroup.net; Tel.: +39-0594823699

Abstract

Microalgal bioprocesses increasingly rely on predictive modelling to support automated control and reduce the cost of laboratory experimentation. Yet, the scarcity of high-quality datasets and the nonlinear nature of microalgal growth severely limit the accuracy and robustness of conventional machine-learning approaches. This study introduces an ensemble-learning framework for forecasting biomass accumulation in *Limnospira platensis* cultures supplemented with Effective Microorganism (EM) consortia. The method combines temporally consistent, strictly causal feature engineering with tree-based ensemble learning under a prospective validation protocol designed to mirror real deployment. Growth measurements are transformed into temporal descriptors that encode phase transitions, short-term growth dynamics, and treatment effects, enabling tree-based ensemble algorithms to capture nonlinear patterns inaccessible to traditional models. When features are restricted to information available strictly before each prediction, and models are evaluated only on real held-out measurements, one-step nowcasting does not surpass a trivial persistence baseline. For genuine multi-horizon forecasting of a monitored culture, the regime with practical value, a horizon-aware gradient-boosting model trained jointly on two cultivation media (Jordan and Zarrouk), attains $R^2 \approx 0.72$ on the held-out Jordan observations (0.72 ± 0.02, mean $\pm$ SD over 30 seeds; RMSE ≈ 0.20 g L^{-1}, Pearson r ≈ 0.85) and remains stable across forecast horizons of 1–28 days, outperforming persistence roughly fourfold in pooled R^2 (0.72 vs. 0.16) at medium-to-long horizons where the naive baseline collapses. Permutation-based feature-importance analysis identifies current biomass and the EM dilution level as the leading predictors, whereas EM treatment identity contributes only modestly. Including

Academic Editor: Hah Young Yoo

Received: 16 June 2026
Revised: 30 July 2026
Accepted: 5 August 2026
Published: 7 August 2026

an independent second-medium experiment (Zarrouk) in joint training did not materially change held-out Jordan performance, indicating that, under these data-limited conditions, neither additional model complexity nor a second training medium substantially improved forecasting once the causal, horizon-aware framework was in place. Among treatments, EM3 most strongly suppressed biomass accumulation. Overall, the study provides an honest, fully reproducible, two-medium benchmark for microalgal biomass forecasting under data-limited conditions, and identifies the forecast horizon as the regime in which ensemble learning adds genuine value over trivial baselines.

Keywords: *Limnospira platensis*; effective microorganisms; biomass prediction; random forest; gradient boosting; multi-horizon forecasting; forecast combination; time-series validation

1. Introduction

Microalgal biotechnology has rapidly evolved into a data-driven discipline, driven by the demand for the sustainable production of proteins, pigments, and biofuels [1–3]. *Limnospira* (*Spirulina*) *platensis*, one of the most industrially relevant strains [4], exemplifies the shift from empirical optimisation to computational approaches capable of predicting complex biological outcomes and supporting automated photobioreactor management. Machine-learning (ML) tools enable real-time optimisation of cultivation parameters and early detection of process anomalies, helping overcome the economic and operational constraints that limit traditional trial-and-error approaches. ML is particularly suited for modelling microalgal systems because it can analyse high-dimensional datasets, capture nonlinear interactions among environmental variables, and improve as new data accumulate [5,6]. Despite these advantages, a central barrier persists: controlled microalgal experiments typically yield only 20–50 time-points, far too few to train robust predictive models. This scarcity, combined with the tendency of many studies to ignore temporal dependencies or growth-phase transitions, leads to overfitting and inflated performance metrics. Research on Effective Microorganisms (EMs) provides additional motivation for predictive tools. EM consortia, comprising lactic-acid bacteria, yeasts, actinomycetes, and photosynthetic bacteria, have traditionally been applied as microbial inoculants in agriculture, aquaculture, bioremediation, and water-quality management [7–12]. In these applications, they are intended to modify nutrient transformations and microbial-community structure. Their effects are not uniformly beneficial, however, and may vary substantially with consortium composition, applied dilution, cultivation medium, environmental conditions, and target microorganism. Accordingly, the literature reports both improved nutrient removal and community restructuring [10–12] and limited or negative responses, including dose-dependent inhibition of cyanobacteria [13]. In the present study, EM supplementation was therefore investigated as an exploratory microbial-modulation strategy rather than as an assumed growth-promoting treatment. The hypothesis of our study is that different EM formulations and dilution levels could alter nutrient availability and microbial interactions, potentially enhancing *L. platensis* culture performance or, conversely, causing competitive or metabolite-mediated inhibition. From an applied perspective, identifying such favourable or inhibitory regimes is essential for assessing whether EM consortia can be considered as process additives for *L. platensis* cultivation. The context-dependent nature of these responses also provides a relevant biological case for developing predictive models capable of separating treatment effects from intrinsic physiological variability. Recent advances in AI for microalgal biotechnology can be grouped into several complementary research directions. Data-driven methods have improved the temporal forecasting and monitor-

ing of algal systems, ranging from hybrid deep-learning models for bloom prediction to explainable ensemble approaches, real-time monitoring, and cultivation time-series forecasting [14–17]. In parallel, machine learning has been increasingly applied to cultivation optimisation and biomass prediction, including the selection of relevant process variables for seawater-based *L. platensis* production and graph-learning approaches for commercial microalgae [18,19]. At the methodological level, recent reviews highlight the growing adoption of hybrid, mechanistic, and deep-learning strategies for representing nonlinear biological dynamics [20–22], while dynamic process models and machine-learning tools have also been developed to support biomass and phycocyanin prediction and downstream operations such as harvesting [1,23]. Taken together, these studies demonstrate the expanding role of AI across the microalgal production chain, from cultivation monitoring and forecasting to process optimisation and biomass recovery. However, they also reveal a persistent limitation: most modelling approaches continue to rely on datasets that are small, noisy, and temporally sparse. The present study addresses this gap by combining ensemble learning with past-only feature engineering and chronological validation to prevent look-ahead bias. Despite this progress, to the best of our knowledge, no previous study has systematically developed ML models for *L. platensis* cultivation under EM treatment. This gap limits the deployment of EM-assisted systems, which remain dependent on costly experimental optimisation. Addressing this need, the present study introduces a predictive framework that combines ensemble learning with temporally consistent, strictly causal feature engineering. By enforcing chronological integrity through a blocked time-series split and computing every predictor from information available before the forecast target, the framework captures nonlinear interactions between biomass history, short-term dynamics, growth phase, and EM treatment effects without inflating performance through look-ahead bias. This approach enables accurate biomass forecasting despite extremely small experimental datasets and provides new insight into the relative influence of EM treatments versus intrinsic physiological drivers.

Two distinct objectives are pursued. First, we aim to develop and validate a strictly prospective, reproducible pipeline for forecasting *L. platensis* biomass under data-limited conditions, using strictly past-only features and evaluation protocols that mimic deployment, in line with established guidance on avoiding look-ahead bias in machine-learning studies [24,25]. Second, we aim to quantify, against appropriate baselines (persistence, linear drift and a Gompertz kinetic model), the forecast regime in which ensemble learning adds value, and to assess the relative contribution of EM treatment and dilution versus intrinsic physiological drivers across two cultivation media (Jordan and Zarrouk).

The novelty of this work lies in several aspects. To our knowledge, this is the first machine-learning framework developed specifically for *L. platensis* cultivated under EM treatment, an application for which no predictive tool previously existed. Methodologically, the study contributes a strictly prospective, multi-horizon forecasting benchmark in which every feature is computed from information available before the prediction target, and every performance metric is reported on real held-out test data, providing an honest estimate of the accuracy attainable on small bioprocess datasets. The EM dilution level is introduced as a genuinely informative predictor, a variable previously conflated with treatment identity rather than exploited in its own right. The framework is validated across two distinct cultivation media (Jordan and Zarrouk), evaluated on held-out Jordan observations, and couples a persistence rule at short horizons with gradient boosting at longer ones within a single horizon-aware forecaster. Finally, performance is quantified separately for each forecast horizon and with explicit seed-level uncertainty, clarifying the regime in which data-driven ensembles add value over trivial baselines.

2. Materials and Methods

This experimental study was conducted to evaluate the effects of EM on the photoautotrophic growth dynamics of *L. platensis* under controlled laboratory conditions. The aim was to compare the influence of three distinct EM formulations, designated EM1, EM2, and EM3, on the biomass accumulation and growth dynamics of *L. platensis* cultivated in modified Jordan medium. Each EM formulation was applied at three dilution levels ($1{:}10^2$, $1{:}10^3$ and $1{:}10^5$; see Table 1), here labelled series A, B, and C. These levels were selected as exploratory loading conditions rather than as an evenly spaced logarithmic dilution series. The $1{:}10^2$ and $1{:}10^3$ treatments provided greater resolution within the comparatively high-EM-loading range, whereas $1{:}10^5$ represented a substantially lower-loading condition. The intermediate $1{:}10^4$ dilution was not included in the original experimental design. Consequently, the experiment enables comparison among the three tested loading regimes but does not permit reconstruction of a continuous dose–response relationship or identification of a precise biological threshold. These are therefore dilution-gradient sub-conditions of each consortium rather than biological replicates of a single condition, and are treated as distinct series throughout; the dilution level is additionally retained as an explicit predictor (Section 2.6).

