Maejo International Journal of
Energy and Environmental Communication
Journal home page: https://ph02.tci-thaijo.org/index.php/MIJEEC

REVIEW ARTICLE

Machine learning and digital twins meet the pond: appraising data-driven modeling of microalgae growth, product quality, and contamination

Santhiya Subramani, Sankarganesh Murugesan*

Department of Chemistry, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, Tamil Nadu 602105, India

ARTICLEINFO

Article history:

Received: 12 September 2025
Revised: 09 October 2025
Accepted: 10 October 2025
Published: 13 October 2025

Keywords:

Microalgae
Machine learning
Digital twin
Soft sensor
Reinforcement learning

Abstract

Microalgae cultivation is limited less by biology than by operation: outdoor cultures face light and temperature that change by the minute, product quality that cannot be measured online, and grazers that can clear a pond in days. Machine learning (ML) and digital twins are promoted as the remedy, yet published accuracy figures are rarely comparable. This review appraises 60 sources published up to 2025 against three questions: which model families predict growth, lipid and contamination; how much and what kind of data they need; and how sound their validation is. Tabular regressors (artificial neural networks, gradient boosting, support vector regression) dominate and report R^2 of 0.87 to 0.98 on laboratory or literature-compiled data, mostly under random hold-out or unstated designs. Sequence models such as long short-term memory networks outperform Monod and Haldane kinetics outdoors because they use light history, and a profile-based deep model reached $R^2 = 0.77$ across 32 raceway ponds in five US states where averaged inputs gave 0.39. Lipid content is estimated non-destructively by near-infrared, hyperspectral and Raman chemometrics, with errors near 4 percentage points of dry weight, but no study reports how a laboratory calibration performs at pilot scale. In laboratory grazing trials, infrared and photosynthetic signals flag contamination 24 to 72 h before a crash, and anomaly detection on redox potential flags it outdoors. Reinforcement learning and adaptive predictive control have run on industrial raceways, but only for pH. The field is scientifically classified into diverse categories, such as data-driven, hybrid modeling methods, and mechanistic methods, along with the framework of offline surrogate to autonomous control. The analysis highlights that robust validation, before the choice of algorithm, remains the key limiting factor. Moreover, a validation scale and standardized reporting checklist are proposed to expand consistency, transparency, and comparability across future studies.

* Corresponding author: msankarajan1990@gmail.com (Murugesan S.)
© 2019. All rights reserved.

1. Introduction

Outdoor microalgae cultivation fails on operation more often than on strain choice. Productivity in open raceway ponds follows light intensity and water temperature that fluctuate through the day [1]. Product quality, especially lipid content, is still measured by extraction methods that take hours to days, so the operator learns the result after the decision it should have informed [2, 3]. Grazers and parasites add a third risk: predator invasions are reported to cost 20 to 30% of biomass in open systems and can clear a culture within 2 to 5 days of detection [4].

Mechanistic models were the first answer. A growth model driven by light and temperature has predicted productivity for three species in outdoor raceways [1]. Such models extrapolate well but need strain-specific parameters, and their classical forms (Monod, Haldane) take instantaneous light as input and so miss photoacclimation [5].

Machine learning (ML) entered the field as a way around both limits. Earlier reviews have cataloged the algorithms used across cultivation, harvesting and conversion [6, 7], their use in algal wastewater treatment

[8], and the concept of an algal digital twin, where the principal barriers named are interconnection, facilities and data acquisition [9]. What these reviews do not provide is a judgment of whether the reported accuracy would survive transfer to another reactor, season or site. A database study that pooled 4,670 instances from 102 articles shows how heterogeneous the underlying evidence is [10].

This review therefore asks three questions.

RQ1. Which ML model families (artificial neural networks, tree ensembles such as random forest, recurrent networks such as LSTM, and reinforcement learning) predict or control growth, lipid and contamination, and where does each one work?

RQ2. What data do these models require: how many observations, at what time resolution, from which sensors and with what reference measurements?

RQ3. How sound is validation practice, and which reported results can be trusted outside the dataset that produced them?

Two organizing devices carry the answers. Figure 1 classifies models by where their knowledge comes from. Figure 2 arranges applications on a maturity ladder from offline surrogate to autonomous control, which is the working definition of digital-twin maturity used here: a twin is a model that stays synchronized with its reactor through live data and, at the upper rungs, acts on it.

1.1 Scope and novelty relative to existing reviews

Five reviews retrieved for this work cover neighboring ground (Table 1). They are organized by algorithm, by process stage or by enabling technology. The present review is organized by strength of evidence: it grades validation design, states data requirements per task, and places each application on a maturity ladder with the study that supports it.

Table 1 Scope of existing reviews and of the present work.

Review	Organizing principle	Main subject
Syed et al. [6]	Algorithm	ML approaches across cultivation systems
Oruganti et al. [8]	Application	AI and ML in microalgal wastewater treatment and biorefinery
Yang et al. [7]	Process stage	ML from identification and cultivation to harvesting and conversion
Sheik et al. [9]	Technology	Algal digital twin: concept, hurdles, case studies
Madkour et al. [11]	Task	Deep learning for image-based classification and detection
This review	**Evidence strength**	**Model choice per target, data requirements, validation grading, maturity ladder**

2. Review methodology

2.1 Literature search and selection criteria

Nine semantic queries were run on 11 November 2026 against a scholarly index that aggregates Semantic Scholar, OpenAlex and PubMed. The queries covered (i) ML prediction of growth and lipid productivity, (ii) digital twins of photobioreactors and raceway ponds, (iii) deep learning for contamination and species identification, (iv) reinforcement learning, predictive control and soft sensors, (v) spectroscopy with ML for lipid estimation, (vi) early detection of grazers and culture crashes, (vii) Internet-of-Things monitoring and automation, (viii) uncertainty-aware and physics-informed growth models, and (ix) ML for wastewater-coupled cultivation. They returned 330 records before de-duplication. Only work published up to December 2025 was retained.

Records were kept when they (a) concerned cultivation of microalgae or cyanobacteria, (b) used a learned, hybrid or observer-based model for prediction, monitoring or control, and (c) reported a quantitative result or a deployment. Studies on downstream conversion (hydrothermal liquefaction, pyrolysis, transesterification) and on natural algal blooms were excluded unless they supplied a method directly transferable to cultivation. Forty-five sources were retained from the search as the most informative for the three research questions: 36 primary studies, 7 reviews, perspectives or meta-analyses, and 2 mechanistic or conventional-control studies used as baselines. Fifteen cultivation studies from the author's group on natural-water, wastewater and open-pond systems [12–26] were added to supply the cultivation context; they are not ML studies and are not graded on the validation scale. One source is a preprint and is marked as such in the reference list.

2.2 Data extraction and validation grading

Each primary study was coded for target variable, model family, scale and setting, size and type of dataset, and validation design. Validation was graded on the six-level scale in Table 2, which separates how well a model fits from how far its claim extends.

Table 2 Validation scale.

Level	Design	What it supports
V0	Fit statistics on the training data only	Nothing beyond description
V1	Random hold-out or k-fold on pooled rows	Interpolation within the sampled conditions
V2	Hold-out blocked by batch, run, day or source article	Prediction of a new run in the same system
V3	New experiment run at model-chosen conditions	A prospective, falsifiable prediction
V4	Different reactor, scale, site or season	Transfer
V5	Model acts on the live process in closed loop	Operational value

2.3 Methodological limitations

Coding was done from abstracts and indexed metadata. Where an abstract does not state the validation design, the entry reads "not stated" and is not counted toward any level. The proportions given in Section 6 are therefore lower bounds, to be confirmed by full-text extraction with a PRISMA flow record before submission.

