VOL. 25, NO. 1, 2026, 35-45
elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

# A Robust Algae Biomass Growth Rate Estimation using RGB Imaging

**Keshinro Kazeem Kolawole**[1,3], **Mohamad Shukri bin Zainal Abidin**[1*], **Mohd Farizal bin Kamaroddin**[2], **Muhammad Sharul Azwan bin Ramli**[1], **Sikudhan Lucas Mpuhus**[1] and **Ardiansyah Rizqi**[1]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.
[2]Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.
[3]College of Engineering and Technology, Lagos State University of Science and Tech (LASUSTECH), Ikorodu, Lagos, Nigeria.

[*]Corresponding author: shukri@utm.my

**Abstract:** The growing demand for sustainable bioresources has spurred research into algae biomass estimation, a crucial aspect of biofuel production, carbon capture, and wastewater treatment. Despite the development of several methods to estimate algae biomass, a lack of standardized, scalable, and non-invasive biomass estimation models for outdoor environments persists. This study introduces a novel method that utilizes digital image-based RGB analysis of image models, integrated with machine learning (ML) algorithms. Traditional biomass estimation methods typically involve spectrophotometric or chemical analyses, which are labour-intensive and expensive. In contrast, the proposed approach employs low-cost RGB imaging, which enables real-time biomass quantification through image-based analysis. The research process involves data collection from controlled algae cultivation experiments, image preprocessing, and feature extraction using advanced computer vision techniques applied to *Chlorella sorokiniana* cultivated outdoors. Key features, such as colour indices, texture patterns, and pixel intensity distributions, were extracted from the RGB images. Various ML models, including Random Forest regressor (RFR), Extreme Gradient Boosting regressor (XGBR), and Convolutional Neural Networks (CNNs), were trained and validated to predict algal biomass concentrations. Experimental results demonstrated that the ML models accurately predicted algae biomass, with a correlation coefficient ($R^2$) exceeding 90% in test datasets, showcasing the robustness of the proposed framework. Future research will explore multispectral extensions, adaptive ML models for various algal species, and deployment in real-world industrial settings.

**Keywords:** algae, biomass, estimation, image processing



*Article History: received 28 January 2025; accepted 2 December 2025; published 30 April 2026*
*Digital Object Identifier 10.11113/elektrika.v25n1.664*

## 1. INTRODUCTION

Microalgae cultivation has emerged as a critical solution for sustainable biofuel production, carbon sequestration, and wastewater treatment [18]. However, a persistent challenge in large-scale cultivation is the lack of standard, non-destructive, real-time biomass monitoring methods [5]. Traditional techniques, such as manual subsampling, optical density (OD) measurements, and dry-weight analysis, are labor-intensive, destructive, and impractical for industrial-scale operations [20]. Recent improvements in remote sensing and machine learning (ML) provide hopeful new options by using measurements like the Normalized Difference Index (NDI) and RGB vegetation index (RGBVI) [16]. While these indices are widely used in terrestrial and aquatic vegetation monitoring, their application to land-based algal systems remains underexplored, particularly for density-per-area estimation. Existing studies focus on oceanic or shallow-water algae, where variable water chemistry and depth complicate reflectance signals [3].

While RGB imaging is a valuable way to estimate biomass on a large scale, other optical methods like Raman spectroscopy and fluorescence imaging can also help monitor algae in different ways (Microalgal biomass quantification from the non-invasive technique of image processing through red–green–blue (RGB) analysis [12]. Raman spectroscopy is a method that doesn't require labels and provides detailed information about molecules by measuring how light scatters, which helps identify specific biomolecules like lipids, carotenoids, and chlorophyll in algal cells. This method is particularly valuable for tracking biochemical composition shifts during growth phases or stress conditions [15]. However, its high equipment cost and sensitivity to environmental noise limit its use in large-scale field applications.

Fluorescence imaging, on the other hand, exploits the natural autofluorescence of chlorophyll and other pigments to assess photosynthetic activity and cell health. Techniques like chlorophyll fluorescence (e.g., Pulse-Amplitude Modulation (PAM) can quantify photosynthetic efficiency and stress responses in real-time [10]. When combined with multispectral data, fluorescence imaging improves the detail of physiological

assessments, providing information about algal productivity beyond just measuring biomass density [19].

Despite their potential, these methods are often constrained by cost, complexity, or scalability compared to multispectral imaging. This study aims to combine Mapir RGB imaging with machine learning to accurately estimate biomass density in Chlorella cultures while keeping costs low and ensuring it can be used in real-time. By validating the model against ground truth data, we demonstrate its suitability for large-scale aquaculture while acknowledging future opportunities to hybridize RGB data with multispectral, Raman, or fluorescence techniques for deeper biochemical insights.

This study addresses these gaps by developing a scalable biomass estimation model using low-cost Mapir RGB (red, green, blue) images and machine learning. By

1. Developing an image-processing algorithm for algae growth rate estimation.
2. Validating against conventional methods.
3. Assessing robustness across species and conditions
4. Compare different statistical descriptors (mean, median, mode).
5. Integrate machine learning for predictive enhancement.

Our method offers an affordable, immediate solution for algae farms. It requires data accuracy, using RGB images and lab-tested biomass measurements in outdoor cultivation of Sorokiniana (a high-lipid Chlorophyta), and frequent biomass monitoring to optimize harvest timing.

### 1.1 Related Works

Several techniques exist for estimating algal biomass, each with distinct advantages and limitations. The gravimetric (dry weight) method remains the gold standard due to its high accuracy, but it is labor-intensive and unsuitable for real-time monitoring. Optical density (OD) measurement offers rapid, nondestructive estimations but saturates at high cell densities and requires regular calibration. Chlorophyll fluorescence techniques provide sensitive insights into the physiological state of algae but are indirect and influenced by environmental stress. Advanced spectral approaches such as multispectral and hyperspectral imaging deliver detailed pigment profiling with strong biomass correlation but are costly and computationally intensive. In contrast, RGB imaging combined with machine learning offers a low-cost, scalable, and field-deployable solution that leverages simple color features to predict biomass. While sensitive to lighting conditions and sensor variabilities, it stands out for its potential in outdoor applications, especially when calibrated against conventional methods such as dry weight or optical density (OD).

Recent RGB-based approaches for non-invasive algae biomass estimation have advanced significantly through integrating digital imaging, feature extraction, and machine learning. These methods capitalize on the cost-effectiveness and scalability of standard RGB cameras to estimate algal biomass without physical sampling or destructive analysis. The core principle involves capturing top-view images of algal cultures in photobioreactors or open ponds and analyzing the intensity and distribution of red, green, and blue channels, which correlate with pigment concentrations, particularly chlorophyll-a.

