Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hyperspectral time-series clustering of microalgal growth kinetics for strain-aware harvest timing

Rajalakshmi Elumalai
SRM University, Andhra Pradesh
Karthikeyan Meenatchi Sundaram
SRM University, Andhra Pradesh
Sunil Chinnadurai
SRM University, Andhra Pradesh
Ankita Juneja
SUNY College of Environmental Science and Forestry
Karthik Rajendran
rajendran.k@srmap.edu.in

SRM University, Andhra Pradesh

Research Article

Keywords: Microalgae, Hyperspectral imaging, Dynamic time warping, Consensus clustering, Growth kinetics

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11094291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hyperspectral time-series clustering of microalgal growth kinetics for strain-aware harvest timing

Rajalakshmi Elumalai[a], Karthikeyan Meenatchi Sundaram[b], Sunil Chinnadurai[b], Ankita Juneja[c], Karthik Rajendran[a, *]

[a]The NET ZERO Lab, Department of Environmental Science and Engineering, SRM University-AP, Amaravati, 522 240, Andhra Pradesh, India

[b]Department of Electronics and Communication Engineering, SRM University-AP, Amaravati, 522 240, Andhra Pradesh, India

[c]Department of Chemical Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY, 13210, USA

Corresponding author: rajendran.k@srmap.edu.in

Abstract

Microalgal cultivation involves strains with distinct growth-kinetic behaviors, making it difficult to define a common cultivation and harvesting strategy across strains. Differences in the timing and progression of growth may result in suboptimal harvest decisions when a fixed harvesting time is used. To address the challenge, the study developed a hyperspectral time-series clustering framework to group microalgal and cyanobacterial strains according to their growth-kinetic behaviour rather than taxonomic relationships. Daily hyperspectral imaging over 31 days across 400–1000 nm with 204 spectral bands generated 1,860 images and 37,200 spectral readings, which were transformed into six biochemical band trajectories for temporal analysis. Three Dynamic Time Warping-based clustering approaches, including Hard-DTW *K*-means, Soft-DTW *K*-means, and agglomerative hierarchical DTW clustering, were evaluated and integrated using a species-wise co-association matrix. Hierarchical DTW produced the highest internal clustering validity among the individual approaches, with a DTW-silhouette value of 0.231. The five consensus clusters differed significantly in six of eight independent pigment and growth-curve metrics excluded from clustering (Kruskal–Wallis, $q < 0.05$), while showing weak agreement with coarse taxonomic grouping (adjusted Rand index = −0.045). Peak-signal timing varied from day 2 to day 25 among clusters, indicating the potential of growth-kinetic grouping to support strain-specific harvest timing. However, the consensus partition was identical to the Hard-DTW partition in the present analysis. Overall, the framework provides a data-driven approach for grouping strains by temporal behaviour and supporting harvest-timing decisions in microalgal cultivation.

Keywords: Microalgae; Hyperspectral imaging; Dynamic time warping; Consensus clustering; Growth kinetics

1. Introduction

Microalgae are used across a wide range of industrial applications, [1] and such applications include biofuel production, pigment and protein extraction, CO_2 fixation, and wastewater treatment [2]. The large-scale cultivation of microalgae involves diverse strains with different physiological characteristics and growth responses [2]. Even under similar cultivation conditions, strains can exhibit substantial differences in biomass productivity, with reported values ranging from 0.04 to 0.37 g L^{-1} d^{-1} [3]. Such variation presents an operational challenge because applying a fixed harvest schedule across strains with different growth kinetics may result in substantial yield losses [4]. Therefore, understanding the temporal growth behaviour of individual strains is important for improving cultivation efficiency and harvest decisions.

Currently, microalgal cultivation is commonly monitored using periodic microscopy [5], optical measurements, and taxonomic characterization [6]. Such approaches provide useful information on cell morphology, biomass accumulation, and strain identity, but they provide limited information on how growth behaviour changes continuously throughout cultivation. The limitation is particularly relevant because microalgae are sensitive to changes in environmental and cultivation conditions and can undergo substantial changes in physiological state during different growth phases [7]. Delayed harvesting after the period of maximum biomass accumulation may also reduce the quality or quantity of valuable cellular products, including pigments and lipids [7]. Therefore, continuous characterization of growth trajectories is required to better distinguish differences in the timing and progression of growth among strains. A monitoring approach that captures both spectral and temporal information could provide a more comprehensive basis for evaluating strain-specific growth behaviour [2].

Hyperspectral imaging (HSI) has emerged as a non-destructive technique for monitoring microalgal cultures because it provides spatial and spectral information across a broad wavelength range [8]. Unlike conventional optical density measurements, which mainly provide an indirect measure of biomass concentration and culture turbidity [6], HSI can capture changes in spectral characteristics associated with pigments and cellular biochemical components during cultivation [9]. Previous studies have demonstrated the potential of hyperspectral and image-based methods for microalgal classification [10], with machine-learning and deep-learning approaches achieving accuracies ranging from 87.04% to 98.98% [11], [12]. Hyperspectral measurements have also been used to distinguish microalgal species and identify different growth stages from spectral information [13]. However, most existing applications have focused on classification or growth-stage identification rather than comparing complete growth trajectories across multiple strains. Because strains may reach comparable physiological stages at different times, temporal analysis of growth trajectories is required to identify strains with similar kinetic behaviour.

Growth-trajectory comparison across strains can therefore be formulated as a time-series clustering problem. Conventional measures such as Euclidean distance are sensitive to temporal shifts and may consider biologically similar growth patterns dissimilar when growth events occur at different times [14]. Dynamic Time Warping (DTW) address the limitation by aligning corresponding temporal features [15] despite differences in their timing [16]. Accordingly, DTW-based methods, including Soft-DTW and hierarchical clustering, have been applied to temporally variable time-series data [15], while consensus clustering can integrate multiple clustering solutions [17]. However, DTW-based *k*-means has not consistently outperformed Euclidean *k*-means in benchmark studies [15], and microalgal time-series studies have mainly focused on single-strain biomass forecasting rather than multi-strain growth-kinetic grouping [18]. Such limitations indicate the need for an integrated framework that combines multiple DTW-based clustering strategies with consensus clustering and independent biological validation.

No prior study has combined multi-algorithm DTW clustering with a consensus step and independent biological validation for microalgae growth-kinetic phenotyping. The study addresses the gap by developing a hyperspectral growth-trajectory framework for twenty microalgal and cyanobacterial strains throughout the cultivation period. The objectives were to: (i) establish a hyperspectral growth-trajectory dataset for twenty microalgal and cyanobacterial strains; (ii) compare three DTW-based clustering strategies and integrate their results into a consensus partition; (iii) independently validate the resulting clusters using taxonomy, pigment indices, and kinetic parameters withheld from the clustering process; and (iv) evaluate the implications of strain-specific

growth kinetics for harvest timing. The proposed framework provides a basis for strain-aware harvest decision-making by identifying groups of strains with similar temporal growth behaviour.

2. Materials and Methods

In the present study, the proposed methodology consisted of six sequential phases: strain selection and cultivation, hyperspectral data acquisition, feature and trajectory tensor formulation, time-series clustering, ensemble consensus clustering, and biological validation, as illustrated in Fig. 1. The pipeline was implemented in Python 3.12.7 with Keras, using scikit-learn [19], tslearn [20], SciPy, and pandas, and was run in Jupyter Notebook 7.2.2 on a GPU-enabled system (Intel Core i7-10750H CPU, NVIDIA GeForce RTX 3060 GPU, 16 GB RAM, 64-bit Windows).

2.1. Microalgae Strain Selection and Cultivation

Twenty microalgal and cyanobacterial strains were selected for the present study, covering Chlorophyta, Cyanobacteria, Ochrophyta, and Haptophyta. The complete list of microalgal species is provided in Supplementary Table S1. The strains were obtained from three different sources: the National Repository for Microalgae and Cyanobacteria–Freshwater (NRMC-F), Tiruchirappalli; the Indian Council of Agricultural Research–Central Marine Fisheries Research Institute (ICAR-CMFRI), Tuticorin; and the National Collection of Industrial Microorganisms (NCIM), Pune.

The selected strains were cultivated using BG-11 [21], ASN-3 [22], and Bold Basal (BB) media [23] at 25–27 °C for 31 days in 250 mL Erlenmeyer flasks under a 12 h light/12 h dark photoperiod. Detailed information on the culture media, cultivation conditions, and other growth parameters is provided in Supplementary Table S3. Each strain was assigned a fixed four-letter canonical code (e.g., ANBA, CHLO, TETR), used consistently throughout data acquisition, model training, and validation.