3. Taxonomy of modeling approaches

Models for microalgae cultivation fall into three families defined by where their knowledge comes from, and twelve classes within them (Figure 1). The family decides how a model fails: mechanistic models fail when a parameter is wrong, data-driven models fail outside their training range, and hybrids fail only where both sources are weak

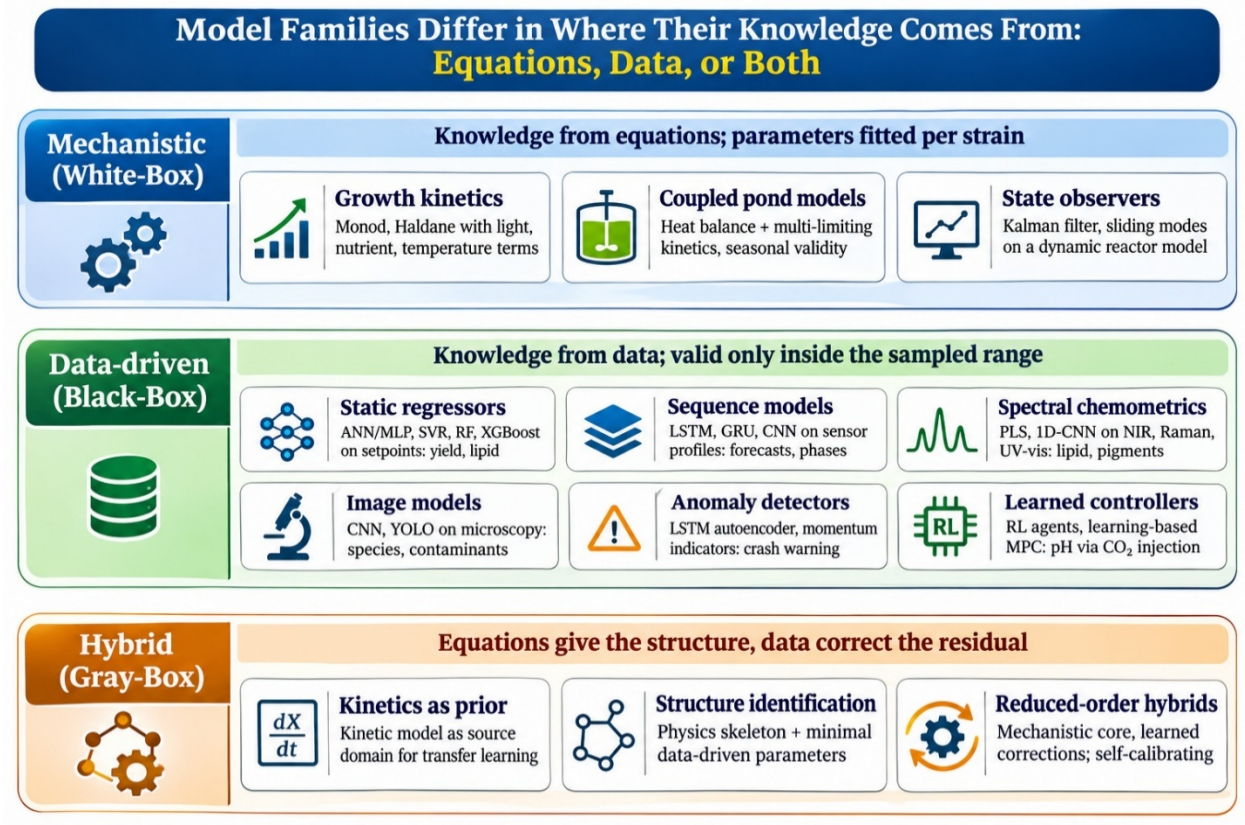

Fig. 1 Taxonomy of models used in microalgae cultivation.

The acronyms in the research questions map onto this taxonomy as follows. Artificial neural networks (ANN, usually a multilayer perceptron) and random forest (RF) are static regressors: they map a row of setpoints to an outcome and hold no memory of earlier time steps. Long short-term memory (LSTM) networks are sequence models that carry state between time steps. Reinforcement learning (RL) is not a predictor but a leaned controller that maps observations to actions. Table 3 sets out what each class needs and where it has been shown to work.

Table 3 Model classes, their data and their demonstrated use in microalgae cultivation.

Class	Typical input	Typical dataset	Demonstrated for	Main weakness	Examples
Static regressors (ANN, SVR, RF, XGBoost)	Setpoints: light, temperature, nutrients, CO_2, time	10^2 to 10^3 rows from designed batches or literature	Biomass and lipid yield, nutrient removal, condition optimization	No dynamics; random splits leak between rows of one run	[10, 27–34]
Sequence models (LSTM, CNN on profiles)	Time series of irradiance, temperature, pH, dissolved oxygen	Weeks to months of logged sensor data	Outdoor growth forecast	Needs long, continuous, labeled records	[5, 35]
Spectral chemometrics (PLS, SVR, 1D-CNN)	NIR, hyperspectral, Raman, UV–vis spectra	Tens to hundreds of spectra with wet-chemistry reference	Lipid, pigments, biomass, composition	Calibration untested across scale and optics	[2, 3, 36–39]
Image models (CNN, YOLO-type detectors)	Microscopy, flow-imaging or culture RGB images	10^3 to 10^5 labeled images	Genus or species identification, culture color to biomass	Scarce annotated data; laboratory images only	[11, 40–43]
Early-warning detectors (autoencoder, physiological and chemical markers)	Redox potential, volatile compounds, fluorescence, infrared spectra	Normal-operation records plus a few crash events	Early warning of contamination and crashes	Few true crash events to test on	[44–47]
Learned controllers (RL, learning-based MPC)	Online pH, CO_2 valve, irradiance	Historical closed-loop logs, then online tuning	pH regulation on raceways	One controlled variable so far	[48–50]
State observers (EKF, sliding mode)	Dissolved oxygen, pH, irradiance plus a dynamic model	A calibrated reactor model	Online biomass estimation	Depends on model fidelity	[51, 52]
Hybrid and uncertainty-aware models	Sparse experiments plus a kinetic structure	10^1 to 10^2 runs	Growth curves, lutein production, small-sample prediction	Requires both modeling and ML skill	[53–56]

Three observations follow from Table 3. First, the classes are not interchangeable: a random forest cannot use light history and an LSTM cannot be trained on 30 Box–Behnken runs. Second, the most cited class, static regressors, is also the one furthest from operation, because it answers "which setpoints are best" for a finished experiment and not "what will the pond do tomorrow". Third, hybrid models are the smallest group although they address the two failure modes that dominate the others, namely small datasets and extrapolation.

4. Predictive performance of machine learning models (RQ1)

No single family wins across the three targets. Growth is best served by tree ensembles on tabular data and by sequence models outdoors, lipid by spectral chemometrics, and contamination by anomaly detection on indirect signals. Table 4 lists the studies that anchor this section.

Table 4 Representative studies by target. Validation levels follow Table 2; "n.s." means not stated in the abstract.

Target	Study	System and scale	Model	Data	Reported result	Validation
Growth	Supriyanto et al. [27]	Polyculture, open raceway, Japan	Three-layer ANN	Pond operating records	$R^2 = 0.93$	n.s.
Growth	Liyanaarachchi et al. [28]	*Chlorella vulgaris*, literature	MLP 23-20-1	484 points from 4 studies	$R^2 = 0.98$	n.s.
Growth	Yeh et al. [5]	*Phaeodactylum tricornutum*, outdoor flat-panel airlift	LSTM, SVR vs Monod, Haldane	50 days of cultivation	ML beats kinetics by using light history	n.s.
Growth	Igou et al. [35]	32 open raceways, 5 US states	Deep model on sensor profiles	598 productivities	$R^2 = 0.77$ vs 0.39 for averaged inputs	n.s. (multi-site data)
Growth	Jiang et al. [53]	*Chlamydomonas reinhardtii*, lab	Kinetic model + TrAdaBoost.R2	Box–Behnken runs	Beats single ML model on small samples	n.s.
Growth	Li et al. [30]	*Porphyridium*, lab	Graph embedding + XGBoost	Cultivation parameters	Train $R^2 > 0.96$, test $R^2 > 0.87$	V1, V3
Growth and lipid	Kumar et al. [33]	Lab photobioreactor	ANN + multi-objective optimization	Designed experiments	Seven-fold lipid productivity	V3
Lipid	Pääkkönen et al. [36]	*C. vulgaris*, closed system	1D-CNN, PLS on NIR imaging	Lipid 2 to 22% of dry weight	RMSE 4 %-points	n.s.