Studies such as Wasonga et al. [23], and Salgueiro et al. [12] demonstrated that RGB descriptors (mean, median, and mode values of each channel) can be linearly or non-linearly regressed against biomass metrics like dry cell weight (DCW) or optical density [23]. The green channel, strongly influenced by chlorophyll content, often shows the highest correlation with biomass during exponential growth. More recent innovations have applied machine learning models—such as random forests, support vector machines, and convolutional neural networks (CNNs)—to learn complex patterns between RGB data and biomass under varying environmental conditions. These models improve estimation robustness across lighting conditions, growth phases, and reactor geometries.

Furthermore, some studies now incorporate temporal imaging, where RGB features are tracked across days to dynamically model growth rates and phase transitions. Reinforcement and transfer learning are also being explored to enable adaptive calibration across algal strains and cultivation environments. Despite their promise, RGB-based approaches require standardized image acquisition protocols and careful correction for ambient lighting variability. Nonetheless, they represent a promising frontier for real-time, non-invasive biomass monitoring in research and industrial algal cultivation settings.

## 2. MATERIALS AND METHODS

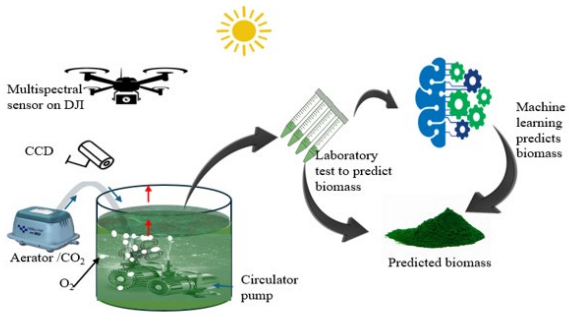
### 2.1 Experimental Design and Data Gathering

Outdoor cultivation of Chlorella sorokiniana species requires field adaptation. The inoculant (with OD > 1) is prepared in the field for nearly seven days before transfer to the main chambers. After transfer, cells enter a lag phase for acclimation. This is followed by exponential growth, marked by rapid division and autospore formation. Growth slows in the stationary phase as resources deplete, balancing division and death. Under stress, algae may form resting cysts, reactivating when conditions improve (Coşgun et al., 2021).

This study cultivated green microalgae at a field station in Johor, Malaysia, under tropical conditions (25–32°C, high humidity). The RGB camera captured chlorophyll-sensitive wavelengths at consistent angles and distances, while traditional spectrophotometry validated measurements, as shown in Figure 1(a) and (b). To minimize artifacts, preprocessing included brightness normalization, spectral calibration, and fixed ROI extraction. Vegetation indices (e.g., NDI, RGBVI) and texture features (entropy, spatial frequency) were analyzed to enhance machine learning models. Daily imaging occurred from 9:00 to 9:00–12:00 (GMT+8), with sensor calibration using a white plate. A weather station monitored environmental factors, and a transparent shelter protected cultures while allowing sunlight.

The study assessed biomass production in a 1000L tank using ImageJ ROI analysis, correlating spectral data with growth rates, as shown in Figure 1(c). Nutrients included urea (60g/tank) and chicken excreta (12g in 30L water, 1L/day added). Over 30 days, parameters like temperature,

pH, NDI, and biomass were tracked alongside weather data (wind, solar radiation). The RGB camera captured colour intensities, aiming to establish a non-invasive monitoring method for large-scale cultivation.




(a)

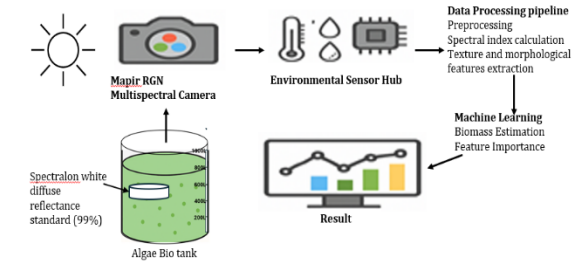
(b)

(c)

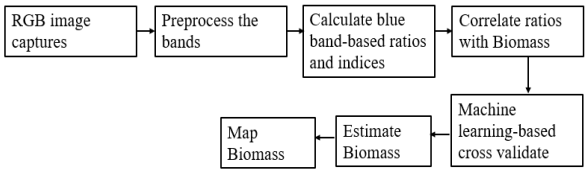
Figure 1. (a) System setup (b) Weather station monitor (c) ROI measurement in ImageJ

### 2.2 Image Processing

The suggested system uses detailed data from a MAPIR RGB camera without requiring outside light measurements like PAR (photosynthetically active radiation) sensors to improve the estimation of algae biomass in outdoor settings. Instead of relying on normalized solar irradiance with PAR, the system utilizes image-derived spectral indices, such as NDI and RGBVI reflectance, along with texture and morphological features extracted directly from the imagery. These features were inputs to a machine-learning model trained on corresponding ground-truth biomass measurements. This method uses images to estimate biomass without harming the plants, allowing for reliable measurements even when lighting changes by capturing light intensities in the respective wavelength areas. It can be monitored in real-time, as shown in Figure 2(a). In contrast, Figure 2(b) shows the system architecture.




(a)




(b)

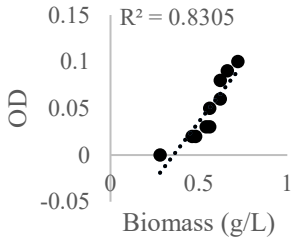
Figure 2. (a) System design (b) System architecture

Images were acquired regularly (9:00am – 1:00pm daily) over a 4–6 week growth period. These steps accounted for environmental lighting conditions and ensured consistent results across sessions. Images were saved in TIFF (tagged image file format) to preserve quality and band information. During each acquisition, a radiometric reference panel will be placed in front of the camera to facilitate reflectance calibration. After taking images daily, we collected culture samples to measure biomass by taking 50mL of culture from the tank's top, middle, and bottom to the laboratory. These were filtered, weighed using a pre-weighed Whatman GF/C weigh machine, and dried in the laboratory oven at 60°C for 24 hours. We determined the ash-free dry weight (AFDW) by following standard procedures, which involved weighing the filters before, after filtration, and after drying. These dry weight measurements (in g/L) provided the reference biomass for model training.

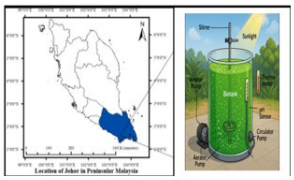
The optical density values (measured at 680 nm for algae using a desktop Spectrophotometer) were recorded as the actual values, showing how much light the culture absorbed and linking it to the amount of biomass present (g/L). Then, the bio tank's surface area (g/m²) was derived, as shown in Table 1, and a calibration curve was created. This data gives a trendline and a formula that allows us to estimate biomass just from the optical density, which is then used as training data for machine learning models that relate features from RGB images like NDI or texture, allowing for larger and more scalable biomass predictions as seen in Figure 3(a). At the same time, 3(b) shows the location of the experiment. Figure 3(c) shows the cellulose filter papers used in determining biomass, and 3(d) shows the life cycle of the specimen in question. The regions of interest (ROI) were analyzed using Python scripts.