2.2. Hyperspectral Data Acquisition & Preprocessing

A Specim IQ hyperspectral camera was used for imaging, covering the 400 to 1000 nm range across 204 spectral bands. The spectral resolution was 7 nm, and the spatial resolution was 512 by 512 pixels [13]. The camera was placed inside a darkroom to block out any outside light during imaging and was fixed at a height of 15 cm above each sample for consistent framing. A tungsten-halogen lamp was placed at a 90-degree angle to reduce glare from the sample surface, and the samples were placed in small 5 mL petri plates next to a white reference tile [13].

Imaging was carried out once each day for 31 days, covering the lag, exponential, and stationary growth phases, with three replicate images taken for each species at every time point. The imaging schedule produced 1,860 hyperspectral images in total, giving 37,200 spectral readings across the twenty species.

Each image was calibrated against the white reference to obtain relative reflectance values for every pixel, following the same white-reference calibration principle used in recent non-invasive microalgae monitoring work [24] (calculation given in supplementary file, Table S2). Spectra were then extracted from fixed regions of interest using a grid-based method, and the mean reflectance across all pixels in each region was taken as the representative spectrum. A Savitzky-Golay filter (window size 11, order 3) was applied next, to reduce instrument noise while keeping the shape of pigment absorption peaks [25]. The filter works by fitting a small polynomial to a moving window of nearby wavelength points and using that polynomial's value at the center point, rather than simply averaging neighbours, so sharp absorption peaks are smoothed without being flattened. All spectra were then resampled to a fixed 1 nm step size, so that every sample shared the same wavelength grid, and reflectance values were finally normalized to a 0 to 1 range across all samples [26].

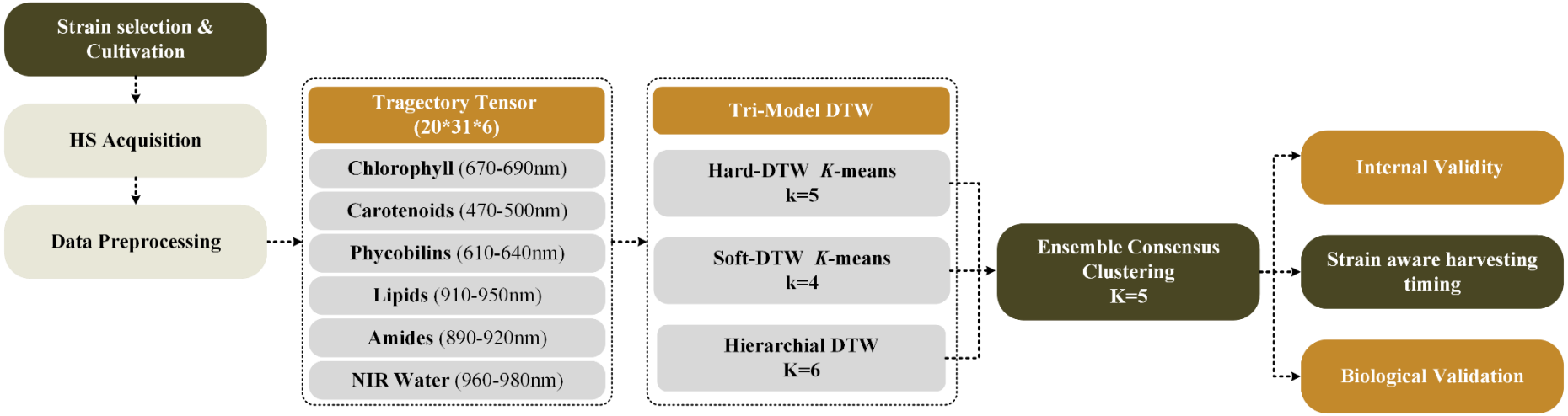


Fig. 1. Overview of the six-phases analytical pipeline: strain selection and cultivation, hyperspectral data acquisition and preprocessing, feature and trajectory tensor formulation, trimodal DTW clustering (Hard DTW *K*-means, Soft-DTW *K*-means, and agglomerative Hierarchical/DTW), ensemble consensus clustering via a species-by-species co-association matrix (K = 5), and independent biological validation against taxonomy, pigment indices, and growth-curve kinetics.

Each spectrum was also treated with a Standard Normal Variate (SNV) step, to remove scattering effects that vary from sample to sample (calculation given in Table S2), rescaling each spectrum to have zero mean and unit variance across its own wavelengths [27]. The SNV step removes multiplicative scaling differences caused by sample thickness or positioning, without needing a separate reference sample for every measurement [27]. A five-point moving average and a first derivative step followed, reducing baseline drift and highlighting slope changes tied to pigment absorption. In addition, pigment-related spectral indices, including the Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), and Simple Ratio (SR), were calculated and standardized separately for each sample [24]. The standardized spectral indices were subsequently concatenated with the principal component features to generate a combined feature representation. The combined representation was used for exploratory visualization and subsequent biological validation.

2.3. Feature and biochemical trajectory tensor formulation

In the present study, six biochemical wavelength bands were defined in advance, based on known pigment and macromolecule absorption features: Chlorophyll (670-690 nm) [28], Carotenoid (470-500 nm) [28], Phycobiliprotein (610-640 nm) [29], Lipid C-H (910-950 nm) [8], Protein Amide (890-920 nm) , and Water/NIR (960-980 nm) [30]. The mean reflectance within each band's wavelength window was then computed to collapse each spectrum to one value per band (calculation given in Table S2), giving one representative number per band per spectrum.

The ROI measurements were averaged within each species-day combination, and the average result produced a dense species by day by band trajectory tensor, so that each strain contributed one growth curve per band. Any species-day combination without a directly measured value was filled by linear interpolation along the day axis, and the ensured every strain contributed a complete kinetic trajectory across the full cultivation period used for clustering [31].

2.3.1. Band decoupling and anti-circularity safeguards

To minimize potential circularity between the features used for clustering and the features subsequently used for biological validation, the Water/NIR band (960–980 nm) was predefined as the sole descriptor of growth kinetics and was not given preferential weighting during cluster formation [32]. The separation ensured that growth-curve-derived metrics, including area under the curve, lag time, time to maximum growth rate, and time to stationary phase, provided an independent assessment of cluster characteristics rather than directly reproducing the criteria used for clustering.

2.3.2. Pigment indices and PCA reduction

Principal Component Analysis (PCA) was applied to reduce the high-dimensional spectral data to a smaller set of principal components before analyses requiring a flat feature representation, including visualization of cluster separation in reduced-dimensional space (shown per model and consensus in Fig. S3). The pigment indices (NDVI, GNDVI, Simple Ratio) were kept as separate covariates rather than being absorbed into the principal components, and keeping their later use in independent validation stayed traceable to a named, interpretable quantity [13].

2.4. Time-Series Clustering Algorithms

Three complementary clustering strategies, all operating on the same standardized species x day x band trajectory tensors, were evaluated for their ability to group strains by shared growth-kinetic behaviour as illustrated in Fig. 2. Each trajectory was first standardized to zero mean and unit variance so that no single band or species dominated the analysis purely through differences in raw signal scale [33].

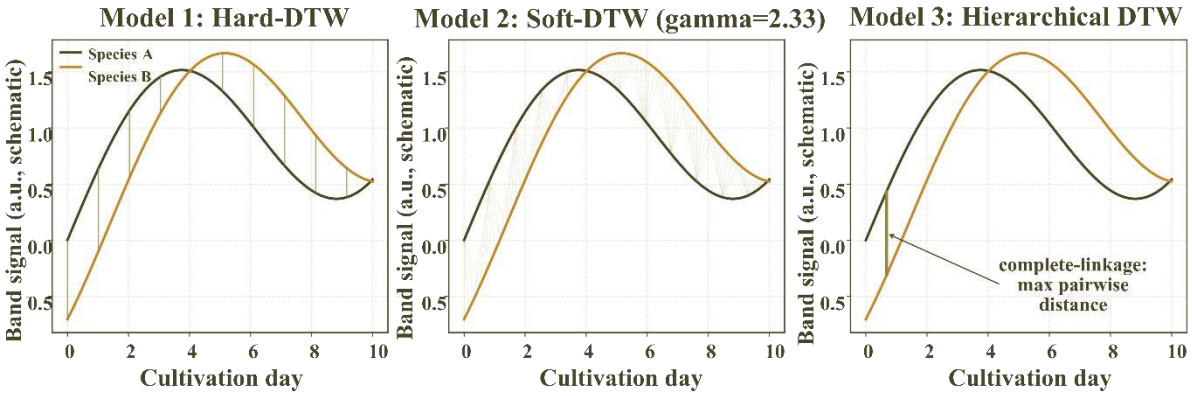


Fig.2. Schematic comparison of the three DTW-based clustering strategies used in the present study, drawn on illustrative example curves rather than real strain data. (a) Hard-DTW K-means: two schematic growth curves ("Species A", "Species B") are aligned along a single, lowest-cost warping path (thin vertical connecting lines), the alignment used by Eq. (1) without smoothing. (b) Soft-DTW (gamma = 2.33): the same two curves are compared using the smoothed distance of Eq. (2); the dense web of thin connecting lines illustrates that many alternative alignment paths are blended into the final distance, rather than relying on one single path as in (a). (c) Hierarchical/DTW: the same two curves are compared using complete-linkage distance, defined as the single maximum pairwise distance between them (annotated arrow), which is the pairwise-distance quantity feeding the agglomerative clustering

Dynamic Time Warping (DTW) was used as the base distance measure for all three models in the present study, since it can measure how similar two growth trajectories are even when one is shifted in time relative to the other [15]. For two series x (length n) and y (length m), the DTW distance is defined by the recursive cost:

$$D(i,j) = d(x_i, y_j) + \begin{matrix} D(i-1,j) \\ \int \quad D(i,j-1) \\ D(i-1,j-1) \end{matrix} \qquad (1)$$

Here, $d(x_i, y_j)$ is the cost of matching point i in x to point j in y (the squared difference between them), and $D(i, j)$ is the lowest total cost of aligning the two series up to those points. In essence, the above equation lets a growth curve stretch or compress in time to match another curve's shape, so two strains that follow the same growth pattern but reach it a day or two apart are still recognized as similar.