Target	Study	System and scale	Model	Data	Reported result	Validation
Lipid	Chu et al. [37]	*Scenedesmus obliquus*, in situ	CARS-PLSR on Vis/SWNIR imaging	Nitrogen-stress series	Rp = 0.976	V1
Contamination	González-Hernández et al. [57]	4 genera; 2 raceways, 1 tubular reactor	Neural network on 300 to 750 nm spectra	Pure-culture spectra	Macro F1 = 98.64%	V4
Contamination	Deore et al. [44]	*Dunaliella tertiolecta* with *Oxyrrhis marina*	FTIR + PLSR	Lab grazing series	R^2 = 0.894; 72 h warning	n.s.
Contamination	Deore et al. [47]	*D. tertiolecta* with two grazers	Non-photochemical quenching signal	Lab grazing series	Signal falls 24 to 48 h before crash	n.s.
Contamination	Soon et al. [45]	300 L outdoor culture	LSTM autoencoder on redox potential	LoRaWAN sensor series	F1 = 0.971	n.s.
Control	Caparroz et al. [48]	Semi-industrial raceway	Data-driven pH model, adaptive control	More than 100 days in 10 months	Tested against fixed-parameter control	V5
Control	Gil et al. [50] (preprint)	Open industrial photobioreactor	RL with behavior cloning	PID logs, daily online tuning	Control effort −54% vs PID (simulation); 8-day run	V5

4.1 Growth and biomass productivity

Static regressors on designed or compiled data. The earliest and still most common design trains an ANN on a response-surface experiment and reports a parity plot. A 23-20-1 perceptron trained on 484 literature points reproduced *C. vulgaris* cell density with R^2 = 0.98 and ranked CO_2 supply as the dominant input [28], and a three-layer ANN reached R^2 = 0.93 for a native polyculture in an open raceway [27]. Later work compares several learners on the same data: boosted trees, ANN and support vector regression were benchmarked for biomass productivity and CO_2 fixation with Bayesian hyperparameter tuning [29], and seven learners were compared for biomass yield in algal wastewater treatment [31].

The pattern across these comparisons is consistent: on a few hundred to a few thousand tabular rows, kernel and tree-based learners match or beat neural networks, and all beat response-surface polynomials. Feature engineering now adds more than a change of learner. A graph embedding of the relations between cultivation parameters lifted XGBoost to a test R^2 above 0.87 for *Porphyridium*, with Shapley analysis pointing to illumination intensity and phosphate as the critical inputs [30].

Sequence models for outdoor cultivation. Static models break down when the input that matters is a history and not a value. In 50 days of outdoor *P. tricornutum* cultivation, LSTM and SVR models outperformed Monod and Haldane kinetics because they could use past light as input, and the LSTM reproduced light acclimation [5]. At network scale, converting real-time sensor records into profiles and learning from them predicted productivity across 32 raceways in five US states with R^2 = 0.77, against 0.39 for a conventional model fed daily averages, and it did so without biomass density, retention time or nutrient data [35].

Hybrid and small-sample approaches. When only a design-of-experiments dataset exists, a kinetic model can stand in for missing data. Using a logistic kinetic model as the source domain and Box–Behnken runs as the target domain in a two-stage TrAdaBoost.R2 scheme outperformed a stand-alone ML model for *C. reinhardtii* growth curves [53]. Automatic identification of the most probable kinetic structure, combined with a minimal set of data-driven parameters, gave a model able to predict, optimize and self-calibrate an algal lutein process [55], and a reduced-order hybrid model predicted biomass across reactor geometries at a cost compatible with online use [56].

Uncertainty-aware models. Unlike traditional point-prediction models, Gaussian Process (GP) models have the advantage of uncertainty quantification through prediction variance, since they output predicted values without directly quantifying their reliability**.** GP models have been shown to have similar prediction accuracy as previously developed ANNs for microalgal growth and lutein production modelling and also provide variance estimates for the reliability of their predictions**.** These uncertainty estimates can inform dynamic optimisation strategies by identifying regions of high predictive uncertainty and preventing the optimisation process from venturing outside the model's reliable operating domain. [54]. Therefore, GP-based modelling remains an efficient way to improve the prediction reliability and ensure more robust optimization of microalgal cultivation and lutein production processes.

Summary. The determination of an appropriate machine-learning approach is dictated by the nature, size and structure of the available dataset. Gradient boosting and support vector regression (SVR) are popular for tabular batch data, while artificial neural networks (ANNs) and deep-learning methods can capture more complex nonlinear relationships with sufficient training data. The sequence- and profile-based models are especially useful when continuous sensor data and temporal process information are at hand. In cases where experimental data is scarce, hybrid modelling approaches are valuable,

combining the strengths of both mechanistic and data-driven methodologies. [35].

4.2 Lipid content and biochemical composition

Two different problems are both called "lipid prediction". The first maps cultivation setpoints to final lipid yield; the second estimates the lipid content of the culture that is in the reactor now.

Prediction from cultivation setpoints. ANN surrogates combined with multi-objective optimization address the inverse relation between growth and lipid accumulation. One such study reports a seven-fold gain in lipid productivity at the model-chosen optimum, confirmed experimentally [33]; another modeled biomass, total lipid, unsaturated lipid and oleic acid of *C. vulgaris* together [34]. A decision-tree analysis of the 4,670-instance literature database found 11 combinations of conditions leading to high biomass and 13 leading to high lipid, which confirms that the two optima seldom coincide [10]. These models are optimization aids for a given strain and reactor. None of the reviewed studies shows a setpoint-to-lipid model transferring to a second strain or scale.

Spectroscopic estimation. Estimating lipid in the living culture is a calibration problem, and here the evidence is stronger. Partial least squares on NIR and FTIR spectra predicted spiked neutral and polar lipids across four species as early as 2011, with NIR the more robust across species (R^2 = 0.97 and 0.95) [2]. Non-invasive NIR hyperspectral imaging of *C. vulgaris* gave a root-mean-square error of 4 percentage points of dry weight over a 2 to 22% range, and a one-dimensional CNN did no better than PLS [36]. Visible and short-wave NIR imaging with variable selection reached Rp = 0.976 for *S. obliquus* lipid in suspension [37]. Raman micro-spectroscopy quantifies lipid in single cells (Rp^2 = 0.925) [38], and a dual-branch attention CNN that fuses Raman spectra with cell images improved R^2 by 6 to 11% over PLSR and SVR [39]. NIR on harvested slurry predicts protein, lipid, ash and water with R^2 above 0.88 within minutes [3].

What is missing is evidence of transfer between scales. Pooling species improved cross-species prediction in the earliest study [2], and combining NIR datasets from different microalgae extended the measurement range without loss of predictive power [3]. No study in the corpus reports how a laboratory spectral calibration performs on a pilot or production reactor.

Summary. Linear chemometrics (PLS with variable selection) remains competitive with deep models on spectra and should be the baseline. Deep models earn their place when modalities are fused [39]. Setpoint-based ANN models optimize; they do not monitor.

4.3 Contamination detection and culture health monitoring

Image-based identification. Convolutional networks classify microalgae with high accuracy on curated images: a model trained on flow-imaging data distinguished six genera with up to 97.27% accuracy [41], and an earlier pair of models estimated the proportions of *Scenedesmus almeriensis* and *C. vulgaris* in mixtures, with the feature-based model more accurate than the image-based one [40]. A systematic review of 51 studies names the limits: large, high-resolution datasets are lacking and classes are imbalanced [11]. Synthetic training images raised detector precision by up to 25% and recall by up to 23% for diatoms [42], which suggests a route for rare contaminants.

Spectroscopic detection. Image methods need a microscope in the loop. A cheaper route treats contamination as a departure from the spectral signature of the pure culture. A neural network trained on 300 to 750 nm absorbance sweeps of four genera reached a macro F1 of 98.64% and was then used on two raceways and one tubular reactor, reading contamination from the softmax output [57]. FTIR bands with PLSR indicated the grazer *O. marina* at 5×10^2 cells mL^{-1}, at least 72 h before the crash [44].

Indirect early-warning signals. The most recent work drops the requirement to know the contaminant. An LSTM autoencoder trained only on normal redox-potential records from a 300 L outdoor culture flagged abnormal patterns with precision 1.000 and recall 0.944 [45]. Physiological and chemical markers offer the same kind of early signal. Non-photochemical quenching fell 24 to 48 h before the crash in *D. tertiolecta* cultures grazed by *O. marina* and by a ciliate [47]. Four volatile organic compounds were confirmed as markers of rotifer grazing on *Microchloropsis salina* and were proposed as inputs for ML-based diagnosis of pond health [46].

Summary. CNNs answer "what is in this image" and are useful for audits. For operations, unsupervised or semi-supervised anomaly detectors on continuous, inexpensive signals give the earliest warning and do not need a labeled library of pests. Their weakness is evidential: each study rests on a handful of crash events.

4.4 Process control and reinforcement learning

RL appears in the research questions as a predictor, but its role is control. The first reported RL deployment on an open photobioreactor combined offline behavior cloning from PID trajectories with daily online fine-tuning. In simulation it cut the integral absolute error by 8% and control effort by 54% against PID, and it then ran for 8 days on the real system (preprint) [50]. Around it sit non-RL learning controllers for the same variable: a data-driven pH model with decision-tree parameter scheduling validated over more than 100 days in 10 months [48], and a learning-based MPC that improved the average error index up to four-fold over nominal MPC [49].