Table 1. Experimental setup showing the combinations of EM dilution and inoculum of *L. platensis* used for each flask.

Experimental Series	Dilution (*v*/*v*)	Jordan Medium + Inoculum (mL)	EM Added (mL)	Total Volume (mL)
EM1 A	$1{:}10^2$	20.0	0.5	20.5
EM2 A	$1{:}10^2$	20.0	0.5	20.5
EM3 A	$1{:}10^2$	20.0	0.5	20.5
EM1 B	$1{:}10^3$	20.0	0.5	20.5
EM2 B	$1{:}10^3$	20.0	0.5	20.5
EM3 B	$1{:}10^3$	20.0	0.5	20.5
EM1 C	$1{:}10^5$	20.0	0.5	20.5
EM2 C	$1{:}10^5$	20.0	0.5	20.5
EM3 C	$1{:}10^5$	20.0	0.5	20.5

2.1. Culture Medium Preparation

The cultivation medium employed in this study was based on the Jordan formulation, a nitrogen-enriched adaptation of the classical Zarrouk medium. The EMJ medium was prepared using distilled water and composed as follows (g L^{-1}): 5 $NaHCO_3$; 1.6 KOH; 5 $NaNO_3$; 0.027 $CaCl_2 \cdot 2H_2O$; 0.4 K_2SO_4, 2 K_2HPO_4; 1 NaCl; 0.4 $MgSO_4 \cdot 7H_2O$; 0.16 EDTA-Na_2; 0.01 $FeSO_4 \cdot 7H_2O$; and 1 mL of Trace elements. The Trace element solution was prepared with the following composition (mg L^{-1}): 250 EDTA-Na_2; 57 H_3BO_3; 110 $ZnSO_4 \cdot 7H_2O$; 25.3 $MnCl_2 \cdot 4H_2O$; 8.05 $CoCl_2 \cdot 6H_2O$; 7.85 $CuSO_4 \cdot 5H_2O$; 5.5 $Mo_7O_{24}(NH_4)_6 \cdot 4H_2O$. The medium was autoclaved at 121 °C for 20 min prior to inoculation and maintained under axenic conditions.

A parallel experiment was carried out in standard Zarrouk medium, following the same EM treatment design (control plus EM1, EM2, and EM3 at the three dilution levels of Table 1), and was sampled at cultivation days 0, 10, 20, 22, 23, and 24. The Zarrouk experiment was used as an additional independent dataset for joint two-medium training and for assessing whether measurements from a second cultivation medium improved forecasting of the held-out Jordan observations. The parallel Zarrouk experiment comprised nine trajectories (an untreated control together with EM1 at three dilution levels, EM2 at two dilution levels, and EM3 at three dilution levels; the EM2 $1{:}10^3$ condition was not available). Under our origin–target construction, in which the earliest observation only seeds the lag features, each trajectory contributed six valid origin–target pairs, for a total of 54 pairs used solely for joint training. These pairs should not be interpreted as independent

biological observations. Because the model was trained on both media and evaluated on held-out Jordan data, this analysis assesses the incremental value of second-medium data rather than strict external cross-medium transfer (Section 2.5, Section 2.6, and Section 3.1).

2.2. Experimental Design

L. platensis inoculum was sourced from a laboratory-maintained unialgal culture, acclimated to EMJ for 5 days prior to the experiment. The culture vessels, glass flasks of 50 mL nominal volume, were filled with 20 mL of EMJ medium containing pre-culture at a known optical density (OD_{565}) concentration of 0.1 and 2.44% (*v*/*v*) of each EM (0.5 mL per flask). Considering the subsequent addition of 0.5 mL of each pre-diluted EM preparation to a final culture volume of 20.5 mL, the approximate final proportions of the original EM stocks were 1:4.1 $\times$ 10^3, 1:4.1 $\times$ 10^4, and 1:4.1 $\times$ 10^6, respectively.

Each EM formulation, designated EM1, EM2, and EM3, was tested at three dilution levels, namely 1:10^2, 1:10^3, and 1:10^5. One flask was assigned to each formulation–dilution combination, yielding nine EM-treated experimental units. The labels A, B, and C identify the three dilution conditions and do not represent biological replicates. Three independent untreated control flasks were additionally maintained, resulting in a total of 12 culture units. Each EM formulation–dilution combination was therefore treated as a distinct experimental series, and the dilution level was retained as an explicit predictor in the modelling framework. For the sake of clarity, a simplification of the experimental setup adopted is shown in Table 1.

The flasks were incubated at a constant temperature of 27 $\pm$ 1 °C. Illumination was provided by cool-white fluorescent lamps delivering an average light intensity of 50 µmol photons m^{-2} s^{-1} on a 24L:0D photoperiod. The cultures were stirred constantly with the use of electric shakers to avoid sedimentation and ensure homogeneity.

Across both media, the experimental design comprised an untreated control together with the three EM consortia (EM1, EM2 and EM3), each applied at three dilution levels and grown in separate flasks (series A, B and C, as in Table 1). The Jordan experiment was monitored over a 28-day cultivation cycle and the parallel Zarrouk experiment over 24 days. At each sampling point, the OD_{565} of every flask was recorded and converted to dry-weight biomass, as described in Section 2.3, and the dispersion observed across the dilution series was retained to characterise the natural variability of each growth trajectory. The resulting time series were densely sampled during the early exponential phase and more sparsely during the late stationary phase, a sampling structure typical of small, controlled microalgal experiments and central to the forecasting task addressed in this work.

The three EM formulations used in this experiment were kindly provided by Prof. Luigi Vertua from Cologne (BS), Italy. The three samples were supplied as liquid mixed-microbial preparations in an aqueous carrier. Their taxonomic composition and the viable abundance of individual microbial groups were not disclosed by the supplier and were not independently characterised in the present study. Accordingly, the possible presence of microbial groups commonly associated with EM formulations, including lactic-acid bacteria, yeasts, photosynthetic bacteria, and actinomycetes, could not be experimentally confirmed. The formulations were therefore treated as operationally distinct EM1, EM2, and EM3 treatments rather than as taxonomically defined inocula. The absence of batch-specific viable counts and compositional characterisation limits exact biological reproducibility and prevents the observed effects from being attributed to a particular microbial taxon.

Each of the three EM formulations was pre-diluted in sterile distilled water to achieve volumetric ratios of 1:100, 1:1000, and 1:100,000, respectively, prior to application. From each diluted solution, 0.5 mL was aseptically added to 20 mL of culture medium, yielding a final dosage of 2.44% (*v*/*v*). This dosage was selected based on literature precedents

describing typical EM application rates in bioremediation and aquaculture contexts, where 2–3% additions are commonly adopted for microbial modulation of aquatic environments.

2.3. Biomass Quantification

Culture growth was monitored over a period of nearly 30 days, with more frequent sampling after day 10. At each sampling point, every flask was analysed independently. Because only one flask was available for each EM formulation–dilution combination, biological variability within the individual EM-treated conditions could not be estimated. The three untreated control flasks represented genuine biological replicates and were summarised using the mean and standard deviation.

L. platensis growth was evaluated by monitoring two key parameters, the optical density (OD) and the dry weight (DW). The OD of the culture at 565 nm was measured using a spectrophotometer (model ONDA V30 SCAN—UV VIS, ZetaLab, Padua, Italy).

Dry-weight biomass was obtained from OD_{565} using a linear calibration established from paired optical-density and gravimetric (membrane-filtration) dry-weight measurements. Over the calibrated range (OD_{565} = 0.14–1.16; dry weight = 0.14–0.92 g L^{-1}; n = 16 paired points), the relationship was DW (g L^{-1}) = 0.776 × OD_{565} + 0.017, with R^2 = 0.995 and a residual RMSE ≈ 0.02 g L^{-1}. Biomass estimates were confined to this calibrated OD range. The same calibration was applied to both media; biomass values were computed directly from OD_{565} using this relationship. Dry biomass concentration was evaluated gravimetrically as follows: (a) a known volume (10 mL) of culture (V) was drawn from the photobioreactors (PBRs); (b) the sample was filtered through a pre-weighed (W1) glass microfiber filter (GF/CTM 55 mm diameter, Whatman, Incofar Srl., Modena, Italy), and the biomass retained on the filter was dried at 105 °C overnight to a constant weight (W2); and (c) the filter paper had been previously dried in a forced-air oven (model 30, Memmert Gmbh, Scwabach, Germany) at 105 °C for 2 h, then cooled to room temperature in a desiccator, and weighed using an analytical scale (model M, Bel Engineering Srl, Monza, MI, Italy).

In compact form, the optical-density-to-dry-weight conversion applied to the two media was as follows:

$$DW = k \times OD_{565}, k = 0.776 \text{ (both media)} \quad (1)$$

where DW is the dry-weight biomass (g L^{-1}), OD_{565} is the optical density at 565 nm, and k is the medium-specific calibration factor.

The final biomass concentration (dry weight), X_f (g L^{-1}), was calculated as follows:

$$X_f = (W_2 - W_1)/V \quad (2)$$

where W_1 and W_2 are the filter weights before and after drying and V is the filtered culture volume.