Hard-DTW *K*-means: A Sakoe-Chiba warping-window constraint of three days was applied to the model, limiting how far the stretching in Eq. (1) can reach and preventing biologically implausible long-range temporal alignment between strains [34]. The number of clusters was chosen by sweeping candidate values and selecting the configuration with the highest silhouette score [35], subject to a balance constraint that prevented any single cluster from containing more than sixty percent of all strains.

Soft-DTW *K*-means: A smoothed variant of DTW was applied for the model, to make the distance measure differentiable and less sensitive to a single best path. The variant replaces the hard minimum in Eq. (1) with a soft minimum:

$$\min(a,b,c) = m - \gamma \log\left(e^{-\frac{a-m}{\gamma}} + e^{-\frac{b-m}{\gamma}} + e^{-\frac{c-m}{\gamma}}\right) \qquad (2)$$

Here, gamma is a smoothing parameter, and a, b, c stand for the three terms being minimized in Eq. (1). As gamma approaches zero, the soft minimum behaves like the ordinary minimum, recovering Hard-DTW; larger gamma values blend more of the alternative alignment paths into the final distance, instead of relying on the single lowest-cost path. Both gamma and the number of clusters were jointly optimized through grid search, with candidate configurations first ranked by silhouette score. The top-ranked configurations were then re-evaluated for stability, measured as the mean pairwise Adjusted Rand Index of cluster assignments across five independent random

initializations. The final gamma and cluster count were selected by combining silhouette score and stability in equal weight, and such criteria favored solutions that were both well-separated and reproducible, rather than optimal on a single run [14].

Agglomerative Hierarchical/DTW clustering: A pairwise DTW distance matrix was computed across all strains for the model, using Eq. (1), and the resulting distance matrix was used as the input to agglomerative hierarchical clustering [16]. The linkage method and cluster count were jointly selected by silhouette score, under the same cluster-balance constraint used for the two *K*-means variants. Because the method is deterministic given a fixed distance matrix, its stability was instead assessed by recomputing the distance matrix and clustering after injecting small Gaussian perturbations into the input trajectories and reporting the mean pairwise Adjusted Rand Index across five such perturbations.

2.5. Ensemble Consensus Clustering

The three models were combined into a single, more reliable partition by building a species-by-species co-association matrix, an approach that follows the general logic of ensemble [36] and consensus clustering established in recent bioinformatics work [17]. For every pair of strains p and q, the co-association value was calculated as:

$$c(p,q) = \frac{1}{M}\sum_{m=1}^{M} 1[label_m(p) = label_m(q)] \quad (3)$$

Here, M is the number of models, $label_m(p)$ is strain p's cluster label under model m and equals 1 when the condition is true and 0 otherwise. In plan terms, the equation counts how many of the three models placed a given pair of strains in the same cluster, giving a value of 1.0 when all three models agree and 0.0 when none do. The co-association matrix was converted to a distance (one minus co-association) and used as the input to average-linkage hierarchical clustering. The number of consensus clusters was selected by evaluating candidate values against a combined criterion of DTW-based silhouette score and mean per-strain co-association agreement, again under the cluster-balance constraint applied to the individual models. Strain pairs with co-association equal to one, meaning agreement across all three models, were distinguished from contested pairs with co-association between 0.3 and 0.7, so that confidence in individual cluster assignments could be communicated transparently.

2.6. Evaluation metrics & Biological Validation framework

Cluster quality was first assessed internally in the present study, using the silhouette score (calculation given in Table S2) and Davies-Bouldin index [37] in both Euclidean and DTW-distance space, together with a cluster-balance report quantifying the size of the largest cluster relative to the full strain set [35]. The silhouette score ranges from −1 to 1. Values close to 1 indicate strong cluster assignment, whereas values near 0 or below indicate overlap with neighboring clusters. [37]. Because internal validity metrics alone cannot establish that a clustering result is biologically meaningful, clusters were additionally validated against two sources of information withheld entirely from the clustering step itself [38].

2.6.1. Taxonomic and pigment validation metrics

Each strain was mapped to a coarse taxonomic group (phylum or class level), and agreement between cluster assignment and taxonomy was quantified using the Adjusted Rand Index and Normalized Mutual Information [39]. Cluster-wise differences in pigment indices (NDVI, GNDVI, Simple Ratio) were tested independently using the Kruskal-Wallis test (calculation given in Table S2) [40]. The test ranks every pigment value across all clusters together, then checks whether any one cluster's ranks are, on average, unusually high or low compared to the others; a large test statistic means the clusters are not just different by chance. Benjamini-Hochberg correction was then applied across the set of pigment comparisons to control the false discovery rate, and clusters containing fewer than three strains were flagged as statistically underpowered for the comparison rather than silently included at equal confidence [41].

2.6.2. Kinetic growth feature extraction

The species growth descriptors were derived from the growth-validation band, to describe each strain's growth behaviour with a small set of interpretable numbers. Area under the curve (AUC) was calculated using the trapezoidal rule (calculation given in Table S2), approximating the total signal accumulated over the cultivation period by treating the curve between each pair of measured days as a straight line and summing the resulting trapezoid areas [42]. The peak signal value and the day on which it occurred were also recorded, along with early- and late-phase growth rates, calculated as the slope of the trajectory before and after its temporal midpoint (calculation given in Table S2). Such descriptors were used both to test for significant separation of consensus clusters on an independent growth metric, and to characterize the timing of lag, maximum growth rate, and stationary phase for each strain, grouped by consensus cluster [43].

3. Results and Discussion

In the present analysis, three DTW-based clustering strategies were applied to group twenty microalgal strains by shared growth-kinetic behavior across six biochemical bands, and their outputs were combined into one consensus partition. The consensus clusters were then checked against taxonomy, pigment indices, and growth-curve timing, none of which were used to build the clusters themselves. The reliability of each model was also checked by testing how stable its output stayed across repeated runs and small changes to the input data.

3.1. Model performance comparison

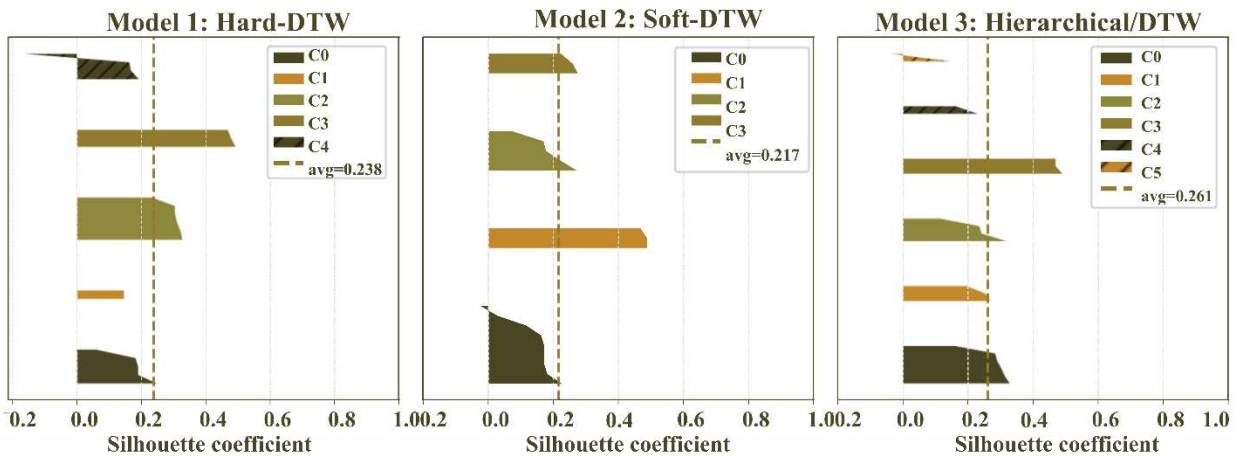


Fig. 3. DTW-based silhouette profile for each of the three clustering models. Hard-DTW (k = 5) reached silhouette = 0.204; Soft-DTW (k = 4, gamma = 2.330) reached 0.230; Hierarchical/DTW (k = 6, complete linkage) reached the highest value at 0.231, with perfect stability across repeated distance-matrix perturbations.