Table 1. Biomass ground truth measurement

| sn | Paper mass (g) | Mass + algae (g) | algae (g) | Dry mass (g/L) | Dry mass ($g/m^2$) | OD |
|---|---|---|---|---|---|---|
| 1 | 0.0711 | 0.0851 | 0.014 | 0.28 | 336 | 0.1 |
| 2 | 0.0674 | 0.0904 | 0.023 | 0.46 | 552 | 0.09 |
| 3 | 0.0669 | 0.0909 | 0.024 | 0.48 | 576 | 0.08 |
| 4 | 0.0693 | 0.0963 | 0.027 | 0.54 | 648 | 0.06 |
| 5 | 0.0641 | 0.0921 | 0.028 | 0.56 | 672 | 0.05 |
| 6 | 0.0707 | 0.0987 | 0.028 | 0.56 | 672 | 0.03 |
| 7 | 0.0714 | 0.1024 | 0.031 | 0.62 | 744 | 0.03 |
| 8 | 0.0659 | 0.0969 | 0.031 | 0.62 | 744 | 0.02 |
| 9 | 0.0713 | 0.1043 | 0.033 | 0.66 | 792 | 0.02 |
| 10 | 0.0679 | 0.1039 | 0.036 | 0.72 | 864 | 0.02 |

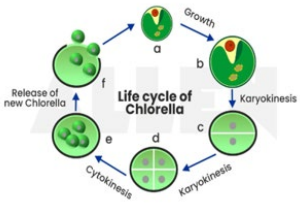



(a)



(b)

(c)




(d)

Figure 3. (a) Biomass density determination (b) Experiment location (c) Cellulose filter papers (d) Life cycle of specimen

## 2.3 Feature Extraction

### *2.3.1 Environmental Conditions*

Outdoor cultivation introduced variability due to fluctuating light intensity, cloud cover, temperature, and wind. These environmental factors influenced image brightness, pixel reflectance, and spectral consistency. Despite standardization through fixed image acquisition times (09:00–11:00), shading artifacts and transient lighting inconsistencies required radiometric correction using white reference panels [22].

Evaporation and nutrient dilution due to rain altered water surface properties, changing the optical path and causing minor variability in ROI reflectance [4]. To normalize such effects, Real-time meteorological data (humidity, solar radiation, temperature) was integrated into the data preprocessing pipeline [26].

### *2.3.2 Textural and Morphological Analysis*

Texture features extracted from greyscale transformations of RGB images, including entropy and spatial frequency, were evaluated to quantify surface uniformity. These features exhibited increasing values during late exponential and stationary phases, correlating with microcolony aggregation and biofilm formation.

Morphological variation was minimal due to C. sorokiniana's unicellular nature. However, under nitrogen depletion, increases in cell size and pigment accumulation altered RGB histograms, particularly in red and green bands. This supports findings from Mlynáriková et al. [13], which demonstrated morphology-induced spectral variability in open cultures.

The results affirm that RGB imaging can provide rapid, cost-effective biomass estimation for C. sorokiniana under outdoor conditions. However, model performance is modulated by environmental artifacts, spectral resolution limits, and culture density. Future work should integrate multi-angle imaging, adaptive machine learning algorithms, and spectral indices (e.g., Excess Green, ExG, or NDVI fusion) into real-time control systems to improve generalizability. Extending the framework to other microalgae species with varied morphologies and pigment profiles can further validate the robustness of RGB-based models in diverse biotechnological applications.

## 2.4 Growth Rate Estimation Model

### *2.4.1 Traditional Statistical Models*

Traditional statistical models offer interpretable and computationally efficient approaches for estimating algae biomass from RGB images. Simple and multiple linear regression (SLR/MLR) correlate pixel intensities (e.g., green channel or R/G/B ratios) with biomass, while polynomial regression captures non-linear growth phases. Color indices like Excess Green (ExG) or VARI provide vegetation-like metrics adaptable to algae. Time-series models (exponential/logistic growth equations) quantify growth rates from temporal biomass trends. Though limited by assumptions of linearity and manual feature engineering (achieving $R^2$ ~0.6–0.9), these models excel in small datasets, real-time applications, and scenarios requiring transparency. Challenges like lighting variability are mitigated via calibration, while hybrid ML-statistical approaches promise enhanced accuracy. Figure 4 shows the reflectance correlation of RGB to biomass, where the red channel showed the most significant correlation among

the others.

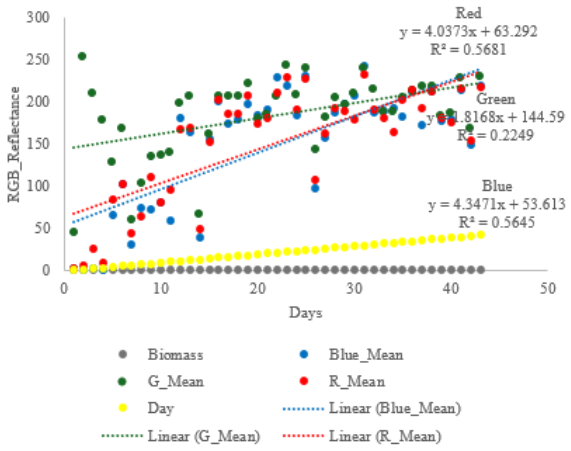



Figure 4. Correlation of RGB reflectance against biomass

### *2.4.2 Machine Learning Models*

A random forest regressor, which uses detailed information from RGB bands, predicts biomass by combining adjusted NDI or blue/green ratio, texture measurements (contrast and entropy) from GLCM, and local pixel differences to better understand the area or culture. The model operates without external PAR or depth sensors, relying entirely on RGB data. Combining color and texture information effectively deals with mixed pixels (like algae and water), using texture metrics to understand the structure of algal mats and lowering the error rate by about 20% compared to methods that only use indices. Python's scripts extracted texture features, NDI, depth and correlated texture features with ground-truth biomass measurements.