Summary. RL is feasible on real raceways when it starts from an existing controller's data and is allowed to adapt online. All evidence concerns pH. No reviewed study applies RL to harvest, dilution or nutrient decisions on a real culture.

4.5 Wastewater-based cultivation systems

Cultivation on wastewater adds nutrient removal as a second target and a variable influent as a disturbance. Seven learners were compared for biomass yield in algal

wastewater treatment, with pH and retention time the critical inputs [31]. For a membrane photobioreactor treating sewage, perceptron models with global sensitivity analysis fitted with R^2 of 0.83 to 0.95 on data that retained the diurnal cycle, and were then used for multi-criteria optimization of retention times and aeration [32]. A meta-analysis of pilot-scale high-rate ponds adds the scale caveat that every ML study in this area should cite: pilot growth rates of about 0.54 d^{-1} are half those observed in the laboratory [58].

Summary. Static ANN and SVR models describe nutrient removal well within one influent and one season. The halving of growth rate between laboratory and pilot [58] means that laboratory-trained models should not be used for pilot design without re-calibration.

4.6 Mixed-culture and low-input cultivation systems

Almost all models reviewed above were trained on single strains in defined media. A large part of practical cultivation looks different. Mixed freshwater communities grown on unsupplemented natural water reach CO_2 uptake rates of 175 mg L^{-1} d^{-1} in laboratory reactors [19] and 162 mg L^{-1} d^{-1} in high-rate ponds, where mixed cultures outperformed pure cultures [24]. Such systems have been run continuously over long periods [23] and perform close to an ideally mixed reactor (159 against 178 mg L^{-1} d^{-1}) [26]. In these ponds pH determines which nitrogen species the algae take up [25], which ties the one variable already under learned control (Section 4.4) to nutrient dynamics. The same low-input logic extends to photobioreactors on natural water for protein production [18] and to open cement ponds on a low-cost fertilizer-based medium for *C. vulgaris* [22].

Waste streams add further variability. Microalgae have been grown on tilapia-pond effluent for bioethanol [13] and on palm oil mill effluent, where a designed experiment located the optimum at 10.9% CO_2 and about 9,960 lux [15]. Macroalgae have been raised on irrigation-canal wastewater for biodiesel [14], and wastewater-grown biomass as a biofuel feedstock is reviewed in [16]. Such design-of-experiment optimizations are the statistical precursors of the static regressors in Section 4.1, and their datasets are of the same kind and size.

For ML these systems are the hardest case and the least studied. Species composition is undefined and shifts with season, the medium is not controlled, and polyculture modeling is represented in the corpus by one ANN for one raceway [27]. They are also where a twin would be most useful, because the operator has few levers and little online information. Models validated only on single-strain laboratory cultures should not be assumed to transfer to them.

5. Data requirements (RQ2)

Data needs are set by the task, not by the algorithm, and the scarce resource is labeled reference measurements, not sensor readings. Table 5 summarizes what the reviewed studies used.

Table 5 Data used per task in the reviewed studies.

Task	Dataset sizes reported	Time resolution	Reference label	Route to reduce data need
Setpoint optimization (growth, lipid)	484 points [28]; 4,670 literature instances [10]; designed experiments [29]	One value per run or per day	Dry weight, gravimetric lipid	Kinetic prior with transfer learning [53]
Outdoor growth forecast	50 days in one reactor [5]; 598 productivities from 32 ponds [35]	Minutes for inputs, daily for output	Dry weight or optical density	Sensor profiles in place of bioprocess variables [35]
Online biomass sensing	Image set with cell counts [43]; 9 months of 30-minute sensor records [59]	Seconds to minutes	Dry weight, cell count, suspended solids	RGB histograms [43]; potentiometric sensors [59]
Lipid and composition by spectra	Lipid range 2 to 22% of dry weight [36]; 9 spike levels in 4 species [2]	Per sample	Gravimetric or chromatographic lipid	Pooling species to widen the range [2, 3]
Identification by image	Pure-culture image libraries [40, 41]	Per sample	Expert annotation	Synthetic images [42]
Crash and contamination warning	Normal-operation series [45]; laboratory grazing trials [44, 46, 47]	Minutes to hours	Microscopy, observed crash	Train on normal data only [45]
Closed-loop pH control	More than 100 days over 10 months [48]; PID logs [50]	Seconds	The controlled variable itself	Behavior cloning from the existing controller [50]

5.1 Dataset size and coverage

Tabular studies work with 10^2 to 10^3 rows, and at this size the estimate of error is itself unstable. In a comparison of seven learners for algal wastewater treatment, shifting the train–test ratio by 20% raised the error by 21%, and the model's predicted optimum was 31.7% above the yield obtained experimentally [31]. Both figures describe a model operating at the edge of its data. Larger sets come mainly from compilation across articles, as in the 4,670-instance database [10], which buys breadth at the price of inconsistent measurement methods between laboratories.

For dynamic models the unit of data is the day of continuous record, and coverage matters more than count. Fifty days in one season supported a comparison of models [5], more than 100 days spread over 10 months supported an adaptive controller [48], and 32 ponds in five states supported the only multi-site claim in the corpus [35]. Because outdoor productivity follows fluctuating light and water temperature [1], a data-driven model trained in one season has not seen the conditions of another.

5.2 Sensor signals and input variables

Inexpensive sensors carry more information than is usually extracted from them. The multi-site productivity model used only remote monitoring records, and the gain from $R^2 = 0.39$ to 0.77 came from keeping the within-day profile in place of the daily mean [35]. Light history, not instantaneous light, explained the advantage of LSTM over kinetic models [5]. Redox potential, seldom logged in algal systems, proved a sensitive indicator of both growth and contamination [45].

Optical and electrochemical signals fill the gap between cheap probes and laboratory assays. Color-histogram models from ordinary RGB images are inexpensive but modest, with $R^2 = 0.77$ for cell count and 0.66 for dry biomass [43]. Microbial potentiometric sensors with ML correlated with suspended solids in an algal cultivation pond ($R^2 = 0.87$) and predicted several water-quality variables with normalized errors mostly below 6.5% [59].

5.3 Reference measurements and labeling

Every supervised model in Table 4 depends on a slow reference: dry weight, gravimetric or chromatographic lipid, cell counts or expert annotation. This is why the lipid range in a calibration set is often narrow, and why a model can be precise within 2 to 22% of dry weight [36] and untested above it. For contamination, the label is an event that operators try to prevent, so positive examples are rare by design. The crash-warning studies each rest on a small number of events [45–47]. For mixed cultures on natural water, direct measurement of total and volatile suspended solids was found to be a sensible biomass index [17], which matters because the reference method fixes what a soft sensor can learn.

5.4 Strategies for reducing data requirements

Four strategies in the corpus reduce the labeled data required.

- **Kinetic priors.** A fitted logistic model used as source domain for transfer learning improved small-sample growth prediction [53].
- **Hybrid structure.** Fixing the kinetic skeleton leaves only a minimal number of parameters to learn from data [55].
- **Synthetic data.** Simulated images improved detection precision and recall by about a quarter [42].
- **Learning from the incumbent.** Offline behavior cloning from PID trajectories let an RL agent start without exploratory experiments on the real reactor [50].

Summary for RQ2. A setpoint model needs a few hundred well-spread runs; a forecasting model needs months of continuous sensor records that span seasons; a spectral model needs tens of reference assays that span the full composition range, and its behavior at a new scale is untested; a warning system needs long normal-operation records and, critically, documented crash events. Open, shared time-series datasets are the missing resource: the corpus points to no benchmark for cultivation dynamics.

6. Validation practice (RQ3)

Validation is the weakest part of the literature: of the 15 anchor studies in Table 4, nine do not state their validation design in the abstract, one relies on random hold-out alone, two add a prospective experiment, one tests transfer to other reactors, and two run in closed loop. High R^2 values and strong claims are therefore attached mostly to designs that support interpolation only.

6.1 Recurrent methodological weaknesses

Non-independent data splitting. Samples taken on successive days of one batch, or rows extracted from one article, are not independent. A random 80/20 split places neighbors of every test point in the training set, and the resulting score measures interpolation along a growth curve the model has already seen. The remedy is to hold out whole runs, days or source articles (level V2). This design is seldom reported; none of the anchor studies states it.