Each of the three clustering strategies was evaluated on DTW-based silhouette score, Davies-Bouldin index, and stability across repeated runs [44], as shown in Table 1. Hierarchical/DTW reached the highest DTW-based silhouette score (0.231) [15] and the highest possible stability (mean pairwise ARI = 1.000, since the method is deterministic given a fixed distance matrix), selecting k = 6 clusters under complete linkage. Soft-DTW reached a similar silhouette score (0.230) at k = 4 (gamma = 2.330), with high but not perfect stability (mean pairwise ARI = 0.835) across five random initializations. Hard-DTW reached the lowest DTW-based silhouette score of the three (0.204) at k = 5, though its own stability was still high (mean pairwise ARI = 0.914). Such results suggest Hierarchical/DTW gave the most internally consistent partition, but all three methods reached silhouette scores in a comparable, moderate range, rather than one model being dramatically better than the other two.

The pattern is consistent with a recent benchmark evaluation of elastic distance measures for time-series clustering, which reported that DTW-based *k*-means did not consistently outperform Euclidean-distance *k*-means across a wide range of general-purpose datasets, even after warping-window tuning as illustrated in (Fig. S2) [15]. The comparatively modest silhouette scores obtained here (0.204-0.231), together with Hard-DTW reaching the lowest score of the three

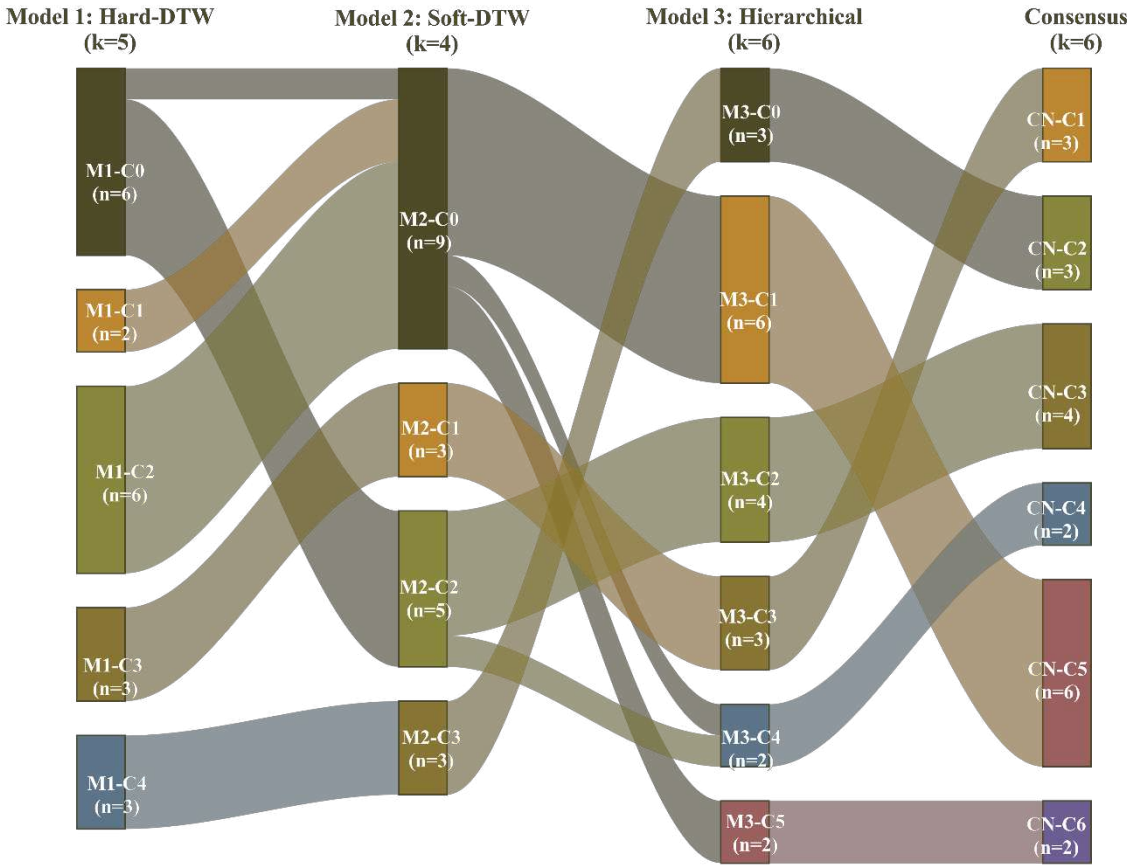


Fig. 4. Sankey diagram tracing each strain's cluster assignment across Model 1 (Hard-DTW, k = 5), Model 2 (Soft-DTW, k = 4), Model 3 (Hierarchical/DTW, k = 6), and the consensus partition (k = 5). Because the consensus partition matches Model 1 exactly, the diagram shows Model 1 and the consensus column as structurally identical, while Models 2 and 3 show genuine reassignment of strains such as SPIR, MESH, and PLAT between columns.

models tested, therefore appear consistent with a previously reported general property of DTW-based clustering rather than a limitation specific to the present hyperspectral dataset.

Table 1. Internal-validity comparison across the three clustering models. Silhouette score is reported in both DTW-distance space and flat Euclidean space; stability is the mean pairwise Adjusted Rand Index across five repeated runs (random seeds for Soft-DTW, Gaussian input perturbation for Hierarchical/DTW).

Si. No	Method	*k*	Silhouette (DTW)	Silhouette (Euclidean)	Davies Bouldin	Dominant Cluster fraction	Stability (mean Pairwise ARI)
1.	Hard-DTW *k*-means	5	0.204	0.238	1.278	0.30	0.914
2.	Soft-DTW *k*-means	4	0.230	0.217	1.357	0.45	0.835
3.	Hierarchical/DTW	6	0.231	0.261	1.031	0.30	1.000

3.2. Consensus cluster profiles and strain taxonomy

A co-association matrix was built to combine the three models into one partition, and cut into five consensus clusters, as shown in Fig. 3. This selection was made objectively: among candidate values of k from 2 to 8, k = 5 gave the best combined score of DTW silhouette and mean species-level agreement (combined score = 0.711),

with zero singleton clusters. The resulting consensus clusters were Cluster 1 (TETR, PLAT, NANO; n = 3), Cluster 2 (SPIR, PAVL, COEL, DESM, TEDE; n = 5), Cluster 3 (NOST, ANBA, CLAM, MESH; n = 4), Cluster 4 (DUNA, COCO, PICH, CHLO, APHA, GLEO; n = 6), and Cluster 5 (MONO, SCNE; n = 2).

The consensus partition should be interpreted with one caveat stated plainly. The pipeline logged a direct check confirming the consensus partition is identical (ARI = 1.000) to Model 1 (Hard-DTW)'s own partition, once cluster numbering is ignored. Model 2 (Soft-DTW) agreed with the consensus at ARI = 0.661, and Model 3 (Hierarchical/DTW) agreed at ARI = 0.850. The consensus procedure selected the Hard-DTW partition as the best-scoring candidate in the present analysis. Therefore, the consensus result was identical to the Hard-DTW partition rather than representing a distinct three-way blended partition. It is worth reporting rather than describing the results as three-way agreement, since the actual evidence for each cluster comes from checking which species pair were unanimous across all three models individually, not from the consensus label alone.

Twenty-eight species pairs were placed together by all three models without exception, making them the strongest similarity claims in the dataset: examples include APHA-CHLO, APHA-COCO, CHLO-DUNA, ANBA-CLAM, ANBA-NOST, COEL-DESM, COEL-PAVL, DESM-TEDE, MONO-SCNE, and NANO-PLAT, NANO-TETR, PLAT-TETR. A further 28 pairs were contested, with co-association between 0.3 and 0.7, and should be treated with caution; such pairs included COEL-SPIR, DESM-SPIR, PAVL-SPIR, and SPIR-TEDE (all at 0.67 agreement), alongside a larger block of pairs at 0.33 agreement linking MESH, MONO, and SCNE to several Cluster 4 strains as shown in (Fig. S1).

Mean trajectories were plotted for each consensus cluster across all six biochemical bands (Fig. 5) to characterize differences in growth behaviour among the fixed clusters. Cluster 2 (SPIR, PAVL, COEL, DESM, TEDE) reached the highest growth-curve area under the curve (AUC = 17.81) and the steepest early-phase growth rate (0.0309 signal units per day) of any consensus cluster. Cluster 1 (TETR, PLAT, NANO) reached the lowest AUC (4.62) and a negative early-phase slope (-0.0064), together with the earliest recorded day of peak signal (day 2), suggesting the cluster's signal declined early rather than accumulating over the cultivation period.