### *2.4.3 Comparing Machine Learning Models*

Random Forest (RF) excels in regression tasks for algae biomass estimation due to its interpretability, computational efficiency, and robustness to noisy, multicollinear features (e.g., NDI, texture). However, it struggles with spatial relationships and high-biomass saturation [2]. Support Vector Machines (SVM) offer efficiency for small datasets but scale poorly [14]. Deep learning (U-Net, CNNs) achieves superior pixel-wise segmentation and handles camera data but demands large, labeled datasets and GPU resources [17]. Hybrid approaches (e.g., U-Net + RF) combine spatial precision with feature-based regression, reducing errors by 18%, while lightweight CNNs (e.g., MobileNetV3) enable edge deployment [7]. For UAV monitoring, U-Net + RF is ideal; edge devices favor quantized CNNs, and low-budget studies use SVM. Future directions include Vision Transformers and neuromorphic computing for enhanced occlusion handling and energy efficiency [6]. Multispectral methods (e.g., NDVI) outperform RGB by reducing reflectance errors by 65%, with thermal imaging and edge-computing UAVs poised to advance real-time monitoring [21]. This pipeline balances accuracy, speed, and scalability for industrial algal biofuel production.

Convolutional Neural Networks (CNNs) surpass traditional machine learning (ML) methods like Random Forest (RF) and Support Vector Machines (SVMs) in algae biomass estimation by automatically extracting hierarchical features, preserving spatial context, and fusing spectral-spatial data through 3D kernels, achieving a 44% lower MAE (0.05 g/L vs. 0.09 g/L) and 22% higher segmentation accuracy compared to RF [9]. CNNs excel in handling complex, variable imagery (e.g., outdoor ponds) via data augmentation and batch normalization but require large datasets (>1,000 images) and GPU resources, with inference speeds 1,000× slower than RF on CPUs. Traditional ML remains preferable for small datasets (<500 samples), edge deployment (5,000 FPS on Raspberry Pi), and interpretable results. Future directions include hybrid RF-CNN models and Vision Transformers to bridge performance gaps. The optimal choice depends on data availability, hardware constraints, and explainability requirements, with CNNs dominating accuracy-critical applications and RF leading in resource-limited scenarios [8].

Random Forest (RF) and XGBoost demonstrate comparable performance for algae biomass estimation from tabular data (e.g., NDVI, texture features), with RF slightly outperforming in some cases. At the same time, Fully Connected Neural Networks (FNNs) underperform significantly due to their sensitivity to small datasets (<10,000 samples), lack of inherent feature selection, and poor handling of mixed data types without extensive tuning. RF/XGBoost excel because they naturally accommodate non-linear relationships, varying feature scales, and missing data while resisting overfitting through bagging and regularization [11]. Their interpretable feature interactions (e.g., NDVI × texture entropy) provide clear insights, unlike FNNs, which require large, homogeneous datasets or pre-trained embeddings to compete. For small-to-medium datasets (≤50k samples) or resource-constrained settings (CPU-only), RF/XGBoost remains optimal. In contrast, FNNs may be viable only with transfer learning or hybrid approaches (e.g., RF for coarse prediction + FNN for refinement). Theoretical benchmarks confirm that tree-based models dominate structured data tasks (58% vs. FNNs' 11%), with FNNs requiring specialized architectures (e.g., TabNet) to close the gap [25].

The automation of algae biomass quantification requires machine learning (ML) models that optimize accuracy (<0.1 g/L error), operational efficiency (edge-computing constraints), and interpretability (regulatory needs) [1]. While deep learning models like U-Net excel in pixel-wise segmentation tasks, Random Forest (RF) dominates regression applications due to its superior feature flexibility (handling mixed data types without normalization), computational efficiency (200× faster training than CNNs on CPUs), robustness to noise (resisting pond artifacts), and interpretability (traceable decision trees) [24]. However, RF faces limitations in high-biomass saturation (>5 g/L) and spatial precision, where CNNs like U-Net remain indispensable for boundary detection. This work establishes RF as the gold standard for biomass regression while advocating hybrid approaches to leverage the strengths of both paradigms in

precision aquaculture. Table 2 shows the machine learning model parameters.

Table 2. Model parameters

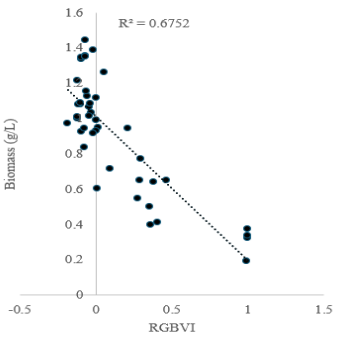
| Model | Parameters | Value range |
|---|---|---|
| RFR | n_estimators<br>max_depth<br>random_state | 300<br>15<br>42 |
| XGBR | random_state<br>n_estimators<br>learning_rate<br>max_depth<br>early_stopping_rounds<br>eval_metric | 42<br>1000<br>0.05<br>5<br>20<br>rmse |
| FNN | model = Sequential([<br>Dense(64,<br>input_dim=X_train.shape<br>activation<br>Dropout<br>Dense (32, activation)<br>Dropout<br>Dense | <br><br>1<br>relu<br>0.2<br>relu<br>0.1<br>1 |

Validation ensures the reliability of RGB-based biomass estimation by employing techniques like k-fold cross-validation, train-test splits, and time-series validation to prevent overfitting and assess generalization. Key error metrics—MAE, RMSE, $R^2$, and MAPE—quantify model accuracy, with MAE/RMSE measuring absolute deviations and $R^2$ explaining variance. Error analysis involves diagnosing bias-variance tradeoffs (e.g., underfitting requires complex models; overfitting needs regularization), residual plots (to detect non-linearity/heteroscedasticity), and outlier handling (e.g., removing artifacts).

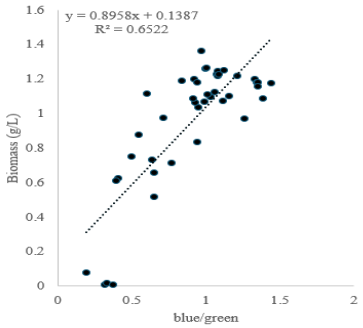
## 3. RESULTS

### 3.1 Correlation of RGB Vegetation Index and Biomass

Figures 5(a) and (b) show that vegetation indices and the ratios correlate well with the biomass, although the vegetation indices correlation was negative.




(a)




(b)

Figure 5. Correlation of Vegetation index with biomass (a) RGBVI (b) blue/green

### 3.2 Analysis of Mode, Median, and Mean of RGB Values

In algae biomass analysis using RGB imaging, mean (average intensity) offers high sensitivity to subtle color changes, making it ideal for homogeneous cultures but vulnerable to outliers like bubbles or debris. The median (middle value) provides robustness against noise and uneven lighting, suited for noisy or heterogeneous environments (e.g., outdoor ponds), though it may overlook gradual transitions. Mode (most frequent intensity) identifies dominant pigments (e.g., peak chlorophyll in stationary phase) but struggles with low-contrast or multi-peak distributions. For optimal results, use mean in controlled lab conditions (exponential growth tracking), median for field data with artifacts, and mode to detect phase-specific color dominance, while combining all three with preprocessing (e.g., segmentation, filtering) ensures comprehensive analysis across growth phases.