Biological observations on *L. platensis* behaviour directly informed the design of the predictive modelling framework. EM identity (δ) and EM dilution level were incorporated as categorical features to capture potential microbial influences on growth dynamics.

To characterise growth dynamics, the specific growth rate between two consecutive sampling days was calculated from the natural logarithm of biomass as follows:

$$\mu_i = [\ln X_i - \ln X_{i-1}]/(d_i - d_{i-1}) \quad (3)$$

2.4. Predictive Modelling with Random Forest and Gradient Boosting

Two ensemble-learning families, random forest (RF) and gradient boosting (GBM), were employed to model biomass accumulation under EM treatments. RF aggregates independent decision trees grown on bootstrap samples [26–28], offering noise robustness and low overfitting risk. GBM constructs trees sequentially, each correcting residual errors of the previous ensemble [29,30], and benefits from advanced implementations such as XGBoost, which include regularisation and learning-rate optimisation [31,32]. Hyperparameters and the model class were selected exclusively on the training data (by a temporal validation split inside the training period for the forecasting task, and by grouped cross-validation for the leave-one-treatment-out task), and the chosen configuration was then evaluated only once on the held-out test set, with no further selection. The frozen final configuration was a gradient-boosting regressor with 500 estimators, a learning rate of 0.05, a maximum depth of 4 and a subsample of 0.7, combined with a horizon-aware forecast rule: at a one-day horizon, the prediction defaults to the most recent observation (persistence), while the gradient-boosting model is used at all longer horizons. The crossover horizon was treated as a further hyperparameter and selected on the training data alone; two independent schemes (grouped cross-validation by treatment and a temporal validation split inside the training period) both selected a one-day threshold, after which the configuration was evaluated only once on the held-out test set. For completeness, regularised linear regression (Ridge), extremely randomised trees, histogram gradient boosting, LightGBM and CatBoost were also evaluated; higher-capacity boosters gave broadly comparable held-out performance and were not selected by the training-only model-selection protocol (Table S1). All models were implemented with the scikit-learn library [33]. A comparative summary of the two approaches, including strengths and limitations for biological datasets, is presented in Table 2.

Table 2. Comparison of the two ensemble model families.

Model	Strengths	Weaknesses
Random Forest (RF)	Robust to experimental noise; low overfitting risk via bootstrap aggregation; minimal tuning; handles nonlinear interactions	Less effective at subtle temporal patterns; performance plateaus with small datasets
Gradient Boosting (GBM/XGBoost)	Strong accuracy on nonlinear, high-interaction data; built-in regularisation aids generalisation; flexible for complex trajectories	Requires careful tuning; more sensitive to noise if poorly regularised; higher computational cost

This modelling work builds on recent applications of ML in microalgal systems, including bloom forecasting [14], *L. platensis* optimisation [18], hybrid mechanistic–ML models [34,35], dynamic population modelling [36], machine-learning prediction of thermochemical biomass valorisation [37], and advanced ensemble frameworks [15–17,19,23,38].

2.5. Feature Engineering for Temporal Biomass Prediction

Feature engineering was designed to translate raw biomass measurements into biologically meaningful predictors that capture the nonlinear, time-dependent nature of *L. platensis* growth [20–22]. A guiding constraint, central to this revision, is that every predictor is computed using only information available strictly before the value being predicted; no feature is a function of the target biomass itself. Each observation X_i was mapped into a feature vector containing both static and time-dependent variables.

The first component, previous biomass, X_{i-1}, encodes biomass inertia. The second component, the recent growth rate, was computed from past values only as

$\Delta X_i = X_{i-1} - X_{i-2}$, together with the slope from the prior step. A third component represented the biomass regime, ϕ_i, assigning the most recent past biomass observation to a low-, intermediate-, or high-biomass range. To remain strictly prospective, the phase label was derived from the most recent past biomass X_{i-1} rather than from the value being predicted. The fourth component was the EM treatment identity δ_i, encoded as control, EM1, EM2, and EM3. A fifth component encoded the EM dilution level (control, $1{:}10^2$, $1{:}10^3$ or $1{:}10^5$) as a discrete treatment variable. No intermediate $1{:}10^4$ condition was inferred or generated, and the model was not used to estimate a continuous dose–response relationship. This fifth component acted as an experimental variable not previously exploited as a predictor, and a sixth encoded the culture medium (Jordan or Zarrouk). Finally, a logarithmic temporal feature $\tau_i = \log(1 + d_i)$ captured diminishing growth increments over time; the forecast horizon (target day minus origin day) was additionally supplied so that a single model serves all horizons.

These components were combined into a single feature vector for each observation:

$$z_i = [x_{i-1}, x_{i-1} - x_{i-2}, \varphi(x_{i-1}), \delta, \rho, \kappa, \log(1 + d_i), h] \quad (4)$$

where x_{i-1} is the previous biomass, $x_{i-1} - x_{i-2}$ is the recent growth increment, $\varphi(x_{i-1})$ is the growth-phase descriptor, δ is the EM treatment identity, ρ is the EM dilution level, κ is the culture medium, d_i is the cultivation day, and h is the forecast horizon.

The categorical growth-phase descriptor $\varphi(\cdot)$ was obtained by thresholding the most recent previous biomass value:

$$\varphi(x) = 0 \text{ if } x < 0.5;\ 1 \text{ if } 0.5 \leq x < 1.2;\ 2 \text{ if } x \geq 1.2 \text{ (g L}^{-1}) \quad (5)$$

where levels 0, 1, and 2 represent low-, intermediate-, and high-biomass regimes, respectively. The cut-offs of 0.5 and 1.2 g L^{-1} were introduced as dataset-specific operational boundaries to partition the biomass range observed in the training trajectories and were not treated as universal physiological thresholds. They were defined before evaluation on the held-out test observations and were not selected using test-set performance. Because the growth phase cannot be determined conclusively from biomass concentration alone, this feature is referred to as a biomass-regime descriptor rather than a direct biological classification of lag, exponential, and stationary phases.

Defining all predictors using information available strictly before the forecast target avoids look-ahead bias and keeps the evaluation representative of real deployment, consistent with established recommendations for time-series machine learning [24,25].

Table 3 summarises the engineered feature set, together with the role of each predictor in the model.

Table 3. The engineered feature set used by the forecasting model. Every predictor is derived from information available strictly before the day being forecasted.

Feature	Definition	Role in the Model
Previous biomass (x_{i-1})	Dry-weight biomass at the previous sampling day	Biomass inertia/ autocorrelation
Recent growth rate ($x_{i-1} - x_{i-2}$)	Change over the two preceding sampling steps	Short-term growth velocity
Biomass regime φ (x_{i-1})	Low, intermediate, or high biomass range inferred from the last past value	Coarse position within the observed biomass trajectory
EM treatment	Control, EM1, EM2 or EM3	Qualitative consortium effect
EM dilution level	$1{:}10^2$, $1{:}10^3$ or $1{:}10^5$ series	Dose-resolved EM signal (novel predictor)

Table 3. *Cont.*

Feature	Definition	Role in the Model
Culture medium	Jordan or Zarrouk	Encodes medium-dependent differences during joint training
Log cultivation time	Log $(1 + d_i)$	Diminishing growth increments over time
Forecast horizon	Target day minus origin day	Lets a single model serve all horizons

2.6. Time-Series Structure, Forecasting Tasks and Validation Design

Because biomass accumulation is inherently temporal and strongly autocorrelated, care was taken to respect chronological order and avoid look-ahead bias between training and testing [24,25]. As illustrated in Figure 1, observations were ordered by cultivation day d_i, and the dataset was partitioned with a blocked temporal split rather than a random split.

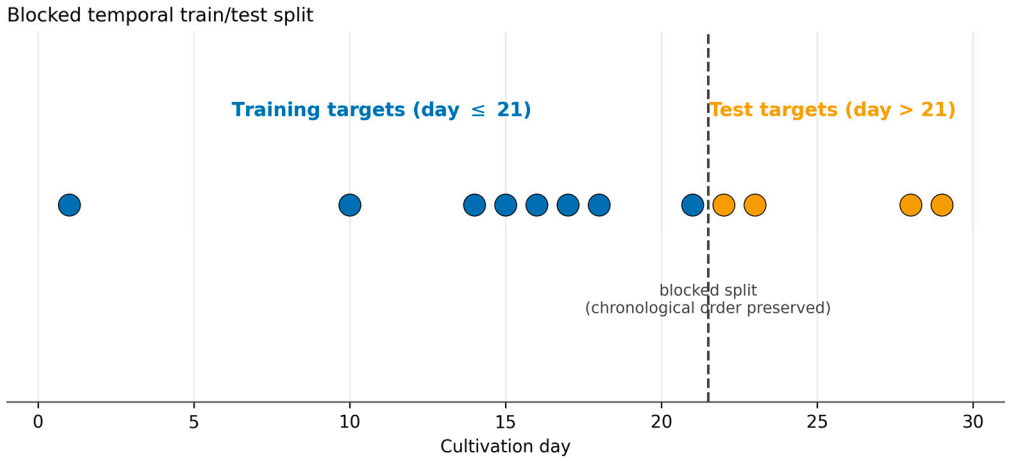

Figure 1. Blocked temporal train/test split. Training targets fall on or before day 21 (blue) and test targets after day 21 (orange); the dashed line marks the boundary. Observations are kept in chronological order across the split, so that every prediction relies only on information available before the day being forecasted.