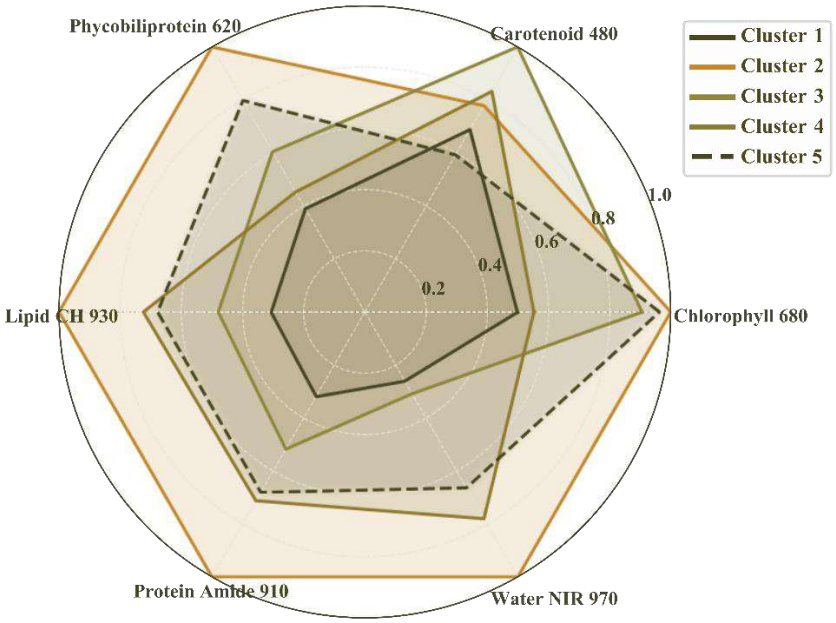


Fig. 5. Mean kinetic trajectory (± SD) for each of the five consensus clusters (Cluster 1: TETR, PLAT, NANO; Cluster 2: SPIR, PAVL, COEL, DESM, TEDE; Cluster 3: NOST, ANBA, CLAM, MESH; Cluster 4: DUNA, COCO, PICH, CHLO, APHA, GLEO; Cluster 5: MONO, SCNE), shown separately for all six biochemical bands. Cluster 2 shows the highest signal accumulation across bands, driven by both a fast early growth rate and a late peak (~day 23); Cluster 1 shows the lowest, with signal declining after an early peak (day 2) rather than accumulating over the cultivation period.

Cluster assignment was compared with each strain's taxonomic group (Fig. 6) to determine whether the consensus clusters corresponded to known taxonomic groupings. Agreement with taxonomy was weak (ARI = -0.045, NMI = 0.229, n = 20 species) [44]. A negative ARI means the consensus clusters agreed with taxonomy no better than chance, and slightly worse. The same pattern held across all three individual models as well, not just the consensus: Hard-DTW reached ARI = -0.045, Soft-DTW reached ARI = 0.019, and Hierarchical/DTW reached ARI = -0.009 (its NMI of 0.289 was the highest of the four, but its ARI was still effectively zero). Chlorophyta strains were distributed across every consensus cluster (2, 3, 2, 3, and 2 strains in Clusters 1 through 5 respectively) rather than concentrating in one, and Cyanobacteria strains were similarly split across Clusters 2, 3, and 4. Taxonomy was completely withheld from the clustering process. Therefore, the observed clustering indicates that the identified groups reflect growth-kinetic patterns that extend across taxonomic boundaries rather than simply reproducing known lineage relationships.

The divergence from taxonomy was notable given that prior hyperspectral studies on microalgae have mainly used spectral information to recover taxonomic identity rather than to group strains by shared growth behaviour: for instance, transmission hyperspectral microscopic imaging combined with PCA and SVM has been used to classify microalgae species with 94.4% accuracy and to estimate discrete growth stages with 98.1% accuracy ($R^2 = 0.998$) using Random Forest [9]. That level of accuracy for taxonomy-aligned classification stands in contrast to the near-chance ARI obtained here, underscoring that growth-kinetic similarity and taxonomic identity are measuring genuinely different properties of a strain, rather than one being a proxy for the other. However, this finding alone does not establish the biological relevance of the clusters. Their biological validity was therefore further evaluated using independent pigment and growth data in Section 3.3.

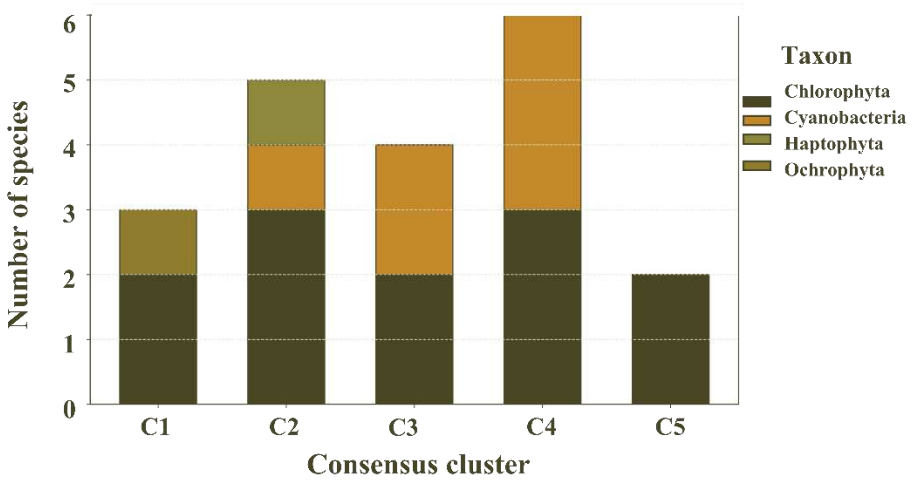


Fig. 6. Taxon composition of each consensus cluster (Cluster 1: TETR, PLAT, NANO; Cluster 2: SPIR, PAVL, COEL, DESM, TEDE; Cluster 3: NOST, ANBA, CLAM, MESH; Cluster 4: DUNA, COCO, PICH, CHLO, APHA, GLEO; Cluster 5: MONO, SCNE). Every consensus cluster contains a mix of Chlorophyta and, in most cases, Cyanobacteria strains, rather than a single taxon dominating any one cluster Chlorophyta strains are spread across all five clusters and Cyanobacteria strains across three, showing that growth-kinetic grouping cuts across taxonomic lineage rather than reproducing it (ARI = -0.045, NMI = 0.229).

3.3. Independent Biological & Statistical Validation

3.3.1. Pigment trajectory separation

NDVI, GNDVI, and growth-curve descriptors were compared across clusters using the Kruskal-Wallis test (Fig. 7) to independently evaluate the consensus clusters using variables withheld from clustering. Clusters differed significantly in NDVI ($H = 14.59$, $p = 0.0056$), GNDVI ($H = 12.43$, $p = 0.0144$), and Simple Ratio ($H = 13.28$, $p = 0.0100$). Growth-curve AUC also differed significantly across clusters ($H = 15.52$, $p = 0.0037$), as did peak signal value ($H = 13.49$, $p = 0.0091$) and early-phase growth rate ($H = 14.77$, $p = 0.0052$). Day of peak signal ($p = 0.0707$) and late-phase growth rate ($p = 0.0724$) did not reach significance. Six of the eight metrics remained significant at $q < 0.05$ after Benjamini-Hochberg correction across all eight tests: NDVI, GNDVI, Simple Ratio,

AUC, peak value, and early-phase growth rate. Because the smallest consensus cluster contained only two strains (MONO, SCNE), such p-values should be interpreted as exploratory rather than confirmatory. A larger strain panel would be required to assess the robustness of the differences.

Vegetation-style spectral indices, including NDVI and GNDVI, have previously been applied for non-invasive monitoring of microalgal cultivation using hyperspectral imaging [24]. In that study, these indices were used to characterize changes in culture status over time within individual cultures. In the present study, NDVI and GNDVI were instead evaluated across five growth-kinetic clusters comprising twenty strains and tested for differences among the resulting groups. Their use as independent validation variables provides an additional assessment of whether the clusters identified from hyperspectral growth trajectories correspond to biologically distinguishable patterns, rather than using the indices solely for descriptive monitoring.