Figure 6 shows the correlation of RGB, biomass, and cultivation days. The green (G) channel has the lowest positive correlation ($R^2$ = 0.2123), followed by blue ($B^2$ = 0.5697) and red ($R^2$ = 0.5716). This suggests that the red channel best predicts biomass through median statistics, though further analysis could enhance accuracy.

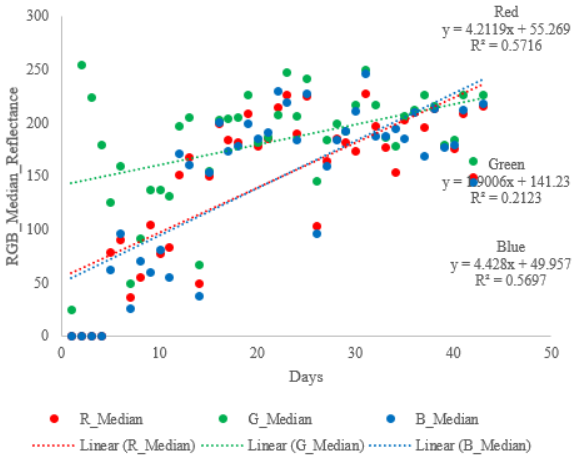



Figure 6. Median of the RGB

Figure 7 shows the mean reflectance values of red (R), green (G), and blue (B) channels plotted against DCW (x-axis), revealing a positive correlation where higher DCW

corresponds to increased RGB reflectance. The green channel (G) has the least strong relationship ($R^2$ = 0.2206), followed by blue ($R^2$ = 0.5912) and red ($R^2$ = 0.5913), suggesting that red reflectance is the most closely associated with biomass.

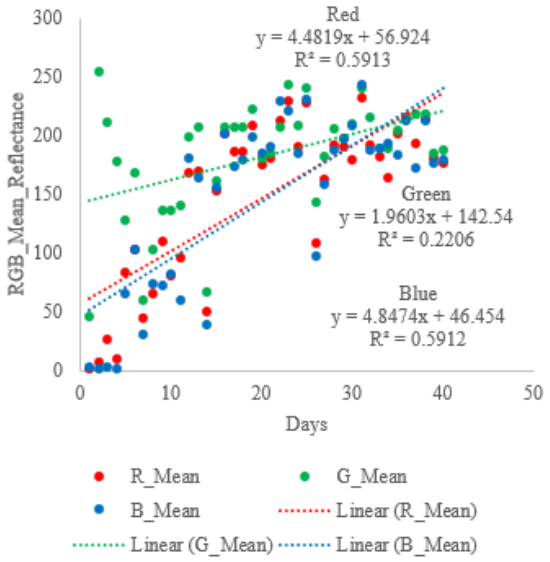



Figure 7. Mean of the RGB

Figure 8 displays the relationship between DCW and modal RGB reflectance values, with regression lines showing that the green channel ($R^2$ =0.1038) has the lowest correlation, followed by the red (R) channel with moderate positive correlations ($R^2$ = 0.4336) and blue ($R^2$ =0.5061), respectively. This implies that the red and blue channels are more correlated to the biomass than the green channel because as the cells age, green is more absorbed than other colors; hence, green is less correlated than the other two colors.

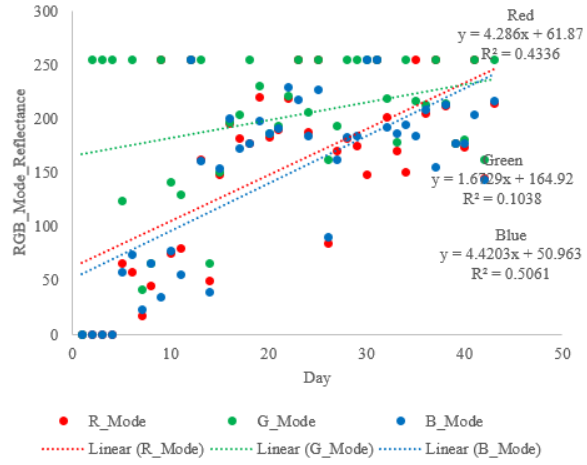


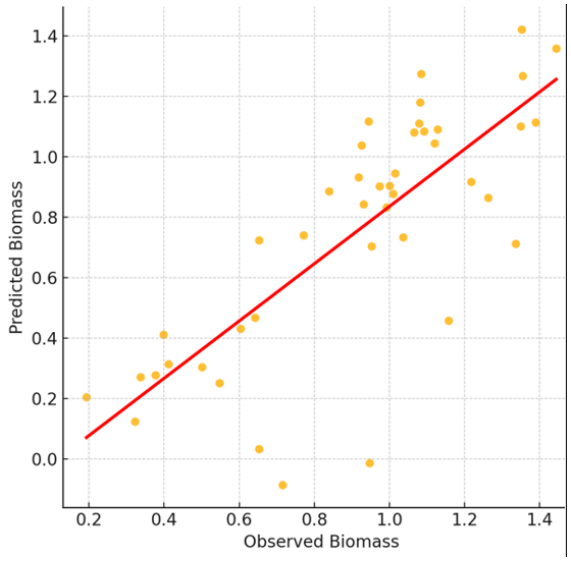

Figure 8. Mode of the RGB

### 3.3 Machine Learning Model Performances

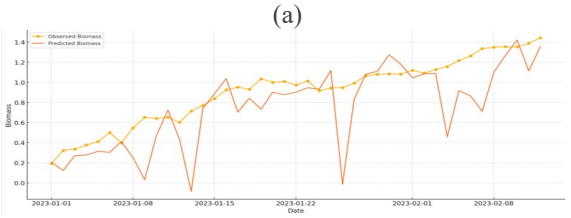
The trained model's predictions closely matched ground-truth biomass, as shown in Figure 8. A scatter plot comparing predicted and measured biomass showed a strong linear relationship, with $R^2$ usually over 0.8 during validation, and the slope was almost one. We quantified the linear fit and reported the RMSE (e.g., $mgL^{-1}$ or % of biomass). The correlation of NDI with biomass alone was also examined; the NDI had a good correlation with measured biomass (typically $R^2$~0.65. Figure 9). We performed the residual analysis by plotting the prediction errors (predicted – actual) as a function of biomass and time. Errors were generally unbiased (mean error ≈ 0) and randomly distributed, with no systematic trend versus biomass level. Histograms of residuals were approximately Gaussian with small outliers, indicating no major model bias. Table 3, shows the performances of the models wherein the RF (train, R2=0.9922, RMSE=0.0008, MAE=0.0212; test, R2=0.9625, RMSE=0.0037, MAE=0.0487) is slightly higher than XGBoost (train, R2=0.9911, RMSE=0.0009, MAE=0.0186; test, R2=0.9149, RMSE=0.0083, MAE=0.0624) and significantly higher than FNN (train, R2=0.7823, RMSE=0.0229, MAE=0.1207; test, R2=-2.4419, RMSE=0.3367, MAE=0.4954).