Unstable error estimates. With a few hundred rows, the score depends on the split. The 21% change in error for a 20% change in split ratio [31], and the gap between training R^2 above 0.96 and test R^2 above 0.87 in an otherwise careful study [30], show the size of the effect. Repeated or nested cross-validation with the spread reported is the minimum.

Absence of baseline models. A learned model should be compared with the simplest credible alternative. Where this is done, the result is informative in both directions: LSTM beat Monod and Haldane kinetics outdoors [5], profile-based learning beat averaged inputs [35], but a one-dimensional CNN did not beat PLS for lipid [36]. Studies that compare only ML models with each other cannot show that ML was needed. Simple and multiple linear regression on routinely monitored physicochemical variables already

describes biomass growth of mixed freshwater cultures [12], so a linear model is the minimum baseline.

Reliance on R^2 alone. R^2 depends on the range of the response and says nothing about usable error. Errors in the units of the decision, such as 4 percentage points of lipid [36] or normalized errors of a few percent for sensor-based estimates [59], let a reader judge fitness for purpose. For warning systems the operational metrics are lead time and false-alarm rate. Lead times are reported in laboratory grazing trials [44, 47], but false-alarm rates over long normal operation are not, and classification accuracy on balanced laboratory samples [41, 57] and perfect precision on a short test record [45] do not yet answer the operator's question.

Unverified optimization results. A surrogate's optimum is an extrapolation until it is run. The better studies do run it: experimental confirmation of a multi-objective optimum [33] and laboratory verification of model-selected conditions [30]. Where the optimum is reported from the model alone, a discrepancy such as the 31.7% over-prediction in [31] remains hidden.

Untested transferability. Only one anchor study applies a model outside the system that produced its data, moving a spectral classifier from laboratory samples to two raceways and a tubular reactor [57]. No study in the corpus reports the accuracy of a laboratory-trained regression model on a pilot reactor, although pilot growth rates are about half of laboratory rates [58]. This is the test every laboratory model will face.

Simulation-based performance claims. In control studies the headline numbers often come from simulation and the experiment is short. The RL controller's 8% and 54% improvements are simulated; its field evidence is an 8-day run [50]. The longest field records in the corpus are for non-RL adaptive control, at more than 100 days [48].

Interpretation of feature importance. Shapley values are now common [30, 36]. They rank what the model used, which is a property of the model and its training range, not of algal physiology. Rankings such as "CO_2 supply has the highest impact" [28] hold within the sampled range only.

Lack of uncertainty quantification. A prediction used for a harvest or dosing decision needs an interval, and few studies give one. The exception shows what is gained: a Gaussian-process model carried its variance into the optimization and avoided over-optimistic operating points outside its range of validity [54]. Reporting sensitivity to the train–test ratio [31] is a cheap first step.

Table 6 Common weaknesses, their effect on reported accuracy, and the corresponding remedy.

Weakness	Effect	Remedy	Validation level reached
Random split of rows from the same run or article	Optimistic error; no evidence for a new run	Group or block the split by run, day or source	V2
Single split on a small dataset	Score depends on the split	Repeated or nested cross-validation with spread	V1, reported honestly
No mechanistic or linear baseline	Cannot show ML was needed	Report Monod-type or PLS baseline on the same split	Any
R^2 only	Fitness for purpose unknown	Error in decision units; lead time and false alarms for detectors	Any
Optimum taken from the surrogate	Extrapolation unverified	Run the predicted optimum, report the gap	V3
One reactor, one season	Transfer unknown	Test on another scale, site or season; report the direct result before adaptation	V4
Controller assessed in simulation	Operational value unknown	Field run long enough to span weather regimes	V5

6.2 Examples of rigorous validation

The multi-site raceway study uses many ponds and locations and benchmarks against a simpler model [35]. The spectral contamination study trains on pure cultures and then tests the classifier on three production reactors [57]. The data-driven pH model is validated over 10 months and then tested in closed loop against a fixed-parameter controller [48].

Summary for RQ3. Reported accuracies are mostly interpolation scores. Results at V3 or above exist for every target but are a minority, and grouped validation (V2) is the missing step that most studies could add at no experimental cost. Until it is routine, R^2 values from different studies should not be ranked against each other.

7. Digital twins and technology maturity

A digital twin of a microalgae culture is reached in steps, and the published evidence thins sharply above the second step (Figure 2). Growth and product models have been shown up to forecasting; closed-loop and autonomous operation have been shown for pH regulation only.

7.1 Definition and scope of a digital twin

A simulation calibrated once and run offline is a digital model; a model updated by live data without acting on the plant is a digital shadow; a twin exchanges data in both

directions, so that the model's output changes the process. A systematic review of algal digital twins confirms that most work sits at the concept stage and lists interconnection and semantics, facilities, and data acquisition among the hurdles [9]. The ladder in Figure 2 makes the distinction operational: rungs L0 to L2 are models and shadows, L3 advises a human, and only L4 and L5 are twins in the strict sense.

7.2 Available system components

The parts of a twin have each been demonstrated separately.

- **State estimation.** An extended Kalman filter built on a dynamic raceway model estimated biomass in an outdoor industrial reactor [51], and a sliding-mode soft sensor was shown to be robust to noise, initial error and parametric uncertainty [52].
- **Predictive core.** Hybrid models that combine a kinetic structure with learned terms support long-horizon prediction and online self-calibration [55], and reduced-order hybrids keep the cost low enough for monitoring and control [56].
- **Decision layer.** Real-time optimization on a first-principles model sets the references for lower-level controllers so as to maximize profit [60].
- **Adaptation.** Daily online fine-tuning lets a learned controller follow changing process dynamics [50], and learning-based MPC corrects model mismatch during operation [49].

What has not been reported is their integration for a biological objective: a twin that senses biomass and composition, forecasts them under tomorrow's weather, and adjusts dilution, depth or nutrient supply in closed loop.

7.3 pH regulation as the first closed-loop application

pH regulation reached the top rungs because it is the easiest case for learning. The variable is measured directly every few seconds, the actuator (CO_2 injection) acts within minutes, the reward is the measurement itself, and an incumbent PID controller supplies safe training data [48, 50]. Biomass and lipid are the opposite: labels are slow and costly, the response to an action unfolds over days, and a wrong action can lose the culture. Contamination adds the rarity of the event. The route upward for these targets therefore runs through L1: a trustworthy online estimate of biomass and composition is the precondition for every higher rung.

7.4 Recommended implementation pathway

For a pilot raceway or photobioreactor the evidence supports the following order of work.

1. Log irradiance, temperature, pH, dissolved oxygen and redox potential at minute resolution for at least one full seasonal cycle [35, 45, 48].
2. Add one optical or electrochemical channel and calibrate it against dry weight across the full density range [43, 59].
3. Fit a kinetic model first, then learn its residual; keep the kinetic model as the baseline in every report [5, 53, 55].
4. Run an anomaly detector on the normal-operation record and log every alarm with its outcome [45].
5. Move to advisory mode for harvest and dilution, and compare its recommendations with operator decisions before closing any loop.
6. Close the loop first on a variable with a safe fallback controller, as was done for pH [50].

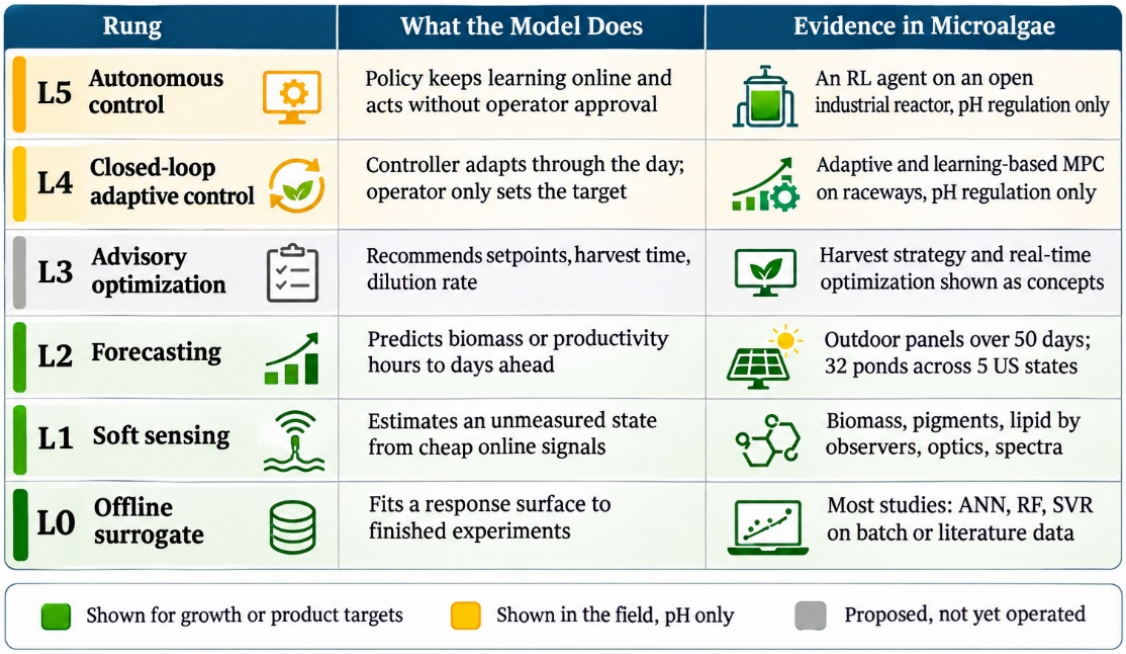

Fig. 2 Maturity ladder from soft sensors to autonomous control, read from the bottom. Color marks the kind of evidence available at each rung.