Formally, the blocked split required every training target day to precede every test target day:

$$\max\{d{:}d \in D_{train}\} < \min\{d{:}d \in D_{test}\} \quad (6)$$

where D_{train} and D_{test} are the sets of target days assigned to the training and test partitions. This blocked design differs from a rolling-origin (expanding-window) evaluation, in which the forecast origin is advanced repeatedly and the model is refitted over successive folds, and from purged or embargoed cross-validation, in which observations adjacent to each fold boundary are discarded to prevent leakage through overlapping feature windows. The one-day exclusion gap enforced between training and validation subsets (Figure 1) embeds the same protective principle as an embargo.

Two forecasting tasks are distinguished. In one-step nowcasting, the biomass at the next sampling time is predicted from the present state. In multi-horizon forecasting, the biomass at a future day is predicted, with the horizon supplied as an input feature so that one model serves all horizons. The deployment scenario of interest is the monitoring of a culture whose early trajectory has been observed; accordingly, the primary evaluation trains on all (origin, target) pairs whose target falls on or before day 21 and tests on pairs whose target falls after day 21. To probe the harder problem of generalising to a never-seen treatment, a leave-one-treatment-out (LOTO) protocol was additionally used, holding out

all pairs of one trajectory and predicting them from a model trained on the remaining nine. Three non-ML baselines were evaluated on identical targets: persistence (carry the last observed value forward), linear drift (extrapolate the recent slope over the horizon), and Gompertz and logistic kinetic models [39], fitted to each trajectory up to the origin and extrapolated.

These baselines are defined as follows. The persistence rule propagates the last observed value forward to the forecast day:

$$\hat{X}(d_t) = X(d_o) \tag{7}$$

where d_t is the target (forecast) day and d_o is the origin day at which the last value $X(d_o)$ is observed.

The Gompertz and logistic kinetic models, fitted to each trajectory and extrapolated, are given respectively by:

$$X(t) = A \cdot \exp\{-\exp[(\mu_m\, e/A)(\lambda - t) + 1]\} \tag{8}$$

$$X(t) = A/\{1 + \exp[(4\,\mu_m/A)(\lambda - t) + 2]\} \tag{9}$$

where X(t) is the biomass at cultivation time t, A is the asymptotic (maximum) biomass, μ_m is the maximum specific growth rate, λ is the lag time, and e is Euler's number.

To further keep lag-based features chronologically separated, a one-day exclusion gap was enforced between the training and validation subsets in a modified *TimeSeriesSplit*. All performance metrics reported below are computed on real held-out observations only.

The unit of prediction in this analysis was an origin–target forecast pair, whereas the number of independent experimental trajectories was substantially smaller. The primary held-out evaluation comprised 340 forecast pairs generated from ten modelled trajectories. Because each trajectory contributed multiple overlapping pairs, including predictions of the same target measurement from different earlier origins, these pairs are not statistically independent. Accordingly, R^2, RMSE, and Pearson's r are reported as pooled descriptive measures of forecast accuracy and should not be interpreted as inferential statistics based on 340 independent biological observations. Likewise, the standard deviation across random seeds quantifies variability arising from model fitting and does not represent biological or trajectory-level uncertainty.

3. Results and Discussion

3.1. Growth Dynamics, Inhibitory Effects, and Biological Implications of EM-Assisted Cultivation

The growth dynamics of *L. platensis* were monitored over the 28-day cultivation period to evaluate the influence of the three EM formulations on biomass accumulation and overall physiological performance. Biomass trajectories, expressed as OD_{560} and corresponding gravimetric dry weight, revealed clear treatment- and dose-dependent differences (Figure 2A).

The untreated control cultures attained different absolute biomass ranges in the two media. The Jordan control reached approximately 1.9 g L^{-1} by day 28, whereas the Zarrouk control approached 3.0 g L^{-1} by day 24, indicating a comparatively higher biomass ceiling in Zarrouk medium. This difference is consistent with the distinct nutrient composition and greater inorganic-carbon availability of the Zarrouk formulation. Nevertheless, because the two experiments followed different sampling schedules and were not designed as a formal factorial comparison of culture media, this observation should not be interpreted as a statistically established medium effect.

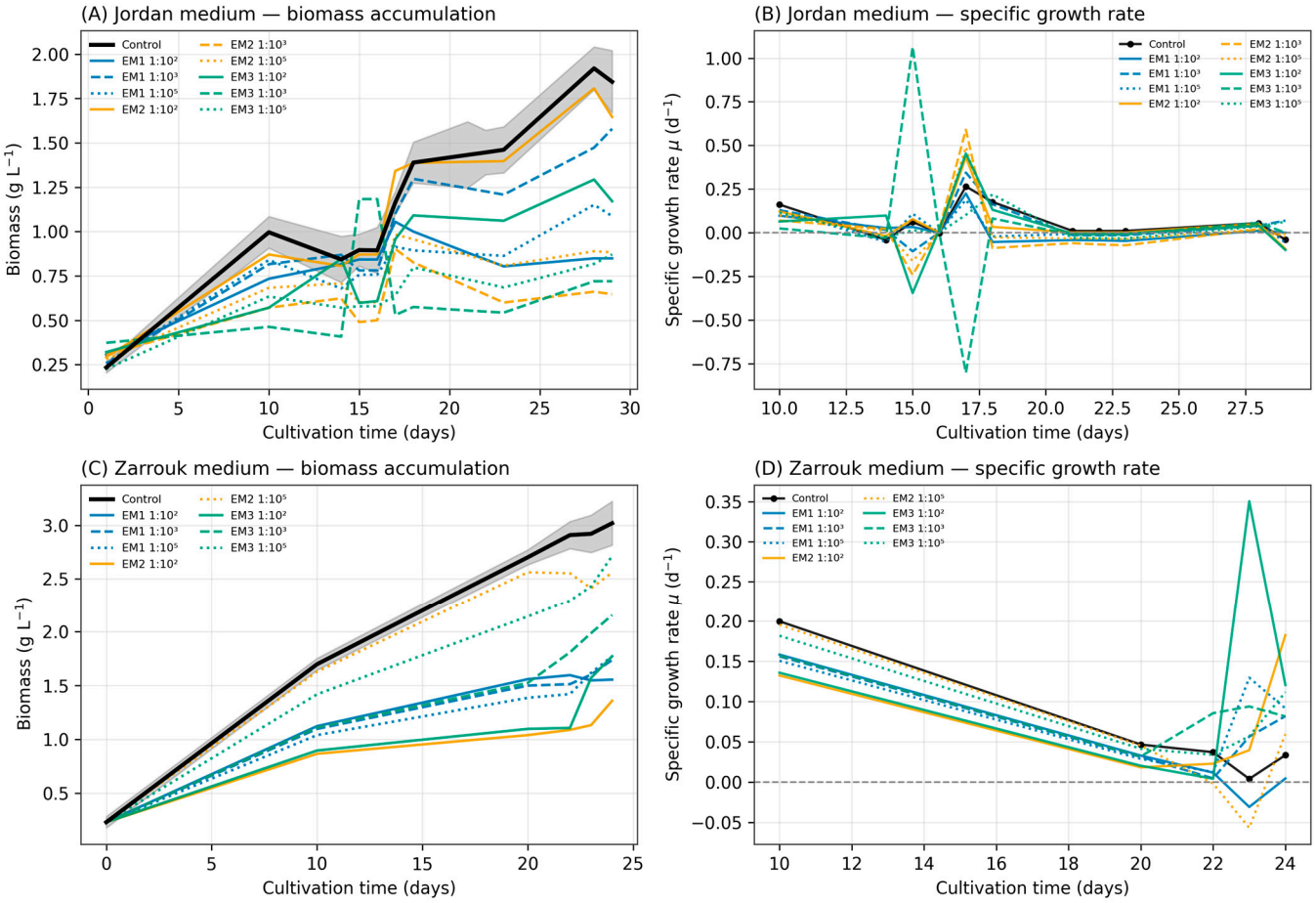

Figure 2. (**A**) Dry-weight accumulation (g L^{-1}) of *L. platensis* cultures in the Jordan medium over the cultivation period under four conditions: control (no EM), EM1, EM2 and EM3. For each EM formulation, the three lines represent the distinct dilution levels 1:10^2, 1:10^3, and 1:10^5; these series are separate experimental conditions and not biological replicates. The control is shown as mean ± standard deviation across three independent biological replicate flasks. No replicate-based error estimate is available for the individual EM formulation–dilution conditions. (**B**) Specific growth rate (μ, d^{-1}) in Jordan medium, computed as $\mu = [\ln X(t) - \ln X(t-1)]/(t - t - 1)$; the dashed horizontal line at μ = 0 separates net biomass growth from stagnation or biomass loss. (**C**,**D**) Corresponding dry-weight accumulation and specific growth rate profiles for the parallel experiment in the Zarrouk medium, sampled at days 0, 10, 20, 22, 23 and 24.