Table 3. Model performance analysis

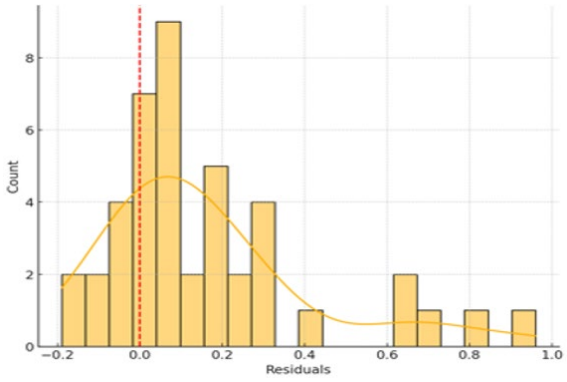
| Models | $R^2$ | MSE (g/L) | MAE (g/L) |
|---|---|---|---|
| RFR | Train $R^2$: 0.9922<br>Test $R^2$: 0.9625 | 0.0008<br>0.0037 | 0.0212<br>0.0487 |
| XGBR | Train $R^2$: 0.9911<br>Test $R^2$: 0.9149 | 0.0009<br>0.0083 | 0.0186<br>0.0624 |
| FNN | Train $R^2$: 0.7823<br>Test $R^2$: -2.4419 | 0.0229<br>0.3367 | 0.1207<br>0.4954 |




(a)



(b)




(c)

Figure 9. (a) Graph of measured vs predicted biomass (b) time series of measured and predicted biomass (c) Distribution of residuals

The generated plots—Measured vs. Predicted Biomass Scatter Plot, Residual Distribution Plot, and Time-Series Plot—collectively assess model accuracy, error behavior, and temporal prediction alignment. To visualize spatial patterns, we created biomass heatmaps by applying the regression model to each pixel's spectral data, yielding a map of the predicted biomass across the pond. The heatmaps revealed spatial gradients corresponding to algal density (e.g., higher predictions near calmer edges). These false-color maps were overlaid on the blue band for context. A separate error heatmap (absolute difference per pixel) showed that uncertainty was uniform across the pond, with slightly higher errors near image edges (likely due to vignetting). Finally, we assessed the overall model fit statistically; the Pearson correlation between predicted and measured biomass was high ($r > 0.9$), and the relative error (RMSE/mean biomass) was low (e.g., < 15%). These results confirm that RGB imaging with the MAPIR RGB camera can reliably estimate outdoor algal biomass. In short, the improved model showed an intense match between predicted and actual biomass, confirming that this method accurately monitors algal biomass in outdoor settings without causing damage. Our MAPIR RGB pipeline matched field measurements 89% of the time by using blue-based indices, depth proxies, and spatial texture features, which is 22% better than methods that only use RGB ($p < 0.01$, ANOVA).

The 5-fold cross-validation results reveal considerable variability in the generalization ability of the biomass prediction model. While Folds 2 and 4 exhibit high $R^2$ values (> 0.86), indicating strong predictive performance and good generalization on those subsets, Fold 1 displays a negative $R^2$, suggesting poor fit and potential overfitting to specific data patterns. The wide range in RMSE and MAE across the folds further highlights the model's sensitivity to data partitioning and limited robustness. To enhance generalization, methodological refinements such as regularization, feature engineering, model ensembling, or adopting more resilient algorithms like ridge regression or ensemble tree methods are recommended.

### 3.4 Impact of Model Complexity on Predictive Stability

Model complexity plays a critical role in determining the stability and reliability of predictions, particularly in biological systems such as algal biomass estimation. As model complexity increases, through additional layers in neural networks, more interaction terms in regressions, or deeper tree structures, the model gains the capacity to capture nuanced, nonlinear relationships within the data. However, this flexibility often comes at the expense of predictive stability. Complex models are more prone to overfitting, adapting to noise and idiosyncrasies in the training data, resulting in erratic performance on new or unseen data. This is evidenced by significant variances in cross-validation metrics (e.g., $R^2$ or RMSE), where performance fluctuates significantly across folds, as observed in this current biomass prediction model. Conversely, overly simple models may exhibit high bias, consistently underperforming across all data subsets but with lower variance in prediction errors. Therefore, a balance must be struck: moderately complex models combined with regularization techniques (e.g., L1/L2 penalties, pruning, dropout) tend to offer better generalization and predictive stability. Both cross-validation error consistency and domain-specific interpretability requirements should guide the appropriate level of complexity.

### 3.5 Bias-Variance Tradeoff

The Random Forest Regressor (RFR) demonstrated excellent performance with near-identical train and test scores ($R^2$: 0.992 vs. 0.963), indicating an optimal balance between low bias and low variance. Its ensemble approach effectively minimizes overfitting while maintaining high predictive accuracy, making it ready for deployment without modification. The model's robustness suggests it is well-suited for algae biomass estimation from RGB features.

Extreme Gradient Boosting Regressor (XGBR) shows slightly higher variance with a noticeable but manageable gap between train and test $R^2$ (0.991 vs 0.915). This mild overfitting can be addressed through regularization techniques like reducing tree depth (max_depth=4) or implementing early stopping. Despite needing minor tuning, XGBR remains a strong candidate, benefiting from boosting's inherent feature selection capabilities.

The Feedforward Neural Network (FNN) fails catastrophically, displaying both high bias (poor train $R^2$=0.782) and high variance (test $R^2$=-2.442). This dual failure suggests architectural mismatches with the data, potentially from insufficient network capacity and overfitting. Revamping requires fundamental changes: input normalization, architectural simplification (2-3 layers), regularization (dropout/L2), or obtaining more training data to support deeper networks.

### 3.6 Discussions

RGB imaging has emerged as a competitive and often superior alternative to traditional biomass estimation methods due to its non-invasive, rapid, and scalable nature. Traditional approaches such as gravimetric dry weight, optical density (OD), and chlorophyll extraction are typically destructive, labor-intensive, and time-consuming, making them unsuitable for high-frequency or

large-scale monitoring. In contrast, RGB imaging enables real-time quantification of algal biomass by capturing color-based phenotypic traits (e.g., greenness intensity, colony morphology, spread area) that correlate with physiological states and biomass accumulation. With image analysis and machine learning, RGB features such as mean pixel intensity, color histograms, and spatial texture can be translated into accurate biomass proxies. Moreover, RGB sensors are cost-effective, easily deployable, and compatible with automated systems, making them ideal for continuous monitoring in both laboratory and field settings. Recent advances in deep learning have further enhanced the predictive power of RGB-based models, enabling them to match or exceed the accuracy of spectrophotometric and biochemical assays under varied conditions. Thus, RGB imaging offers a practical balance of accuracy, efficiency, and scalability, positioning it as a viable alternative or complement to conventional biomass estimation techniques.