Table 7 Evidence at each rung of the maturity ladder.

Rung	Defining capability	Best evidence in the corpus	Open gap
L0 Offline surrogate	Maps setpoints to outcome for a finished experiment	ANN, SVR and boosting on designed or compiled data [10, 27–34]	Grouped validation; transfer
L1 Soft sensing	Estimates biomass, pigment or lipid online	Extended Kalman filter on an industrial raceway [51]; sliding-mode sensor [52]; potentiometric sensors with ML [59]; RGB imaging [43]; spectral lipid models [36, 37]	Fouling and drift; lipid sensing in the reactor at scale
L2 Forecasting	Predicts the coming hours to days from live inputs	LSTM with light history over 50 outdoor days [5]; profile-based model across 32 ponds [35]	Weather-forecast inputs; uncertainty bounds
L3 Advisory optimization	Recommends harvest, dilution or setpoints	Harvest strategy derived from a learned growth model [5]; steady-state real-time optimization of raceway profit [60]	Field trials against operator practice
L4 Closed-loop adaptive control	Acts on the process and adapts to conditions	Data-driven adaptive pH control over 10 months [48]; learning-based MPC [49]	Variables other than pH
L5 Autonomous control	Learns online and operates without approval of each action	RL with behavior cloning, 8-day run [50] (preprint)	Long runs; safety constraints; biological objectives

7.3 pH regulation as the first closed-loop application

pH regulation reached the top rungs because it is the easiest case for learning. The variable is measured directly every few seconds, the actuator (CO_2 injection) acts within minutes, the reward is the measurement itself, and an incumbent PID controller supplies safe training data [48, 50]. Biomass and lipid are the opposite: labels are slow and costly, the response to an action unfolds over days, and a wrong action can lose the culture. Contamination adds the rarity of the event. The route upward for these targets therefore runs through L1: a trustworthy online estimate of biomass and composition is the precondition for every higher rung.

7.4 Recommended implementation pathway

For a pilot raceway or photobioreactor the evidence supports the following order of work.

1. Log irradiance, temperature, pH, dissolved oxygen and redox potential at minute resolution for at least one full seasonal cycle [35, 45, 48].
2. Add one optical or electrochemical channel and calibrate it against dry weight across the full density range [43, 59].
3. Fit a kinetic model first, then learn its residual; keep the kinetic model as the baseline in every report [5, 53, 55].
4. Run an anomaly detector on the normal-operation record and log every alarm with its outcome [45].
5. Move to advisory mode for harvest and dilution, and compare its recommendations with operator decisions before closing any loop.
6. Close the loop first on a variable with a safe fallback controller, as was done for pH [50].

7.5 Data infrastructure and economic considerations

The ladder cannot be climbed without a data layer. Long-range low-power networks already carry redox and pH data from outdoor cultures [45], and months of logged sensor records underpin the multi-site forecasting and adaptive-control studies [35, 48].

Two practical points follow. Laboratory cultures do not represent production reactors: pilot growth rates are about half of laboratory rates [58], so a twin has to be calibrated on the reactor it mirrors. The value of an early warning is set by the loss it prevents: predator invasions are reported to cost 20 to 30% of biomass in open systems [4].

The value of a twin should therefore be reported as avoided loss and saved labor per hectare or per cubic meter, not as R^2. No study in the corpus reports the cost of the sensing and computing layer against the gain it delivered over a season, and this is the figure a producer needs before investing.

8. Future research directions and reporting recommendations

8.1 Research priorities

Five gaps follow directly from Sections 4 to 7.

Open benchmark datasets. The corpus points to no shared time-series dataset with sensor records, weather, biomass, composition and documented crashes. A shared set covering several ponds, seasons and at least a few crash events would allow models to be ranked on the same split, which is not possible today.

Online composition sensing at scale. Lipid and composition models work on the bench and on harvested slurry [3, 36, 37]. The step to an in-line probe on a pilot reactor, with the direct result reported before re-calibration, has not been documented. The composition to be sensed

follows the product route: carbohydrate for bioethanol [13], lipid for biodiesel [14, 15], and biomass for biogas co-digestion [20] or biohydrogen [21].

Hybrid predictive models. The few hybrid studies address small data and extrapolation together [53, 55, 56]. Extending them with light-history terms learned by sequence models [5] and with the light and temperature dependence of validated pond models [1] would give a forecast model with a physical floor.

Cost-aware contamination warning. Lead times of 24 to 72 h are reported from spectroscopic and photosynthetic signals in laboratory grazing trials [44, 47], and anomaly detection has been shown on one outdoor culture [45]. The missing figures are false-alarm rates over long normal operation and the economic value of an early harvest. Pooling events across sites is the only way to reach meaningful sample sizes.

Learning-based control beyond pH. Offline-to-online RL has been shown for one variable in one preprint [50]. Dilution rate, harvest timing and nutrient dosing are slower and riskier, so they need a forecast model for safe planning and constraints that keep the culture viable. Advisory operation (L3) with a logged comparison against operators is the necessary intermediate step.

8.2 Reporting checklist

The following ten items would make ML studies in microalgae cultivation comparable. Each can be reported in one or two sentences.

1. **System.** Strain, reactor type and volume, indoor or outdoor, location and dates.
2. **Unit of data.** What one row is (a run, a day, a time step, a spectrum, an image) and how many independent runs the rows come from.
3. **Reference methods.** How each target was measured, with its own repeatability.
4. **Range.** Minimum and maximum of every input and target in the training data.
5. **Split.** Random, grouped by run or source, or blocked in time; who or what is in the test set.
6. **Baseline.** A mechanistic or linear model evaluated on the same split.
7. **Metrics.** Error in the units of the decision, with spread over repeated splits; for detectors, lead time and false-alarm rate.
8. **Validation level.** V0 to V5 as in Table 2, stated explicitly.
9. **Transfer.** The direct result on any second system before adaptation, and the data needed to adapt.
10. **Availability.** Data, code and trained model, or the reason they cannot be shared.

8.3 Limitations of the review

Four limits apply. First, the corpus was assembled by semantic search in one aggregated index and screened by a single reviewer, so relevant studies in conference proceedings or outside the query wording may be missing. Second, coding relied on abstracts and metadata; validation designs described only in the full text are recorded as not stated, which makes the picture in Section 6 conservative. Third, the corpus is restricted to work published up to 2025, and the only reinforcement-learning deployment in it is a preprint whose results may change on peer review. Fourth, reported metrics could not be pooled quantitatively because targets, units and splits differ between studies, which is itself the main finding on validation practice.

9. Conclusions

The three research questions have clear answers on current evidence.

RQ1. The right model depends on the target. For growth, gradient boosting and support vector regression are the sound default on tabular batch data, and sequence models are needed outdoors, where they beat Monod-type kinetics by using light history and are the only class with multi-site evidence. For lipid, spectral chemometrics is the credible route, with PLS still competitive against deep networks; setpoint-based ANN models optimize conditions but do not monitor the culture. For contamination, spectroscopic, photosynthetic and redox signals give 1 to 3 days of warning in laboratory trials and on one outdoor culture, without a pest library. Reinforcement learning is a controller, not a predictor, and has been deployed for pH only.

RQ2. Data needs are set by the task. Setpoint models need a few hundred well-spread runs; forecasting needs months of minute-resolution sensor records across seasons; spectral models need reference assays over the full composition range and re-calibration at each scale; warning systems need documented crash events, which are the scarcest data of all. Transfer learning, kinetic priors, hybrid structure, synthetic data and learning from incumbent controllers each cut the requirement.