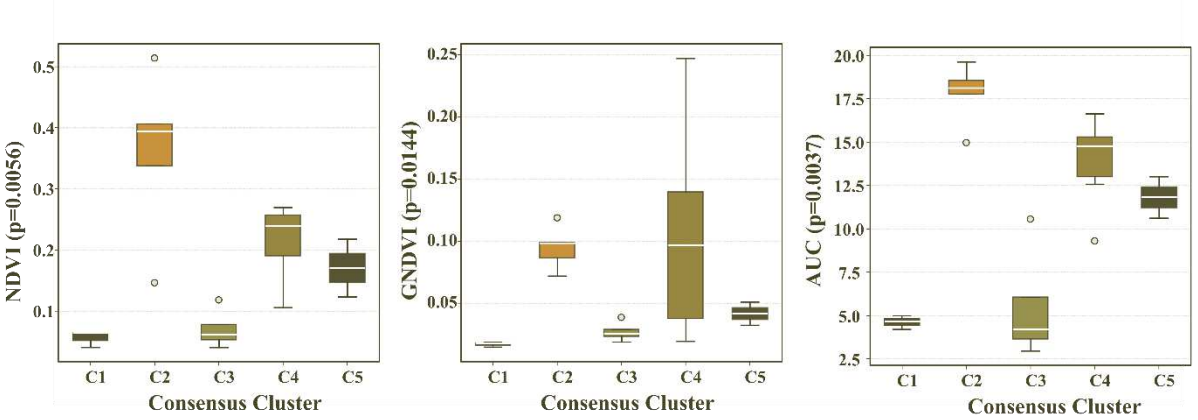


Fig. 7. Boxplots of NDVI, GNDVI, and growth-curve AUC across the five consensus clusters, with Kruskal-Wallis p-values in each panel title. Cluster 2 shows the highest NDVI and growth-curve AUC, while Cluster 4 shows the highest GNDVI; Cluster 1 shows the lowest values across all three metrics. None of the three metrics were used during clustering, so this separation reflects an independent biological difference between clusters rather than a restatement of the clustering criterion.

3.3.2. Bio-kinetic growth dynamics

Day of peak signal was compared across consensus clusters (Fig. 8) to characterize differences in growth timing rather than final signal magnitude. Cluster 1 (TETR, PLAT, NANO) peaked earliest, at a mean day of 2.0, with zero variance across its three strains, indicating an early and consistent decline pattern rather than sustained growth. Cluster 3 (NOST, ANBA, CLAM, MESH) peaked latest, at a mean day of 25.25. Clusters 2, 4, and 5 peaked at intermediate days of 23.4, 24.7, and 21.5 respectively. The observed variation in peak timing, from day 2 to day 25 across the 31-day cultivation period, indicates that the consensus clusters are associated with distinct growth-kinetic patterns rather than differences only in final signal levels.

Prior time-series studies of microalgal cultivation have primarily focused on biomass forecasting for individual strains [18], providing limited insight into differences in growth timing across multiple strains cultivated under comparable conditions. In contrast, the present study compares peak-signal timing across five clusters comprising twenty strains, providing a multi-strain perspective on variation in growth kinetics and potential harvest windows. The temporal variation is relevant to process economics because harvest timing can influence biomass composition and production costs. For example, techno-economic analysis of a *Dunaliella sp*.-based process identified late-exponential-phase harvesting as the most cost-favourable scenario among the conditions evaluated [45]. While that study considered harvest timing for a single strain, the day -2-to-day-25 variation in peak-signal timing observed across the present clusters indicates that a common harvest day may not be equally appropriate for all growth-kinetic groups. Thus, cluster-specific growth trajectories could provide a basis for evaluating whether harvest timing can be better aligned with strain-group-specific cultivation behaviour.

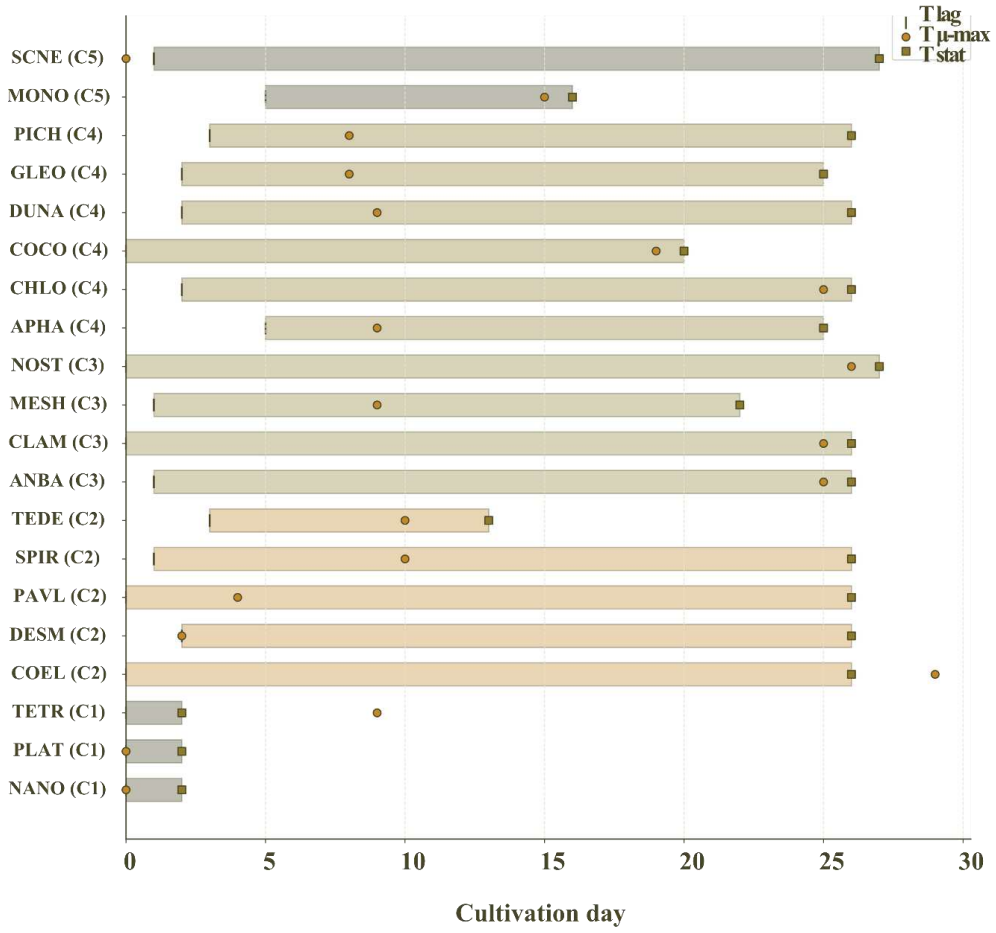


Fig. 8. Day of peak signal for every strain, grouped by consensus. Cluster 1 peaks on day 2 with no variance across its three strains, consistent with the early decline pattern rather than sustained growth; the remaining four clusters peak later and more variably, between a mean of day 21.5 (Cluster 5) and day 25.25 (Cluster 3) of the 31-day cultivation period. This spread in peak timing shows the consensus clusters differ in growth-kinetic phenotype, not only in final signal magnitude.

3.4. Implications for biomanufacturing process engineering

The present study combines three DTW-based clustering strategies with independent biological validation, enabling the resulting consensus clusters to be evaluated using growth-related measurements in addition to internal cluster-validity metrics. Six of eight independent pigment and growth-curve metrics differed significantly across clusters after multiple-testing correction. These findings provide evidence that the identified clusters are associated with biologically distinguishable growth and pigment characteristics. Although the consensus partition in the current run matched Model 1 exactly rather than blending all three models, and although the smallest cluster contained only two strains, the spread of peak-signal timing from day 2 to day 25 still points to genuinely different growth-kinetic phenotypes across the strain panel. Future research should prioritize controlled harvest-validation trials that compare the predicted optimal harvest day for each consensus cluster with biomass or pigment yield. Such experiments would provide direct evidence for the utility of the clustering framework in harvest-timing decisions. Addressing this, along with a larger strain panel and a repeated consensus check across multiple training runs, would extend the value of the approach beyond the twenty strains tested here. Moreover, applying the same pipeline to a reduced set of biochemical bands could clarify whether a smaller, cheaper optical sensor package would retain enough signal to support the same harvest-timing decisions. The application of DTW-based growth-

413

414 **Table 2.** Related studies applying computational methods to microalgae phenotyping and time-series clustering