Control cultures maintained comparatively high and stable μ values during the early and intermediate phases of cultivation, consistent with efficient assimilation of bicarbonate and nitrogen sources characteristic of *L. platensis*. Conversely, EM-treated cultures exhibited more irregular growth-rate profiles. EM1 and EM2 were associated with lower growth-rate maxima and shorter exponential-growth phases, while EM3 frequently produced μ values that approached zero or became negative during the later stages of cultivation, indicating growth stagnation or net biomass loss. These responses may result from multiple interacting mechanisms, including microbial competition for nutrients or micronutrients, EM-induced changes in pH and water chemistry, and the accumulation of organic metabolites or extracellular compounds capable of indirectly constraining cyanobacterial growth or photosynthetic performance. Such mechanisms remain hypothetical, however, because nutrient competition, metabolite production, and photosynthetic inhibition were not measured directly in the present study. The observed variability is nevertheless consistent with previous reports showing that interactions between microbial consortia and photosynthetic microorganisms are strongly dependent on consortium composition, cultivation

medium, environmental conditions, and application dose [10–13]. One plausible mechanism involves acidogenic members of the EM consortium. Lactic-acid bacteria can convert available organic substrates into lactic acid, and the resulting decrease in pH depends on microbial activity, substrate availability, and the buffering capacity of the medium [40]. This could be particularly relevant for *L. platensis*, a strongly alkaliphilic cyanobacterium for which maximal biomass production is generally observed under alkaline conditions around pH 9–10, while acidification can rapidly inhibit growth and photosynthetic activity [41,42]. Thus, if EM3 contained a comparatively active acidogenic fraction, especially at the highest EM concentration, organic-acid production could have consumed part of the medium's alkalinity and contributed to the observed growth stagnation. Nevertheless, because culture pH, organic-acid concentrations, and EM community composition were not measured, this explanation remains hypothetical and requires direct experimental verification. In agreement with these biological observations, the predictive modelling analysis indicated that EM consortium identity contributed only modestly to model performance, whereas the EM dilution level emerged as one of the most influential predictors. Therefore, the results do not indicate that EM-related effects were negligible; rather, they suggest that the EM dose carried more biologically relevant information than consortium identity alone.

The greater importance of EM dilution relative to consortium identity may reflect the intensity of the ecological loading imposed on the *L. platensis* culture. The least diluted EM preparations introduced a comparatively larger inoculum of exogenous microorganisms and associated metabolites, potentially producing a stronger transient perturbation through nutrient competition, heterotrophic oxygen demand, changes in water chemistry, or the release of biologically active compounds. By contrast, under the most diluted conditions, the introduced microbial population and metabolite load were likely lower, reducing their capacity to alter the dominant cyanobacterial growth dynamics. Dilution may therefore act as an integrative proxy for the magnitude of the biological disturbance, whereas EM identity captures only qualitative differences among consortia of incompletely characterised compositions. Nevertheless, this result should be interpreted as a dose-dependent ecological loading gradient rather than a confirmed threshold, because microbial abundance, dissolved oxygen, organic-carbon loading, and consortium persistence were not directly measured.

Treating the different dilution series as conventional experimental replicates would consequently obscure systematic dose-dependent responses. Overall, EM3 exerted the strongest inhibitory effect, although none of the tested formulations caused complete culture collapse during the 28-day experimental period. Because the EM formulations were not taxonomically or quantitatively characterised, differences among EM1, EM2, and EM3 cannot be mechanistically assigned to particular microbial groups. The present findings should therefore be interpreted as formulation- and batch-specific responses, and future studies should incorporate viable counts, taxon-resolved profiling, and relevant physicochemical characterisation of the inocula. These findings support a phase- and dose-aware interpretation of EM-assisted cultivation, in which the effects of EM formulations may vary throughout the growth cycle and according to the applied concentration. Furthermore, because the experiments were conducted in 50 mL flasks under controlled light and temperature conditions, extrapolation to industrial PBRs, where cultures experience spatial and temporal gradients in light, pH, temperature, gas transfer, and mixing, should be regarded as a hypothesis requiring direct validation rather than as an established operational recommendation.

3.2. Multi-Horizon Forecasting Across Two Cultivation Media

A single horizon-aware gradient-boosting model, trained on the Jordan and Zarrouk experiments jointly and predicting across horizons of 1–28 days, achieved a held-out R^2 of 0.72 (0.72 ± 0.02 over 30 random seeds; RMSE ≈ 0.20 g L^{-1}, Pearson r ≈ 0.85) on the real Jordan test points, against 0.16 for a persistence baseline on the same targets (Figures 3 and 4, Table 4). These metrics were calculated across 340 held-out origin–target pairs derived from ten modelled trajectories. Because multiple forecast pairs originated from the same experimental trajectory and could share origin or target observations, the effective number of independent experimental units was considerably lower than 340. The reported metrics should therefore be interpreted as pooled predictive-performance summaries for this dataset. Reporting over 30 seeds provides an explicit measure of estimation uncertainty; including the second medium in joint training did not materially change held-out Jordan performance (see below). Before joint modelling, OD measurements were converted to dry-weight biomass using the single OD_{565}-to-dry-weight calibration described in Section 2.3, thereby avoiding systematic scale bias associated with the different optical properties of the Jordan and Zarrouk media. The horizon-aware tree ensemble accommodated the distinct biomass plateaus through the combined use of current biomass, lagged biomass dynamics, cultivation time, forecast horizon, and culture-medium identity. Current and lagged biomass encoded the position of each culture relative to its trajectory-specific saturation region, while the categorical medium feature allowed nonlinear medium-dependent interactions to be learned. The low permutation importance assigned to culture medium therefore does not indicate biologically equivalent growth in the two formulations. Rather, it suggests that medium identity added relatively little predictive information once the observed biomass history was available. Adding the correctly labelled Zarrouk experiment to joint training left held-out Jordan performance essentially unchanged (R^2 ≈ 0.72 with and without the second medium), indicating no material benefit from the second medium under the present design, consistent with the absence of a held-out cross-medium test. Resolving performance by horizon is informative and motivates the model's design. At short horizons (1–4 days), persistence is excellent (R^2 ≈ 0.97), so the forecaster defers to the last observation at a one-day horizon; at medium horizons (5–10 days), persistence degrades but remains positive (R^2 ≈ 0.43), and only at long horizons (11–28 days) does it collapse to negative skill (R^2 ≈ −0.36), whereas gradient boosting remains stable (R^2 ≈ 0.75 and 0.68). Combining the two (persistence at the shortest horizon and gradient boosting beyond it) yields a single forecaster that is competitive at short range (R^2 ≈ 0.78) and robust at medium-to-long range, where simple extrapolation fails. The Gompertz and logistic kinetic models, although mechanistically appealing, were unstable when extrapolated from limited early data and did not match the ensemble at medium-to-long horizons.

Table 4. Performance summary (held-out Jordan test; real held-out data; ML values seed-averaged over 30 seeds).

Task/Method	R^2	RMSE (g L^{-1})	Pearson r	*n*
Persistence (multi-horizon)	0.16	0.35	0.64	340
Linear drift	<0	1.43	0.19	340
Gompertz (kinetic)	−0.07	0.39	0.55	340
Horizon-aware GBM, Jordan only	0.72 ± 0.01	0.21	0.87	340
Horizon-aware GBM, Jordan + Zarrouk	0.72 ± 0.02	0.20	0.85	340
Short/medium/long	0.77/0.75/0.68	-	-	-
GBM, LOTO (unseen treatment)	0.33	-	-	550

Note: *n* denotes the number of evaluated origin–target forecast pairs, not the number of independent biological trajectories. The 340 pairs in the primary evaluation were derived from ten modelled trajectories and are correlated within trajectories.

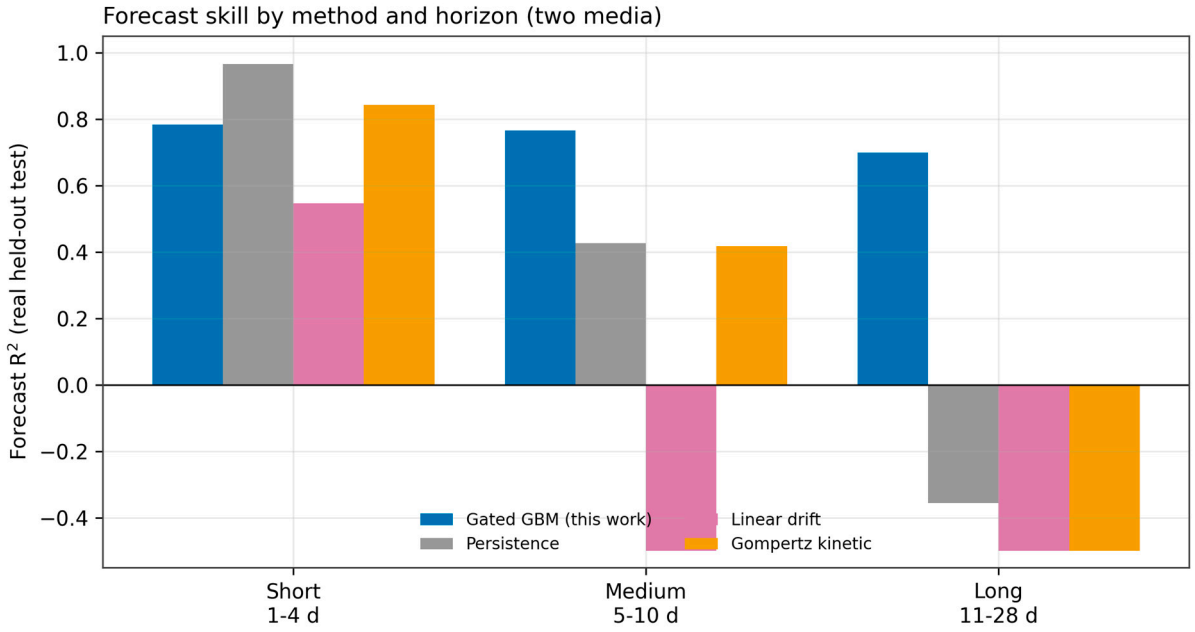

Figure 3. Forecast R^2 by method and horizon (blocked temporal test on the Jordan series, real data only; model trained on both media). Persistence is excellent at short range but collapses to negative skill at long range, whereas the horizon-aware gradient-boosting forecaster of this work remains stable across all horizons. Linear drift and the kinetic baselines are outperformed at medium-to-long horizons.