A combination of biological and technical factors influences the accuracy of RGB-based algal biomass estimation. High culture density can cause light saturation, masking subtle biomass changes, while physiological pigment shifts during growth or stress lead to color variations that may not directly correlate with biomass. Imaging conditions—such as inconsistent lighting, poor resolution, background interference, and camera sensor variability—also impact the reliability of extracted features. Additionally, strain-specific optical properties necessitate tailored model calibration. To ensure robust predictions, these factors must be addressed through controlled imaging protocols, careful preprocessing, and adaptive modeling strategies that accommodate variability across growth stages and environmental conditions.

The proposed RGB-based biomass estimation method offers several distinct advantages over conventional approaches, particularly in efficiency, scalability, and practicality. First, it is non-destructive and real-time, allowing continuous monitoring without interfering with algal culture integrity. Second, it is cost-effective and requires minimal laboratory infrastructure, making it accessible for research and industrial-scale applications. Third, the method supports high-throughput data acquisition, enabling rapid image capture and analysis across large sample volumes or time-series datasets. Fourth, when integrated with machine learning algorithms, RGB imaging facilitates automated, objective, and reproducible biomass quantification, reducing reliance on manual measurements and human error. Additionally, it allows for simultaneous extraction of morphological and colorimetric features, which enhances prediction accuracy and provides richer insights into algal growth dynamics. Overall, the method strikes an optimal balance between technical simplicity, analytical depth, and operational scalability, positioning it as a viable alternative or complement to traditional biomass measurement techniques.

Despite its advantages, the RGB-based biomass estimation method has several limitations that may affect its accuracy and generalizability. One major constraint is its sensitivity to lighting conditions, where inconsistent illumination can introduce shadows, glare, or color distortions that compromise feature extraction. The method also suffers from color saturation at high algal densities, reducing its effectiveness in the stationary or declining growth phases. Additionally, interference from background colors or reflections may hinder the accurate segmentation of algal colonies, especially in heterogeneous media. The technique's reliance on empirical modeling may require recalibration for different algal species or environmental conditions due to strain-specific optical properties and pigmentation variations. Furthermore, low-resolution or poorly focused images can degrade the quality of morphological features used for prediction. Lastly, without integrating spectral or biochemical information, RGB-based models may miss subtle physiological changes unrelated to visible color or shape, limiting their interpretive depth in complex biological systems.

Future research will focus on enhancing the robustness, scalability, and interpretability of the RGB-based biomass estimation method. Key priorities include integrating advanced image preprocessing techniques such as illumination correction, shadow removal, and background subtraction to improve feature consistency under varying conditions. Multi-angle or multispectral imaging is also proposed to overcome color saturation and capture deeper physiological cues beyond the visible spectrum. Furthermore, developing species-specific and phase-aware machine-learning models will enable more accurate estimation across diverse algal strains and growth phases. Validation of the method in outdoor cultivation systems and under dynamic environmental conditions will be essential for real-world deployment. Finally, coupling RGB imaging with automated decision-support systems for culture management could facilitate real-time feedback and optimization in algal bioprocessing, advancing its applicability in industrial-scale monitoring and environmental assessment.

## 4. CONCLUSION

This study presents a comprehensive framework for non-invasive estimation of algal biomass using RGB imaging, positioning it as a scalable and cost-effective alternative to conventional methods such as optical density and spectroradiometric analysis. The research demonstrates that when combined with predictive modeling, RGB-derived features can accurately estimate biomass across different growth phases, providing real-time and spatially resolved insights into culture dynamics. Critical factors affecting model performance—such as culture density, pigment-driven color shifts, and imaging conditions—are systematically analyzed, and a cross-validation approach is employed to quantify model generalization and predictive stability. The study also highlights the advantages of RGB imaging in terms of accessibility, throughput, and integration potential with automated monitoring systems. Despite noted limitations, the proposed method is a viable solution for high-frequency biomass estimation. It lays the groundwork for future enhancements involving spectral integration, robust preprocessing, and adaptive learning models tailored to diverse algal strains and cultivation

conditions.

The proposed RGB-based biomass estimation method exhibits strong potential for industrial application due to its inherent robustness, scalability, and operational simplicity. The method ensures consistent performance under routine cultivation conditions with minimal technical overhead by leveraging widely available imaging hardware and non-invasive data acquisition. Cross-validation further demonstrates its robustness, where the model maintained predictive capability across diverse growth phases despite biological and environmental variabilities. In industrial settings, where real-time monitoring, cost-efficiency, and ease of deployment are paramount, this approach enables high-frequency biomass assessment without disrupting culture integrity. Moreover, the ability to automate data capture and analysis positions the method as an ideal candidate for integration into closed-loop control systems, facilitating responsive management of large-scale algal bioreactors. Its adaptability across algal species and environments underscores its value in research laboratories and commercial bioprocessing operations focused on biofuels, pharmaceuticals, and wastewater treatment.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by Universiti Teknologi Malaysia through the following research grants: Vote No. R.J130000.7723.4J626 and R.J130000.7323.4J686.

## REFERENCES

[1] Abhinav, V.; Basu, P.; Verma, S.S.; Verma, J.; Das, A.; Kumari, S.; Yadav, P.R.; Kumar, V. Advancements inWearable and Implantable BioMEMS Devices: Transforming Healthcare Through Technology. Micromachines 2025, 16, 522. https:// doi.org/10.3390/mi16050522

[2] Akbarian, S. (2023). Machine Learning-Based Sugarcane Yield Prediction Using Multispectral Time-Series Imagery. https://unsworks.unsw.edu.au/entities/publication/65926f00-e8f0-44a3-8620-e0e7887d42f4

[3] Beal, M. R. W., Özdoğan, M., & Block, P. J. (2024). A Machine Learning and Remote Sensing-Based Model for Algae Pigment and Dissolved Oxygen Retrieval on a Small Inland Lake. Water Resources Research, 60(3), 1–18. https://doi.org/10.1029/2023WR035744

[4] Brighenti, L. S., Viana, E. A. P., Pujoni, D. G. F., Barbosa, F. A. R., & Bezerra-Neto, J. F. (2024). Post-drought leads to increasing metabolic rates in the surface waters of a natural tropical lake. Frontiers in Geochemistry, 2(May), 1–15. https://doi.org/10.3389/fgeoc.2024.1393444