Si.No.	Focus	Data	Method(s)	Quantitative result	Ref.
1	Species classification and growth-stage estimation from hyperspectral microscopy	Transmission hyperspectral microscopic images; 3 microalgae species	PCA + SVM (classification); Random Forest (growth stage)	SVM: 94.4% accuracy; Random Forest growth-stage prediction: 98.1% accuracy ($R^2 = 0.998$)	[9]
2	Non-invasive cultivation monitoring using hyperspectral imaging	Hyperspectral imager; green microalgae cultivations	Vegetation-style spectral indices vs. manual measurement	Reported alongside Xu et al.'s 94.4% classification benchmark as supporting evidence for non-destructive monitoring	[24]
3	Comparative ML/DL image classification for autotrophic batch cultivation	Microscope images of cultivated microalgae	SVM, k-NN, CNN	Best model (SVM): 97.93% accuracy	[46]
4	Species identification using microalgal mucilage in marine ecosystems	Light microscopy and SEM images; 6 classes	MobileNetV2, GoogleNet	Best model (MobileNetV2): 87.04% accuracy; F1 = 0.87	[11]
5	Classification of cyanobacteria and chlorophyta microalgae groups	Microscopic images	CNN-SVM hybrid, GoogleNet	Best model (GoogleNet): 98.98% accuracy; F1 = 0.97	[12]
6	Environmental microorganism classification	Fluorescent microscopy images; 21 classes	ResNet50, VGG16, Inception-v3	Best model (Inception-v3): 92.9% accuracy	[47]
7	Classification of dead and living *Chlorella vulgaris* cells	Optical microscope images	Bioimage Informatics + Inception-v3	Inception-v3: 95% accuracy	[48]
8	Strain selection for mass production based on lipid/fatty-acid profiles	Tropical microalgae strain panel	Comparative lipid and fatty-acid profiling	Identified high-lipid strains suitable for scale-up; qualitative and compositional results without a single accuracy metric	[2]

Si.No.	Focus	Data	Method(s)	Quantitative result	Ref.
9	Benchmark evaluation of elastic distance measures for time-series clustering	UCR time-series archive; general benchmark, not algae-specific	DTW, Soft-DTW, and seven other elastic distances with k-means/k-medoids	DTW-based *k*-means did not outperform Euclidean-distance *k*-means, even after warping-window tuning	[15]
10	Comprehensive review of *k*-means clustering variants	Cross-domain review; algorithmic, not algae-specific	*k*-means and major variants	Synthesized big-data-era advances and limitations of *k*-means-family algorithms	[49]

kinetic clustering, combined with independent biological validation [7], ultimately points toward strain-aware harvest scheduling as a practical route to reducing yield loss in microalgae biomanufacturing [2].

4. Conclusion

In the present study, three DTW-based clustering strategies, Hard-DTW *k*-means, Soft-DTW *k*-means, and agglomerative Hierarchical/DTW clustering, were applied to hyperspectral growth-kinetic trajectories from twenty microalgal and cyanobacterial strains. The comparison across the three models showed that Hierarchical/DTW reached the highest internal cluster validity (DTW-silhouette = 0.231, Davies-Bouldin = 1.031) among the three, with all models reaching silhouette scores in a broadly comparable range (0.204-0.231). Hard-DTW and Hierarchical/DTW maintained high stability across repeated perturbations (mean pairwise ARI of 0.914 and 1.000, respectively), while Soft-DTW remained reasonably stable at 0.835. Notably, the resulting consensus partition, despite matching the Hard-DTW partition exactly in the current run rather than reflecting a genuinely new three-way blend, still separated the twenty strains into five clusters that differed significantly on six independent pigment and growth-curve metrics withheld from clustering. The developed pipeline may provide a basis for strain-specific harvest-timing decisions in microalgal biomanufacturing, provided that predicted harvest windows are validated through controlled harvest trials and the approach is evaluated using a larger strain panel.

Abbreviations

ARI	Adjusted Rand Index
AUC	Area Under the Curve
BB	Bold Basal (medium)
BG-11	Blue-Green medium 11 (standard cyanobacterial growth medium)
CMFRI	Central Marine Fisheries Research Institute
CNN	Convolutional Neural Network
DTW	Dynamic Time Warping
GNDVI	Green Normalized Difference Vegetation Index
GPU	Graphics Processing Unit
HSI	Hyperspectral Imaging
ICAR	Indian Council of Agricultural Research
k-NN	k-Nearest Neighbors
NCIM	National Collection of Industrial Microorganisms
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
NMI	Normalized Mutual Information
NRMC-F	National Repository for Microalgae and Cyanobacteria-Freshwater

PCA	Principal Component Analysis
RAM	Random Access Memory
ROI	Region of Interest
SD	Standard Deviation
SNV	Standard Normal Variate
SR	Simple Ratio
SVM	Support Vector Machine
UCR	University of California, Riverside (UCR Time Series Archive)

Data availability

The strain sources and cultivation conditions, the complete hyperspectral dataset (1,860 images; 37,200 spectral readings), and the code files used for trajectory-tensor construction, preprocessing, clustering, stability analysis, consensus clustering, and biological validation. Complete results for the three DTW-based clustering models, including internal validity, stability, co-association, consensus selection, taxonomy comparisons, and statistical validation of eight pigment and growth-curve metrics, are also provided.

Acknowledgement

Rajalakshmi Elumalai acknowledges the University Research Fellowship (URF) received from SRM University-AP to conduct this work.

Funding

This work acknowledges the support from SRM University-AP through the SRM University-AP Research Grant under Grant No. SRMAP/URG/ GENERAL-D/2025-26/024 and SRM/TRUST/AP/Jan/IC/2223/PUR-00007.

Author Information

Authors and Affiliations

The NET ZERO Lab, Department of Environmental Science and Engineering, SRM University-AP, Amaravati, 522 240, Andhra Pradesh, India

Rajalakshmi Elumalai & Karthik Rajendran

Department of Electronics and Communication Engineering, SRM University-AP, Amaravati, 522 240, Andhra Pradesh, India

Karthikeyan Meenatchi Sundaram & Sunil Chinnadurai

Contributions

Rajalakshmi Elumalai: Conceptualization, Data curation, Methodology, Data Curation, Software, Visualization, Writing – Original Draft. Karthikeyan Meenatchisundaram: Methodology, Data curation, Validation, Software. Sunil Chinnadurai: Data curation, Investigation. Ankita Juneja: Investigation. Karthik Rajendran: Investigation, Supervision, Writing – review & editing.

Corresponding author

Correspondence to rajendran.k@srmap.edu.in

Ethics declarations

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References:

[1] F. Camacho, A. Macedo, and F. Malcata, "Potential industrial applications and commercialization of microalgae in the functional food and feed industries: A short review," *Mar. Drugs*, vol. 17, no. 6, 2019, doi: 10.3390/md17060312.

[2] A. R. Andrew, W. T. L. Yong, M. Misson, A. Anton, and G. J. W. L. Chin, "Selection of Tropical Microalgae Species for Mass Production Based on Lipid and Fatty Acid Profiles," *Front. Energy Res.*, vol. 10, no. June, pp. 1–10, 2022, doi: 10.3389/fenrg.2022.912904.

[3] L. Rodolfi *et al.*, "Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor," *Biotechnol. Bioeng.*, vol. 102, no. 1, pp. 100–112, 2009, doi: 10.1002/bit.22033.

[4] L. M. L. Laurens *et al.*, "Acid-catalyzed algal biomass pretreatment for integrated lipid and carbohydrate-based biofuels production," *Green Chem.*, vol. 17, no. 2, pp. 1145–1158, 2015, doi: 10.1039/c4gc01612b.

[5] M. G. Savvidou, A. Ferraro, E. Hristoforou, D. Mamma, D. Kekos, and F. N. Kolisis, "Incorporation of Magnetic Nanoparticles into Protoplasts of Microalgae Haematococcus pluvialis: A Tool for Biotechnological Applications," *Molecules*, vol. 25, no. 21, 2020, doi: 10.3390/molecules25215068.

[6] M. J. Griffiths, C. Garcin, R. P. van Hille, and S. T. L. Harrison, "Interference by pigment in the estimation of microalgal biomass concentration by optical density," *J. Microbiol. Methods*, vol. 85, no. 2, pp. 119–123, 2011, doi: 10.1016/j.mimet.2011.02.005.

[7] B. Guieysse and M. Plouviez, "Microalgae cultivation: closing the yield gap from laboratory to field scale," *Front. Bioeng. Biotechnol.*, vol. 12, no. February, pp. 1–7, 2024, doi: 10.3389/fbioe.2024.1359755.

[8] B. Chu *et al.*, "Visible/Short-wave near-infrared hyperspectral analysis of lipid concentration and fatty acid unsaturation of Scenedesmus obliquus in situ," *Comput. Electron. Agric.*, vol. 182, no. February, p. 106046, 2021, doi: 10.1016/j.compag.2021.106046.

[9] Z. Xu, Y. Jiang, J. Ji, E. Forsberg, Y. Li, and S. He, "Classification, identification, and growth stage estimation of microalgae based on transmission hyperspectral microscopic imaging and machine learning," *Opt. Express*, vol. 28, no. 21, p. 30686, 2020, doi: 10.1364/oe.406036.

[10] J. W. R. Chong *et al.*, "Artificial intelligence-driven microalgae autotrophic batch cultivation: A comparative study of machine and deep learning-based image classification models," *Algal Res.*, vol. 79, no. January, p. 103400, 2024, doi: 10.1016/j.algal.2024.103400.