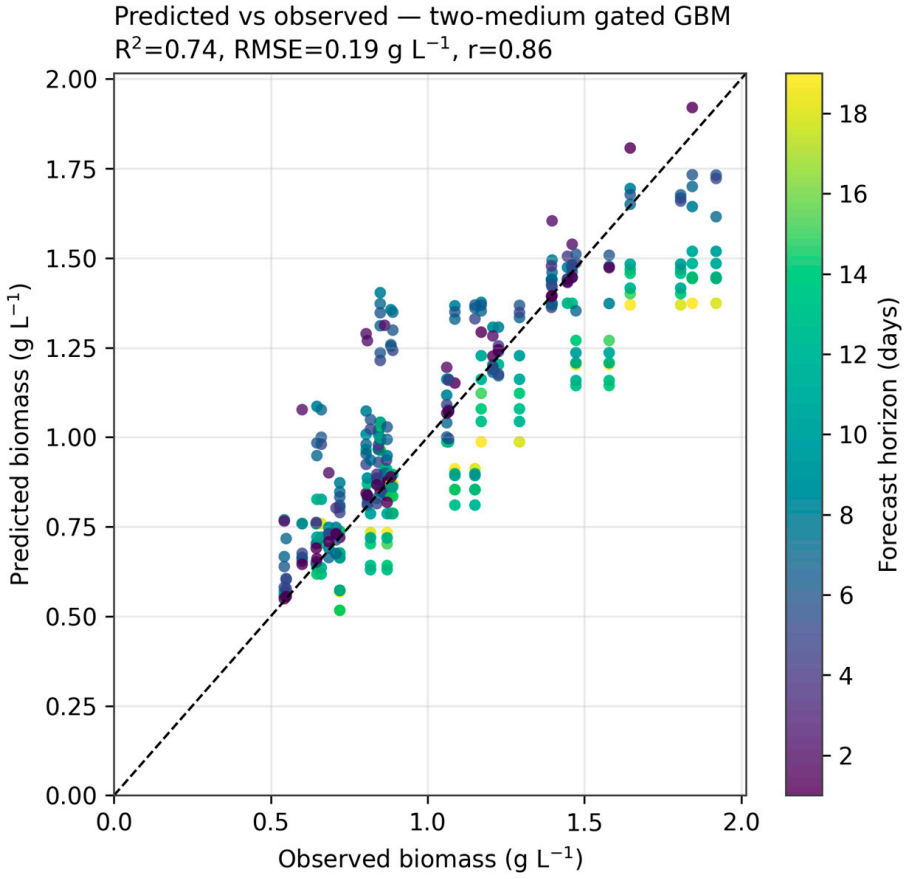

Figure 4. Observed versus forecasted biomass for the two-medium horizon-aware gradient-boosting model (seed-averaged) on the held-out Jordan test, coloured by forecast horizon (R^2 = 0.72, RMSE ≈ 0.20 g L^{-1}).

3.3. Feature Importance: Biomass, Dilution and Medium

Permutation importance for the two-medium gradient-boosting model (Figure 5), computed on held-out data, is dominated by the current biomass, with the EM dilution level and the forecast horizon among the next most influential predictors. To our knowledge, this is the first study to use EM dilution as an explicit predictor and to demonstrate its informativeness for this system, thereby addressing the conflation of dilution with treatment identity in earlier analyses. The culture medium and EM identity contribute at lower levels, and the recent lagged biomass adds incremental information. Because these importances are permutation-based and computed on held-out data, they are not subject to the impurity-based bias that can spuriously inflate correlated or high-cardinality predictors. Because these importances are computed over the whole held-out set, they are horizon-averaged: as the forecast horizon lengthens, the informativeness of the most recent observation necessarily decays. Stratifying the analysis by horizon band (Figure S1, Supplementary Materials) makes this explicit, with the importance of current biomass declining from short to long horizons while that of the EM dilution level increases, becoming the second most influential predictor at the longest horizons.

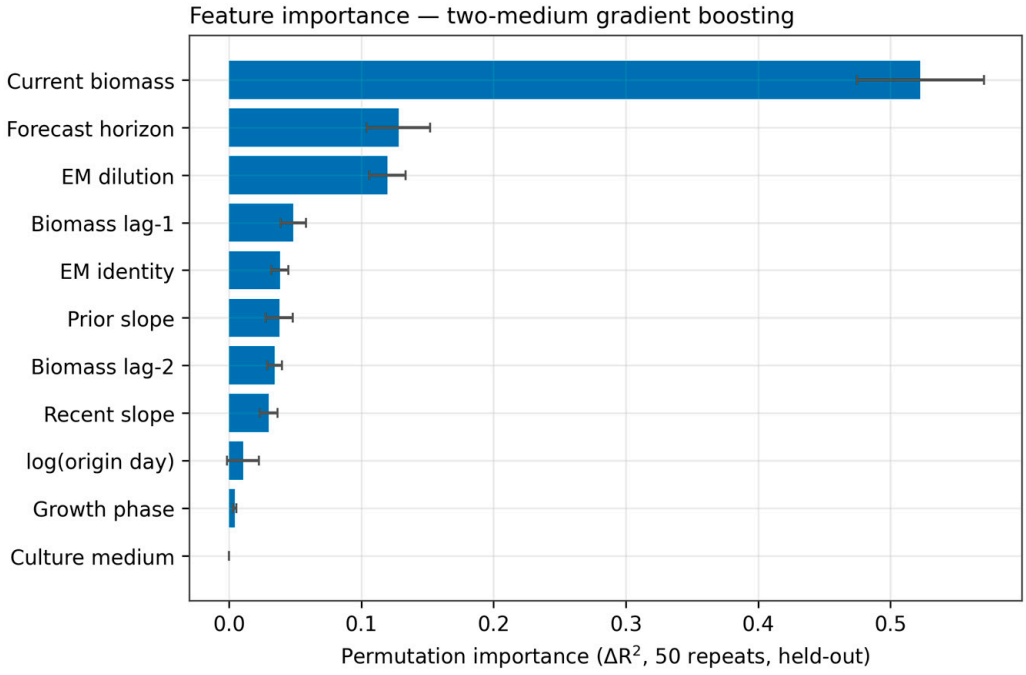

Figure 5. Permutation importance (ΔR^2, 50 repeats, held-out data) for the two-medium gradient-boosting model. Current biomass, EM dilution, and forecast horizon are the leading predictors; error bars show the standard deviation across permutations.

3.4. Generalisation to Unseen Treatments

Generalising to an entirely unseen treatment is substantially harder. Under the LOTO protocol, the pooled R^2 falls to about 0.33, and the regularised linear model is competitive with or better than the tree ensembles, a reminder that, on very small datasets, higher-capacity models do not automatically generalise better. The contrast between the monitoring task (where the early trajectory of the target culture is available and ensembles excel) and the LOTO task (where it is not) delimits the realistic scope of the model: it forecasts the continuation of a monitored culture well, but should not be expected to predict a novel EM formulation without further data. This pattern has a straightforward architectural explanation. Random forests and gradient boosting are piecewise-constant predictors: within each terminal leaf they return an average of the training targets, and they

therefore cannot extrapolate beyond the range of target values encountered during training. When a held-out treatment follows a trajectory or reaches biomass levels absent from the training leaves, the ensemble can only return the nearest observed average, whereas a regularised linear model fits a continuous slope and projects it forward. This is why Ridge regression attains a pooled R^2 of about 0.33 under the LOTO protocol while the tree ensembles collapse (Extra Trees ≈ 0.06, random forest ≈ -0.02, gradient boosting ≈ 0.07). A related identifiability problem explains the poor performance of the sigmoidal kinetic baselines: the Gompertz and logistic models require joint estimation of the asymptotic biomass, the maximum specific growth rate and the lag time, and when they are fitted only to the sparse early portion of a trajectory the asymptote is essentially unidentifiable. The fit then becomes ill-conditioned, so small perturbations in the early observations produce large swings in the extrapolated curve, and the problem is aggravated under inhibitory conditions in which the trajectory departs from the assumed monotonic sigmoidal shape.