RQ3. Validation is the binding constraint. Most reported accuracies come from random hold-out on dependent rows or from unstated designs and support interpolation only. Prospective, transfer and closed-loop validations exist but are a minority, and no study reports how a laboratory-trained regression model performs on a pilot reactor. Hybrid and uncertainty-aware models are the two developments most likely to raise the standard.

On the maturity ladder, microalgae cultivation has working soft sensors and forecasts, concept-level advisory tools, and closed-loop or autonomous control for a single variable. The shortest route to a digital twin that manages biomass and composition runs through trustworthy online sensing, hybrid predictive models with a kinetic floor, grouped validation as standard practice, and shared time-series data that include failures.

Abbreviations: ANN, artificial neural network; CNN, convolutional neural network; EKF, extended Kalman filter; GP, Gaussian process; IoT, Internet of Things; LSTM, long short-term memory; ML, machine learning; MLP, multilayer perceptron; MPC, model predictive control; NIR, near-infrared; PID, proportional-integral-derivative; PLS(R), partial least squares (regression); RF,

random forest; RL, reinforcement learning; RMSE, root-mean-square error; SVR, support vector regression; VOC, volatile organic compound.

10. Acknowledgement

Funding and acknowledgements: The authors thank the Director, Principal, Deans, Heads, faculty members and research colleagues of the Saveetha School of Engineering, SIMATS, Saveetha University for providing constant encouragement, financial assistance, and the necessary research facilities.

Data availability. No new data were generated. The list of screened records and the coding sheet will be provided as supplementary material.

11. Declaration of generative AI in scientific writing

An AI assistant was used for literature retrieval, drafting and figure preparation. The author reviewed and edited the content and takes full responsibility for the publication.

12. CRediT Author Statement

Santhiya Subramani: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review and editing.

Sankarganesh Murugesan – Conceptualization, Writing-review and Editing, Supervision

13. Research Involving Human and Animals Rights

This article is a review of published literature and does not contain any studies with human participants or animals performed by the author.

14. Ethics Approval and Consent to Participate

Not applicable.

15. Declaration of Competing Interest

The author declares no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

16. References

[1] Huesemann, M., Crowe, B., et al. (2015). A validated model to predict microalgae growth in outdoor pond cultures subjected to fluctuating light intensities and water temperatures. Algal Research. https://doi.org/10.1016/j.algal.2015.11.008

[2] Laurens, L., & Wolfrum, E. (2011). Feasibility of spectroscopic characterization of algal lipids: Chemometric correlation of NIR and FTIR spectra with exogenous lipids in algal biomass. BioEnergy Research. https://doi.org/10.1007/s12155-010-9098-y

[3] Bleisch, R., Mühlstädt, G., et al. (2024). A robust, non-invasive and fast routine for the quantification of the nutritional composition of microalgae biomass slurries based on near-infrared spectroscopy. Algal Research. https://doi.org/10.1016/j.algal.2024.103882

[4] Deore, P., Beardall, J., & Noronha, S. (2020). A perspective on the current status of approaches for early detection of microalgal grazing. Journal of Applied Phycology. https://doi.org/10.1007/s10811-020-02241-x

[5] Yeh, Y.-C., Syed, T., et al. (2023). Improving microalgae growth modeling of outdoor cultivation with light history data using machine learning models: A comparative study. Bioresource Technology. https://doi.org/10.1016/j.biortech.2023.129882

[6] Syed, T., Krujatz, F., et al. (2024). A review on machine learning approaches for microalgae cultivation systems. Computers in Biology and Medicine. https://doi.org/10.1016/j.compbiomed.2024.108248

[7] Yang, C.-T., Kristiani, E., et al. (2024). Machine learning in microalgae biotechnology for sustainable biofuel production: Advancements, applications, and prospects. *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2024.131549

[8] Oruganti, R., Biji, A. P., et al. (2023). Artificial intelligence and machine learning tools for high-performance microalgal wastewater treatment and algal biorefinery: A critical review. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2023.162797

[9] Sheik, A. G., Kumar, A., et al. (2024). Reinvigorating algal cultivation for biomass production with digital twin technology: A smart sustainable infrastructure. Algal Research. https://doi.org/10.1016/j.algal.2024.103779

[10] Çoşgun, A., Günay, M. E., & Yıldırım, R. (2020). Exploring the critical factors of algal biomass and lipid production for renewable fuel production by machine learning. Renewable Energy. https://doi.org/10.1016/j.renene.2020.09.034

[11] Madkour, D. M., Shapiai, M. I., et al. (2023). A systematic review of deep learning microalgae classification and detection. IEEE Access. https://doi.org/10.1109/access.2023.3280410

[12] Ramaraj, R., Tsai, D. D.-W., & Chen, P. H. (2014). An exploration of the relationships between microalgae biomass growth and related environmental variables. Journal of Photochemistry and Photobiology B: Biology, 135, 44–47. https://doi.org/10.1016/j.jphotobiol.2014.04.001

[13] Bhuyar, P., Trejo, M., Dussadee, N., Unpaprom, Y., Ramaraj, R., & Whangchai, K. (2021). Microalgae cultivation in wastewater effluent from tilapia culture pond for enhanced bioethanol production. Water Science and Technology. https://doi.org/10.2166/wst.2021.194

[14] Whangchai, K., Souvannasouk, V., Bhuyar, P., Ramaraj, R., & Unpaprom, Y. (2021). Biomass generation and biodiesel production from macroalgae grown in the irrigation canal wastewater. Water Science and Technology. https://doi.org/10.2166/wst.2021.195

[15] Bhuyar, P., Sundararaju, S., Rahim, M. H. A., Ramaraj, R., Maniam, G. P., & Govindan, N. (2019). Microalgae cultivation using palm oil mill effluent as growth medium for lipid production with the effect of CO_2 supply and light intensity. Biomass Conversion and Biorefinery. https://doi.org/10.1007/s13399-019-00548-5

[16] Palanisamy, K. M., Bhat, O. A., Oteikwu, M. O., Govindan, N., Maniam, G. P., Ramaraj, R., & Unpaprom, Y. (2022). Production of biofuel from microalgae grown in wastewater: A review. Maejo International Journal of

Energy and Environmental Communication, 4(3), 17–26. https://doi.org/10.54279/mijeec.v4i3.248256

[17] Ramaraj, R., Tsai, D. D.-W., & Chen, P. H. (2015). Biomass of algae growth on natural water medium. Journal of Photochemistry and Photobiology B: Biology, 142, 124–128. https://doi.org/10.1016/j.jphotobiol.2014.12.007

[18] Tsai, D. D.-W., Ramaraj, R., et al. (2023). Autotrophic production of bio-functional proteins from freshwater microalgae using natural water medium for an economical and ecofriendly approach. Maejo International Journal of Energy and Environmental Communication, 5(2). https://doi.org/10.54279/mijeec.v5i2.250835

[19] Ramaraj, R., Tsai, D. D.-W., & Chen, P. H. (2014). Freshwater microalgae niche of air carbon dioxide mitigation. Ecological Engineering. https://doi.org/10.1016/j.ecoleng.2014.03.058

[20] Wannapokin, A., Ramaraj, R., Whangchai, K., & Unpaprom, Y. (2018). Potential improvement of biogas production from fallen teak leaves with co-digestion of microalgae. 3 Biotech, 8, 123. https://doi.org/10.1007/s13205-018-1084-7

[21] Pathy, A., Nageshwari, K., Ramaraj, R., Pragas Maniam, G., Govindan, N., & Balasubramanian, P. (2022). Biohydrogen production using algae: Potentiality, economics and challenges. Bioresource Technology. https://doi.org/10.1016/j.biortech.2022.127514

[22] Ramaraj, R., Unpaprom, Y., & Dussadee, N. (2016). Cultivation of green microalga, Chlorella vulgaris for biogas purification. (Journal name, volume and pages to be added by the author.)