[11] M. E. Sonmez, N. E. Gumus, N. Eczacioglu, E. E. Develi, K. Yücel, and H. B. Yildiz, "Enhancing microalgae classification accuracy in marine ecosystems through convolutional neural networks and support vector machines," *Mar. Pollut. Bull.*, vol. 205, no. March, p. 116616, 2024, doi: 10.1016/j.marpolbul.2024.116616.

[12] M. E. Sonmez, N. Eczacıoglu, N. E. Gumuş, M. F. Aslan, K. Sabanci, and B. Aşikkutlu, "Convolutional neural network - Support vector machine based approach for classification of cyanobacteria and chlorophyta microalgae groups," *Algal Res.*, vol. 61, no. November 2021, 2022, doi: 10.1016/j.algal.2021.102568.

[13] P. Salmi, M. A. Eskelinen, M. T. Leppänen, and I. Pölönen, "Rapid quantification of microalgae growth with hyperspectral camera and vegetation indices," *Plants*, vol. 10, no. 2, pp. 1–12, 2021, doi: 10.3390/plants10020341.

[14] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognit.*, vol. 44, no. 3, pp. 678–693, 2011, doi: 10.1016/j.patcog.2010.09.013.

[15] C. Holder, M. Middlehurst, and A. Bagnall, *A review and evaluation of elastic distance functions for time series clustering*, vol. 66, no. 2. Springer London, 2024. doi: 10.1007/s10115-023-01952-0.

[16] K. Zhang, S. Lin, H. Sun, L. Ma, and J. Xu, "Dynamic Time Warping Based Clustering for Time Series Analysis," *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 415 LNICST, pp. 376–385, 2022, doi: 10.1007/978-3-030-94182-6_29.

[17] Y. Shan, S. Li, F. Li, Y. Cui, and M. Chen, "Dual-level clustering ensemble algorithm with three consensus strategies," *Sci. Rep.*, vol. 13, no. 1, pp. 1–19, 2023, doi: 10.1038/s41598-023-49947-9.

[18] K. Meenatchi Sundaram, D. Kumar, J. Lee, S. Barathi, and K. Rajendran, "Time series forecasting of microalgae cultivation for a sustainable wastewater treatment," *Process Saf. Environ. Prot.*, vol. 196, Apr. 2025, doi: 10.1016/j.psep.2025.106845.

[19] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," vol. 12, pp. 2825–2830, 2011.

[20] R. Tavenard *et al.*, "Tslearn, A Machine Learning Toolkit for Time Series Data," *J. Mach. Learn. Res.*, vol. 21, pp. 1–6, 2020.

[21] R. Y. Stanier, R. Kunisawa, M. Mandel, and G. Cohen-Bazire, "Purification and properties of unicellular blue-green algae (order Chroococcales).," *Bacteriol. Rev.*, vol. 35, no. 2, pp. 171–205, 1971, doi: 10.1128/mmbr.35.2.171-205.1971.

[22] R. Rippka, J. Deruelles, and J. B. Waterbury, "Generic assignments, strain histories and properties of pure cultures of cyanobacteria," *J. Gen. Microbiol.*, vol. 111, no. 1, pp. 1–61, 1979, doi: 10.1099/00221287-111-1-1.

[23] Culture Collection of Algae and Protozoa, "3N-BBM+V (Bold Basal Medium with 3-fold Nitrogen and Vitamins; modified)," vol. 12, no. 1, pp. 1–2, 2016, [Online]. Available: http://www.ccap.ac.uk/media/documents/3N_BBM_V.pdf

[24] S. Pääkkönen *et al.*, "Non-invasive monitoring of microalgae cultivations using hyperspectral imager," *J. Appl. Phycol.*, vol. 36, no. 4, pp. 1653–1665, 2024, doi: 10.1007/s10811-024-03256-4.

[25] M. Schmid, D. Rath, and U. Diebold, "Why and How Savitzky-Golay Filters Should Be Replaced," *ACS Meas. Sci. Au*, vol. 2, no. 2, pp. 185–196, 2022, doi: 10.1021/acsmeasuresciau.1c00054.

[26] F. J. Esquivel, J. L. Romero-Béjar, and J. A. Esquivel, "Preprocessing of spectroscopic data to highlight spectral features of materials," *Anal. Sci. Adv.*, vol. 5, no. 9–10, pp. 14–18, 2024, doi: 10.1002/ansa.202400018.

[27] C. Yan, "A review on spectral data preprocessing techniques for machine learning and quantitative analysis," *iScience*, vol. 28, no. 7, p. 112759, 2025, doi: 10.1016/j.isci.2025.112759.

[28] R. Bentivogli, L. Pezzolesi, N. Caputo, B. Casarotto, and S. Silvestri, "Laboratory-based hyperspectral reflectance analysis for phytoplankton species identification," *Ecol. Inform.*, vol. 94, no. January, p. 103626, 2026, doi: 10.1016/j.ecoinf.2026.103626.

[29] A. Aizpuru and A. González-Sánchez, "Traditional and new trend strategies to enhance pigment contents in microalgae," *World J. Microbiol. Biotechnol.*, vol. 40, no. 9, pp. 1–26, 2024, doi: 10.1007/s11274-024-04070-3.

[30] D. Davydov *et al.*, "Water, Collagen, and Lipid Content in the Human Skin and Muscles Assessed with Near-Infrared Diffuse Reflectance Spectroscopy and Multi-Spectral Optoacoustic Tomography," *Adv. Sci.*, vol. 12, no. 41, pp. 1–13, 2025, doi: 10.1002/advs.202505619.

[31] D. Carpenter, E. Deyle, J. Park, E. Saberski, and G. Sugihara, "Repairing gaps in ecological time series," *Methods Ecol. Evol.*, vol. 16, no. 3, pp. 573–583, 2025, doi: 10.1111/2041-210X.14491.

[32] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[33] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

[34] M. Herrmann, C. W. Tan, and G. I. Webb, “Parameterizing the cost function of dynamic time warping with application to time series classification,” *Data Min. Knowl. Discov.*, vol. 37, no. 5, pp. 2024–2045, 2023, doi: 10.1007/s10618-023-00926-8.

[35] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[36] S. Coleman, P. D. W. Kirk, and C. Wallace, “Consensus clustering for Bayesian mixture models,” *BMC Bioinformatics*, vol. 23, no. 1, 2022, doi: 10.1186/s12859-022-04830-8.

[37] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, “The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters,” *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.3309.

[38] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, 2013, doi: 10.1016/j.patcog.2012.07.021.

[39] L. Hubert and P. Arabie, “Comparing partitions,” *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985, doi: 10.1007/BF01908075.

[40] L. Yang, P. Wang, and J. Chen, “2dGBH: Two-dimensional group Benjamini–Hochberg procedure for false discovery rate control in two-way multiple testing of genomic data,” *Bioinformatics*, vol. 40, no. 2, 2024, doi: 10.1093/bioinformatics/btae035.

[41] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[42] Q. Dai, D. W. Sun, Z. Xiong, J. H. Cheng, and X. A. Zeng, “Recent advances in data mining techniques and their applications in hyperspectral image processing for the food industry,” *Compr. Rev. Food Sci. Food Saf.*, vol. 13, no. 5, pp. 891–905, 2014, doi: 10.1111/1541-4337.12088.

[43] F. S. Midani, J. Collins, and R. A. Britton, “AMiGA: Software for Automated Analysis of Microbial Growth Assays,” *mSystems*, vol. 6, no. 4, pp. 1–14, 2021, doi: 10.1128/msystems.00508-21.

[44] M. Gagolewski, “Normalised Clustering Accuracy: An Asymmetric External Cluster Validity Measure,” *J. Classif.*, vol. 42, no. 1, pp. 2–30, 2025, doi: 10.1007/s00357-024-09482-2.

[45] Y. Sui, Y. Jiang, M. Moretti, and S. E. Vlaeminck, “Harvesting time and biomass composition affect the economics of microalgae production,” *J. Clean. Prod.*, vol. 259, p. 120782, 2020, doi: 10.1016/j.jclepro.2020.120782.

[46] J. W. R. Chong *et al.*, “Artificial intelligence-driven microalgae autotrophic batch cultivation: A comparative study of machine and deep learning-based image classification models,” *Algal Res.*, vol. 79, no. February, p. 103400, 2024, doi: 10.1016/j.algal.2024.103400.

[47] C. M. Liang, C. C. Lai, S. H. Wang, and Y. H. Lin, “Environmental microorganism classification using optimized deep learning model,” *Environ. Sci. Pollut. Res.*, vol. 28, no. 24, pp. 31920–31932, 2021, doi: 10.1007/s11356-021-13010-9.

[48] R. Reimann *et al.*, “Classification of dead and living microalgae Chlorella vulgaris by bioimage informatics and machine learning,” *Algal Res.*, vol. 48, no. December 2019, p. 101908, 2020, doi: 10.1016/j.algal.2020.101908.

[49] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI1.pdf
- SI2.pdf