3.5. Originality and Practical Utility

Three aspects distinguish this work from earlier machine-learning studies of microalgal cultivation. First, to our knowledge, it is the first leakage-free, multi-horizon forecasting benchmark for *L. platensis* under EM treatment, with every feature computed before the prediction target and every metric reported on real held-out data; this yields an honest estimate of the accuracy attainable on small bioprocess datasets, rather than the optimistic values that arise when target-derived features inflate evaluation. Second, the EM dilution level (previously conflated with consortium identity) emerges as a genuinely informative predictor, supporting a dose-resolved rather than a purely qualitative view of EM effects. Third, joint training across two media (Jordan and Zarrouk) showed that including an independent second-medium dataset did not materially change held-out Jordan forecasting. Generalisation to a completely unseen cultivation medium remains to be evaluated, and no held-out cross-medium test was performed. A further limitation is the small number of independent experimental trajectories underlying the pooled forecast metrics. Although the origin–target construction provides many evaluation cases across different forecast horizons, it does not create additional independent biological units. Confirmation of the reported performance will therefore require additional independently cultivated batches and trajectory-level validation.

In practical terms, the framework is directly useful for the monitoring of ongoing cultures: given an observed early trajectory, it forecasts biomass up to four weeks ahead with bounded error (Figure 6), which can inform harvest scheduling, resource planning, and the early detection of stalled or inhibited cultures. Equally important for practitioners, the study makes explicit the regime in which data-driven models add value (medium-to-long horizons) and the regime in which a trivial persistence rule already suffices (very short horizons), so that modelling effort can be directed where it pays off. More broadly, the validation protocol (past-only features, real-data evaluation, persistence and kinetic baselines, and performance resolved by forecast horizon) provides a transferable template for trustworthy machine learning on the small, noisy datasets that are typical of controlled bioprocess experiments.

Figure 6 illustrates this directly: for three representative held-out cultures spanning the biomass range, the multi-horizon forecast issued at the end of the training window tracks the subsequent real observations, and the spread across 30 random seeds provides an explicit and notably narrow uncertainty band.

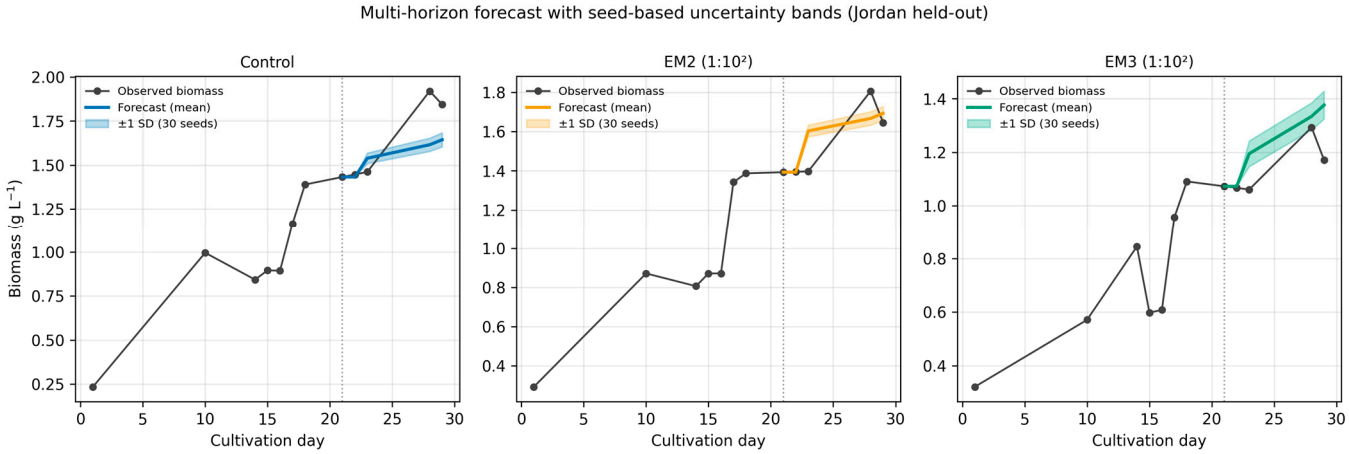

Figure 6. Multi-horizon forecast with uncertainty for three representative Jordan held-out cultures (control, and EM2 and EM3 at the 1:10^2 dilution). Grey markers are the observed biomass; the coloured line is the seed-averaged forecast issued from the last training observation (dotted line, day 21) for all subsequent days; the shaded band spans ± 1 standard deviation of the prediction across 30 random seeds. The band is narrow ($\approx$0.02 g L^{-1}), reflecting the stability of the ensemble, whereas the gap between the line and markers reflects genuine forecast error on unseen days.

4. Conclusions

Working from two small EM-treated *Limnospira platensis* experiments (Jordan and Zarrouk media), this study shows that honest, strictly prospective forecasting of microalgal biomass is feasible and useful at the accuracy the data genuinely support. For multi-horizon forecasting of a monitored culture, a single horizon-aware gradient-boosting model delivers stable performance ($R^2 \approx 0.72$, RMSE $\approx$ 0.20 g L^{-1}) across horizons of up to four weeks, outperforming a persistence baseline more than fourfold, precisely where simple extrapolation fails; at short horizons, persistence is already adequate and is adopted directly by the forecaster. The EM dilution level was identified as an informative and previously overlooked predictor. Adding the Zarrouk observations to the Jordan training data produced essentially no change on the held-out Jordan targets, with R^2 from 0.72 ± 0.01 to 0.72 ± 0.02 and RMSE remaining close to 0.20 g L^{-1}. This indicates no material benefit from adding the second-medium observations under the present experimental design, consistent with the absence of a held-out cross-medium evaluation. It should not be interpreted as validated transfer to an unseen medium, because no held-out Zarrouk or leave-one-medium-out evaluation was performed.

Beyond the specific system, the study offers a transferable template for ML-based bioprocess forecasting under data scarcity: use only information available at prediction time; evaluate only on real held-out data; benchmark against persistence and a kinetic baseline; match the validation protocol to the deployment scenario; and report performance resolved by forecast horizon together with seed-level uncertainty. Future work should pursue external validation on independent cultivation batches and incorporate continuous pH and optical pigment measurements as dynamic-state variables. Online pH is particularly relevant because it integrates inorganic-carbon assimilation, carbonate-system dynamics, gas exchange, and possible acidogenic activity by EM-associated microorganisms; it could therefore provide the direct mechanistic evidence needed to test whether changes in medium chemistry have contributed to the observed biomass suppression. In parallel, optical chlorophyll- or phycocyanin-related signals may detect changes in photosynthetic status and cellular stress before they are reflected in intermittently measured dry-weight biomass. From a modelling perspective, these high-frequency signals could serve as leading indicators of metabolic transitions, improve phase identification, and support earlier

forecasting of growth inhibition or culture failure. The framework should ultimately be extended to dynamically controlled photobioreactors, where such state-aware forecasting would be most valuable for harvest scheduling and early fault detection.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/pr14162536/s1, Figure S1: Permutation importance (ΔR^2, 50 repeats) computed on the held-out Jordan test set and stratified by forecast-horizon band (short, 1–4 days; medium, 5–10 days; long, 11–28 days). The importance of the current biomass declines as the horizon lengthens, whereas that of the EM dilution level increases, becoming the second most influential predictor at the longest horizons. This stratification complements Figure 5 of the main text, in which the importances are pooled over all horizons and are therefore horizon-averaged; Table S1. Held-out performance of the additional model classes evaluated during model selection, reported on the same held-out Jordan test set (340 origin–target pairs; mean over 30 random seeds). Model selection was performed on the training data alone, so no model was chosen on the basis of test-set performance. The scores refer to the model classes as such; the deployed configuration adds the one-day persistence gate, giving $R^2 \approx 0.72$.

Author Contributions: Conceptualization, B.C., A.C., G.A.L., R.S.-R. and C.V.P.; Data Curation and Writing—Original Draft Preparation, B.C., A.C. and G.A.L.; Methodology, D.S., G.C., G.D., A.M., M.V., A.P. and B.C.; Investigation, R.M., R.A. and L.U.; Software and Validation, B.C.; Writing—Reviewing and Editing, B.C., A.C. and G.A.L.; Supervision, A.C., G.A.L., R.S.-R. and C.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3—Call for tender No. 1561 of 11.10.2022 of Ministero dell'Università e della Ricerca (MUR); it was also funded by the European Union—Next Generation EU: Award Number: Project code PE0000021, Concession Decree No. 1561 of 11.10.2022 adopted by Ministero dell'Università e della Ricerca (MUR), CUP—I53C22001450006, Project title "Network 4 Energy Sustainable Transition—NEST".