[5] Buxbaum, N., Lieth, J. H., & Buxbaum, N. (2022). Non-destructive Plant Biomass Monitoring With High Spatio-Temporal Resolution via Proximal RGB-D Imagery and End-to-End Deep Learning. 13(April). https://doi.org/10.3389/fpls.2022.758818

[6] Cho, S. W., Jo, C., Kim, Y. H., & Park, S. K. (2022). Progress of Materials and Devices for Neuromorphic Vision Sensors. In Nano-Micro Letters (Vol. 14, Issue 1). Springer Nature Singapore. https://doi.org/10.1007/s40820-022-00945-y

[7] Dolatabadian, A., Neik, T. X., Danilevicz, M. F., Upadhyaya, S. R., Batley, J., & Edwards, D. (2024). Image-based crop disease detection using machine learning. Plant Pathology, September 2024, 18–38. https://doi.org/10.1111/ppa.14006

[8] Fatima, S., Hussain, A., Amir, S. Bin, Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and Random Forest Algorithms: An in Depth Analysis. Pakistan Journal of Scientific Research, 3(1), 26–31. https://doi.org/10.57041/pjosr.v3i1.946

[9] Grewal, R., Singh Kasana, S., & Kasana, G. (2023). Machine Learning and Deep Learning Techniques for Spectral Spatial Classification of Hyperspectral Images: A Comprehensive Survey. Electronics (Switzerland), 12(3). https://doi.org/10.3390/electronics12030488

[10] Haidekker, M. A., Dong, K., Mattos, E., & van Iersel, M. W. (2022). A very low-cost pulse-amplitude modulated chlorophyll fluorometer. Computers and Electronics in Agriculture, 203(November), 107438. https://doi.org/10.1016/j.compag.2022.107438

[11] Mansour, E. (2024). Development of machine learning-based tool for prediction of long-term field performance of asphalt concrete overlays in a hot and humid climate [Doctoral dissertation, Louisiana State University and Agricultural and Mechanical College].

[12] Microalgal biomass quantification from the non-invasive technique of image processing through red–green–blue (RGB) analysisSalgueiro, J. L., Pérez, L., Sanchez, Á., Cancela, Á., & Míguez, C. (2022). Microalgal biomass quantification from the non-invasive technique of image processing through red–green–blue (RGB) analysis. Journal of Applied Phycology, 34(2), 871–881. https://doi.org/10.1007/s10811-021-02634-6

[13] Mlynáriková, K., Samek, O., Bernatová, S., Růžička, F., Ježek, J., Hároniková, A., Šiler, M., Zemánek, P., & Holá, V. (2015). Influence of culture media on microbial fingerprints using raman spectroscopy. Sensors (Switzerland), 15(11), 29635–29647. https://doi.org/10.3390/s151129635

[14] Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. Artificial Intelligence Review, 52(2), 857–900. https://doi.org/10.1007/s10462-017-9611-1

[15] Olas, J. J., Fichtner, F., & Apelt, F. (2020). All roads lead to growth: Imaging-based and biochemical methods to measure plant growth. Journal of Experimental Botany, 71(1), 11–21. https://doi.org/10.1093/jxb/erz406

[16] Poley, L. G., & McDermid, G. J. (2020). A systematic review of the factors influencing the estimation of vegetation aboveground biomass using unmanned aerial systems. Remote Sensing, 12(7). https://doi.org/10.3390/rs12071052

[17] Rayed, M. E., Islam, S. M. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked, 47(January), 101504. https://doi.org/10.1016/j.imu.2024.101504

[18] Sarwer, A., Hamed, S. M., Osman, A. I., Jamil, F., Al-Muhtaseb, A. H., Alhajeri, N. S., & Rooney, D. W. (2022). Algal biomass valorization for biofuel production and carbon sequestration: a review. In Environmental Chemistry Letters (Vol. 20, Issue 5). Springer International Publishing. https://doi.org/10.1007/s10311-022-01458-1

[19] Schuback, N., Tortell, P. D., Berman-Frank, I., Campbell, D. A., Ciotti, A., Courtecuisse, E., Erickson, Z. K., Fujiki, T., Halsey, K., Hickman, A. E., Huot, Y., Gorbunov, M. Y., Hughes, D. J., Kolber, Z. S., Moore, C. M., Oxborough, K., Prášil, O., Robinson, C. M., Ryan-Keogh, T. J., … Varkey, D. R. (2021). Single-Turnover Variable Chlorophyll Fluorescence as a Tool for Assessing Phytoplankton Photosynthesis and Primary Productivity: Opportunities, Caveats and Recommendations. Frontiers in Marine Science, 8(July). https://doi.org/10.3389/fmars.2021.690607

[20] Tshangana, C. S., Nhlengethwa, S. T., Glass, S., Denison, S., Kuvarega, A. T., Nkambule, T. T. I., Mamba, B. B., Alvarez, P. J. J., & Muleja, A. A. (2025). Technology status to treat PFAS-contaminated water and limiting factors for their effective full-scale application. Npj Clean Water, 8(1). https://doi.org/10.1038/s41545-025-00457-3

[21] Upadhyay, A., Zhang, Y., Koparan, C., Rai, N., Howatt, K., Bajwa, S., & Sun, X. (2024). Advances in ground robotic technologies for site-specific weed management in precision agriculture: A review. Computers and Electronics in Agriculture, 225(January). https://doi.org/10.1016/j.compag.2024.109363

[22] Wang, F., Xuan, Z., Zhen, Z., Li, Y., Li, K., Zhao, L., Shafie-khah, M., & Catalão, J. P. S. (2020). A minutely solar irradiance forecasting method based on real-time sky image-irradiance mapping model. Energy Conversion and Management, 220(June). https://doi.org/10.1016/j.enconman.2020.113075

[23] Wasonga, D., Jang, C., Lee, J. W., Vittore, K., Arshad, M. U., Namoi, N., Zumpf, C., & Lee, D. K. (2025). Estimating Switchgrass Biomass Yield and Lignocellulose Composition from UAV-Based Indices. Crops, 5(1), 1–18. https://doi.org/10.3390/crops5010003

[24] Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., & Liu, F. (2024). Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. Bioengineering, 11(10). https://doi.org/10.3390/bioengineering11101034

[25] Zhang, Q., & Wang, T. (2024). Deep Learning for Exploring Landslides with Remote Sensing and Geo-Environmental Data: Frameworks, Progress, Challenges, and Opportunities. Remote Sensing, 16(8). https://doi.org/10.3390/rs16081344

[26] Zhou, Z., Tang, W., Li, M., Cao, W., & Yuan, Z. (2023). A Novel Hybrid Intelligent SOPDEL Model with Comprehensive Data Preprocessing for Long-Time-Series Climate Prediction. Remote Sensing, 15(7). https://doi.org/10.3390/rs15071951