[23] Ramaraj, R., Tsai, D. D.-W., & Chen, P. H. (2015). Carbon dioxide fixation of freshwater microalgae growth on natural water medium. Ecological Engineering. https://doi.org/10.1016/j.ecoleng.2014.11.033

[24] Tsai, D. D.-W., Ramaraj, R., & Chen, P. H. (2016). Carbon dioxide bio-fixation by algae of high rate pond on natural water medium. Ecological Engineering. https://doi.org/10.1016/j.ecoleng.2016.03.021

[25] Tsai, D. D.-W., Ramaraj, R., & Chen, P. H. (2012). Growth condition study of algae function in ecosystem for CO_2 bio-fixation. Journal of Photochemistry and Photobiology B: Biology, 107, 27–34. https://doi.org/10.1016/j.jphotobiol.2011.11.005

[26] Tsai, D. D.-W., Chen, P. H., & Ramaraj, R. (2017). The potential of carbon dioxide capture and sequestration with algae. Ecological Engineering. https://doi.org/10.1016/j.ecoleng.2016.10.049

[27] Supriyanto, Noguchi, R., et al. (2018). Artificial neural networks model for estimating growth of polyculture microalgae in an open raceway pond. Biosystems Engineering. https://doi.org/10.1016/j.biosystemseng.2018.10.002

[28] Liyanaarachchi, V. C., Nishshanka, G., et al. (2020). Development of an artificial neural network model to simulate the growth of microalga Chlorella vulgaris incorporating the effect of micronutrients. Journal of Biotechnology. https://doi.org/10.1016/j.jbiotec.2020.02.010

[29] Hossain, S. M. Z., Sultana, N., et al. (2022). Modeling and multi-objective optimization of microalgae biomass production and CO_2 biofixation using hybrid intelligence approaches. Renewable and Sustainable Energy Reviews. https://doi.org/10.1016/j.rser.2021.112016

[30] Li, H., Chen, L., et al. (2024). Graph-learning-based machine learning improves prediction and cultivation of commercial-grade marine microalgae Porphyridium. Bioresource Technology. https://doi.org/10.1016/j.biortech.2024.131728

[31] Meenatchisundaram, K., Gowd, S. C., et al. (2024). Data-driven model development for prediction and optimization of biomass yield of microalgae-based wastewater treatment. Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2024.103670

[32] Szeląg, B., González-Camejo, J., et al. (2024). Multi-criteria analysis of the continuous operation of a membrane photobioreactor to treat sewage: Modeling and sensitivity analysis. Chemical Engineering Journal. https://doi.org/10.1016/j.cej.2024.154202

[33] Kumar, R., Sarkar, D., & Sen, R. (2024). Simultaneously maximizing microalgal biomass and lipid productivities by machine learning driven modeling, global sensitivity analysis and multi-objective optimization for sustainable biodiesel production. Applied Energy. https://doi.org/10.1016/j.apenergy.2023.122597

[34] Liyanaarachchi, V. C., Nishshanka, G., et al. (2021). Artificial neural network (ANN) approach to optimize cultivation conditions of microalga Chlorella vulgaris in view of biodiesel production. Biochemical Engineering Journal. https://doi.org/10.1016/j.bej.2021.108072

[35] Igou, T. K. S., Zhong, S., et al. (2023). Real-time sensor data profile-based deep learning method applied to open raceway pond microalgal productivity prediction. Environmental Science & Technology. https://doi.org/10.1021/acs.est.2c07578

[36] Pääkkönen, S., Pölönen, I., et al. (2024). Lipid monitoring of Chlorella vulgaris using non-invasive near-infrared spectral imaging. Journal of Applied Phycology. https://doi.org/10.1007/s10811-024-03397-6

[37] Chu, B.-Q., Chen, K., et al. (2021). Visible/short-wave near-infrared hyperspectral analysis of lipid concentration and fatty acid unsaturation of Scenedesmus obliquus in situ. Computers and Electronics in Agriculture. https://doi.org/10.1016/j.compag.2021.106046

[38] Li, X., Sha, J., et al. (2019). Quantitative visualization of intracellular lipids concentration in a microalgae cell based on Raman micro-spectroscopy coupled with chemometrics. Sensors and Actuators B: Chemical. https://doi.org/10.1016/j.snb.2019.04.048

[39] Zhan, B.-S., Tan, Z., et al. (2025). Deep learning with multimodal Raman spectral fusion framework: An analytical approach for microalgal lipid quantification. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. https://doi.org/10.1016/j.saa.2025.127009

[40] Otálora, P., Guzmán, J. L., et al. (2021). Microalgae classification based on machine learning techniques. Algal Research. https://doi.org/10.1016/j.algal.2021.102256

[41] Otálora, P., Guzmán, J. L., et al. (2023). An artificial intelligence approach for identification of microalgae cultures. New Biotechnology. https://doi.org/10.1016/j.nbt.2023.07.003

[42] Venkataramanan, A., Faure-Giovagnoli, P., et al. (2023). Usefulness of synthetic datasets for diatom automatic detection using a deep-learning approach. Engineering Applications of Artificial Intelligence. https://doi.org/10.1016/j.engappai.2022.105594

[43] Uğuz, S., Şahin, Y. S., et al. (2025). Real-time algal monitoring using novel machine learning approaches. Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc9060153

[44] Deore, P., Beardall, J., et al. (2021). FTIR combined with chemometric tools: A potential approach for early screening of grazers in microalgal cultures. Journal of Applied Phycology. https://doi.org/10.1007/s10811-021-02543-8

[45] Soon, C. F., Jamiaan, M., et al. (2025). Smart sensing and anomaly detection for microalgae culture based on LoRaWAN sensors and LSTM autoencoder. Aquaculture International. https://doi.org/10.1007/s10499-025-02104-7

[46] Fisher, C. L., Lane, P. D., et al. (2020). Low molecular weight volatile organic compounds indicate grazing by the marine rotifer Brachionus plicatilis on the microalgae Microchloropsis salina. Metabolites. https://doi.org/10.3390/metabo10090361

[47] Deore, P., Karthikaichamy, A., et al. (2020). Non-photochemical quenching, a non-invasive probe for monitoring microalgal grazing: An early indicator of predation by Oxyrrhis marina and Euplotes sp. Applied Phycology. https://doi.org/10.1080/26388081.2019.1651218

[48] Caparroz, M., Guzmán, J. L., et al. (2024). A novel data-driven model for prediction and adaptive control of pH in raceway reactor for microalgae cultivation. New Biotechnology. https://doi.org/10.1016/j.nbt.2024.04.001

[49] Pataro, I., Gil, J. D., et al. (2023). A learning-based model predictive strategy for pH control in raceway photobioreactors with freshwater and wastewater cultivation media. Control Engineering Practice. https://doi.org/10.1016/j.conengprac.2023.105619

[50] Gil, J. D., del Río Chanona, E. A., et al. (2025). Reinforcement learning meets bioprocess control through behaviour cloning: Real-world deployment in an industrial photobioreactor [Preprint]. arXiv. https://doi.org/10.48550/arxiv.2509.06853

[51] García-Mañas, F., Guzmán, J. L., et al. (2018). Biomass estimation of an industrial raceway photobioreactor using an extended Kalman filter and a dynamic model for microalgae production. Algal Research. https://doi.org/10.1016/j.algal.2018.11.009

[52] Delgado, E., Moreno, J. C., et al. (2024). Soft-sensor based on sliding modes for industrial raceway photobioreactors. Biosystems Engineering. https://doi.org/10.1016/j.biosystemseng.2024.07.015

[53] Jiang, M., Cao, X.-P., et al. (2023). A kinetic-assisted growth curve prediction method for Chlamydomonas reinhardtii incorporating transfer learning. Bioresource Technology. https://doi.org/10.1016/j.biortech.2023.130246

[54] Bradford, E., Schweidtmann, A. M., et al. (2018). Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. Computers & Chemical Engineering. https://doi.org/10.1016/j.compchemeng.2018.07.015

[55] Zhang, D., Savage, T. R., & Cho, B. A. (2020). Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation. Biotechnology and Bioengineering. https://doi.org/10.1002/bit.27512

[56] Shahhoseyni, S., Greco, L., et al. (2024). A reduced-order hybrid model for photobioreactor performance and biomass prediction. Algal Research. https://doi.org/10.1016/j.algal.2024.103750

[57] González-Hernández, J., Ciardi, M., et al. (2024). A low-cost methodology based on artificial intelligence for contamination detection in microalgae production systems. Algal Research. https://doi.org/10.1016/j.algal.2024.103849

[58] Velásquez-Orta, S. B., Yáñez-Noguez, I., et al. (2024). Pilot-scale microalgae cultivation and wastewater treatment using high-rate ponds: A meta-analysis. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-024-34000-7

[59] Saboe, D., Ghasemi, H., et al. (2020). Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.142876

[60] Otálora, P., Skogestad, S., et al. (2024). Modeling, control and online optimization of microalgae-based biomass production in raceway reactors. IFAC-PapersOnLine. https://doi.org/10.1016/j.ifacol.2024.08.342