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this work, the authors used ChatGPT Plus in order to rephrase some written paragraphs, with the sole aim of making the whole manuscript more fluent and readable. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Conflicts of Interest: Authors Riccardo Allodi, Riccardo Allodi, Luca Usai and Giovanni Antonio Lutzu were employed by the company Teregroup Srl. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Otálora, P.; Guzmán, J.L.; Acién, F.G.; Berenguel, M.; Reul, A. An artificial intelligence approach for identification of microalgae cultures. *New Biotechnol.* **2023**, *77*, 58–67. [CrossRef]
2. 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]
3. Wu, Y.; Shan, L.; Zhao, W.; Lu, X. Harnessing artificial intelligence to revolutionize microalgae biotechnology: Unlocking sustainable solutions for high-value compounds and carbon neutrality. *Mar. Drugs* **2025**, *23*, 184. [CrossRef]
4. Cosenza, A.; Lima, S.; Gurreri, L.; Mancini, G.; Scargiali, F. Microalgae in the Mediterranean area: A geographical survey outlining the diversity and technological potential. *Algal Res.* **2024**, *82*, 103669. [CrossRef]
5. Xu, K.; Zhu, Z.; Yu, H.; Zou, X. Prediction of microalgae harvesting efficiency and identification of important parameters for ballasted flotation using an optimized machine learning model. *Algal Res.* **2025**, *87*, 103985. [CrossRef]
6. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]

7. Mayer, J.; Scheid, S.; Widmer, F.; Fließbach, A.; Oberholzer, H.R. How effective are 'effective microorganisms' (EM)? *Appl. Soil Ecol.* **2010**, *46*, 230–239. [CrossRef]
8. Olle, M.; Williams, I.H. Effective microorganisms and their influence on vegetable production: A review. *J. Hort. Sci. Biotechnol.* **2013**, *88*, 380–386. [CrossRef]
9. Safwat, S.M.; Matta, M.E. Environmental applications of effective microorganisms: A review. *J. Eng. Appl. Sci.* **2021**, *68*, 48. [CrossRef]
10. Lananan, F.; Abdul Hamid, S.H.; Wan Din, W.N.S.; Ali, N.; Khatoon, H.; Jusoh, A.; Endut, A. Symbiotic bioremediation of aquaculture wastewater in reducing ammonia and phosphorus utilizing effective microorganism (EM-1) and microalgae (*Chlorella* sp.). *Int. Biodeterior. Biodegrad.* **2014**, *95*, 127–134. [CrossRef]
11. Dondajewska, R.; Kozak, A.; Rosińska, J.; Gołdyn, R. Water quality and phytoplankton structure changes under the influence of effective microorganisms (EM) and barley straw: Lake restoration case study. *Sci. Total. Environ.* **2019**, *660*, 1355–1366. [CrossRef] [PubMed]
12. Gao, J.; Shao, N.; Sun, Y.; Nie, Z.; Yang, X.; Dai, F.; Xu, G.; Xu, P. Impact of effective microorganisms and *Chlorella vulgaris* on *Eriocheir sinensis* and water microbiota in ponds experiencing cyanobacterial blooms. *Sustainability* **2023**, *15*, 7362. [CrossRef]
13. Lürling, M.; Tolman, Y.; van Oosterhout, F. Cyanobacteria blooms cannot be controlled by effective microorganisms (EM®) from mud- or Bokashi-balls. *Hydrobiologia* **2010**, *646*, 133–143. [CrossRef]
14. Liu, M.; He, J.; Huang, Y.; Tang, T.; Hu, J.; Xiao, X. Algal bloom forecasting with time-frequency analysis: A hybrid deep learning approach. *Water Res.* **2022**, *219*, 118591. [CrossRef] [PubMed]
15. Mermer, O.; Zhang, E.; Demir, I. A comparative study of ensemble machine learning and explainable AI for predicting harmful algal blooms. *Big Data Cogn. Comput.* **2025**, *9*, 138. [CrossRef]
16. Uguz, S.; Sahin, Y.S.; Kumar, P.; Yang, X.; Anderson, G. Real-time algal monitoring using novel machine learning approaches. *Big Data Cogn. Comput.* **2025**, *9*, 153. [CrossRef]
17. Sundaram, K.M.; Kumar, D.; Lee, J.; Barathi, S.; Rajendran, K. Time series forecasting of microalgae cultivation for a sustainable wastewater treatment. *Process Saf. Environ. Prot.* **2025**, *196*, 106845. [CrossRef]
18. Li, H.; Guo, L.; Chen, L.; Zhang, F.; Wang, W.; Lam, T.K.-Y.; Xia, Y.; Cai, Z. Machine learning-assisted optimisation of food-grade *Spirulina* cultivation in seawater-based media: From laboratory to large-scale production. *J. Environ. Manag.* **2024**, *369*, 122279. [CrossRef] [PubMed]
19. Li, H.; Chen, L.; Zhang, F.; Cai, Z. Graph-learning-based machine learning improves prediction and cultivation of commercial-grade marine microalgae *Porphyridium*. *Bioresour. Technol.* **2025**, *416*, 131728. [CrossRef] [PubMed]
20. Coşgun, A.; Günay, M.E.; Yıldırım, R. Machine learning for algal biofuels: A critical review and perspective for the future. *Green Chem.* **2023**, *25*, 3354–3373. [CrossRef]
21. Syed, T.; Krujatz, F.; Ihadjadene, Y.; Mühlstädt, G.; Hamedi, H.; Mädler, J.; Urbas, L. A review on machine learning approaches for microalgae cultivation systems. *Comput. Biol. Med.* **2024**, *172*, 108248. [CrossRef] [PubMed]
22. Imamoglu, E. Artificial intelligence and/or machine learning algorithms in microalgae bioprocesses. *Bioengineering* **2024**, *11*, 1143. [CrossRef] [PubMed]
23. Fu, Y.; Zhang, Q.; Tan, Z.; Yu, S.; Zhang, Y. Predicting harvesting efficiency of microalgae with magnetic nanoparticles using machine learning models. *J. Environ. Chem. Eng.* **2025**, *13*, 115406. [CrossRef]
24. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
25. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. [CrossRef]
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Liu, Y.; Wang, Y.; Zhang, J. New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 246–252. [CrossRef]
28. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
29. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
30. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef] [PubMed]
31. Ali, Z.A.; Abduljabbar, Z.H.; Tahir, H.A.; Sallow, A.B.; Almufti, S.M. Extreme gradient boosting algorithm with machine learning: A review. *Acad. J. Nawroz Univ.* **2023**, *12*, 320–334. [CrossRef]
32. Zhang, Z.; Zhu, X.; Liu, D. Model of gradient boosting random forest prediction. In *2022 IEEE International Conference on Networking, Sensing and Control (ICNSC), Shanghai, China, 15–18 December 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, A.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

34. Aowtrakool, N.; Sopitthummakhun, A.; Laomettachit, T.; Ruengjitchatchawalya, M. A dynamic model for predicting biomass and phycocyanin yields in *Arthrospira (Spirulina) platensis*: A guidance for effective batch cultivation. *Algal Res.* **2024**, *83*, 103709. [CrossRef]
35. Atzori, F.; Cosenza, B.; Zedda, F.; Sidorowicz, A.; Fais, G.; Cao, G.; Concas, A. Hybrid modeling of photocatalytic contaminant degradation using nanomaterials synthesized with microalgal extracts. *Chem. Eng. Sci.* **2025**, *316*, 121945. [CrossRef]
36. Cosenza, B.; Miccio, M.; Pannocchia, G.; Sammarco, I. Advanced simulation model for studying biofuel-producing microalgae populations. *Chem. Eng. Trans.* **2024**, *109*, 559–564. [CrossRef]
37. Cosenza, B.; Picone, A.; Volpe, M.; Messineo, A. Improving hydrothermal carbonization prediction by machine learning: Towards a more accurate and less expensive process modelling through data augmentation. *Energy* **2025**, *333*, 137373. [CrossRef]
38. Khan, A.; Ullah, S.; Ali, R.; Rehman, M.; Moshawih, S.; Goh, K.W.; Ming, L.C.; Gew, L.T. Machine learning optimization of environmental factors influencing biomass and nutritional composition in local algal species. *R. Soc. Open Sci.* **2025**, *12*, 241336. [CrossRef] [PubMed]
39. Zwietering, M.H.; Jongenburger, I.; Rombouts, F.M.; van 't Riet, K. Modeling of the bacterial growth curve. *Appl. Environ. Microbiol.* **1990**, *56*, 1875–1881. [CrossRef] [PubMed]
40. Passos, F.V.; Fleming, H.P.; Ollis, D.F.; Felder, R.M.; McFeeters, R.F. Kinetics and modeling of lactic acid production by *Lactobacillus plantarum*. *Appl. Environ. Microbiol.* **1994**, *60*, 2627–2636. [CrossRef] [PubMed]
41. Ismaiel, M.M.; El-Ayouty, Y.M.; Piercey-Normore, M. Role of pH on antioxidants production by *Spirulina* (*Arthrospira*) *platensis*. *Braz. J. Microbiol.* **2016**, *47*, 298–304. [CrossRef] [PubMed]
42. Liu, W.; Wang, J.; Liu, T. Low pH rather than high CO_2 concentration itself inhibits growth of *Arthrospira*. *Sci. Total Environ.* **2019**, *666*, 572–580. [CrossRef] [PubMed]