Evaluating the Effects of LED Light on Plant Growth Using XGBoost and Interpretable Machine Learning with SHAP

Muhammad Nadzmi Bin Mohd Nizam1, a), Shih Yin Ooi1, 2, b), Ying Han Pang 1, 2, c) and It Ee Lee3, 4, d)

*1Faculty of Information Science and Technology (FIST), Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia*

*2Centre for Advanced Analytics (CAA), COE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia*

*3Faculty of Artificial Intelligence and Engineering (FAIE), Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia*

*4Centre for Wireless Technology (CWT), COE for Intelligent Network, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia*

*b) Corresponding author: syooi@mmu.edu.my*

*a) muhammad.nadzmi.mohd@student.mmu.edu.my*

*c) yhpang@mmu.edu.my*

*d) ielee@mmu.edu.my*

**Abstract.** LED lights are artificial lighting that has gained growing attention in precision agriculture due to its potential to promote plant growth under controlled conditions. This paper explores the effects of various LED wavelengths on plant growth by using the XGBoost machine learning algorithm. This paper further investigates the results of the algorithm using SHapley Additive exPlanations (SHAP). A few datasets are collected and analyzed before being trained by the model to achieve high predictive accuracy. The results show significant relationships between LED wavelengths and exposure to the growth of plants.

# introduction

In this paper we choose XGBoost to effectively model complex agricultural dataset [1]. This model is advanced at handling sophisticated relationship between inputs and outputs thus making predictions across various agricultural scenarios robust [2]. Although XGBoost often delivers remarkable predictive accuracy, analysts regularly note its considerable opacity. By overlaying SHAP values onto the models’ output, researchers can begin to unpack the local and global drivers of predictions, a need that arises frequently in data drawn from agricultural experiments.

Research on LED lighting and plant growth spans decades, yet few investigations incorporate modern machine-learning approaches with demonstrable accuracy and transparency. The present study closes that loop by reprocessing publicly archived datasets through XGBoost paired with SHAP, thereby revealing the distinct roles various LED settings play in phytonutrient development [3]. A central aim is to benchmark the XGBoost model against Random Forest, support-vector-machine, and classic-decision-tree implementations, providing an empirical test of predictive resilience across frameworks. SHAP-derived feature scores further clarify which light parameters matter most, underscoring how interpretable algorithms can demystify opaque agricultural datasets.

# Literature Review

## Impact of LED Lighting on Plant Growth

Light-emitting diode lamps have increasingly supplanted traditional horticultural fixtures and now serve as the primary source of illumination in many controlled environments. Their modular spectrum enables growers in both greenhouses  [4] and vertical farms to fine-tune light recipes, often lifting growth rates and crop quality [5]. Modifying the LED light spectrum holds promise for laboratory horticulture. Red and blue peaks can be dialed in with surprising precision, and that flexibility nudges the plant toward peak photosynthetic output [6]. The tunable spectrum permits deliberate manipulation of physiological processes in crops, fostering luxuriant vegetative expansion under blue wavelengths while stimulating flowering and fruit set when red light predominates [7].

Compared with older technologies such as high-pressure sodium lamps or fluorescent tubes, light-emitting diodes draw considerably less power and emit much less waste heat. The lower thermal burden extends fixture lifetime and eases the cooling load in enclosed spaces [8]. Light-emitting diodes enjoy a remarkably long service life and incur minimal electricity expenses, features that are steadily drawing farmers toward them in twenty-first-century field science [9]. The upfront price tag of LED lighting arrays can indeed appear steep, yet their lasting dividends in crop yield and environmental stewardship justify the outlay for many modern farms.

## Comparison of XGBoost with Other ML Models in Agriculture

Researchers have repeatedly noted that XGBoost outclasses Support Vector Machines and K-Nearest Neighbors when the volume of data swells. The gradient-boosting framework tends to scale smoothly, absorbing new records with a surprising lack of fuss. Moreover, the model resists the urge to bog analysts down in minute sacrifices of tuning, yielding solid predictions even with a coarse parameter grid. [10]. Artificial Neural Networks are mathematically adept at seizing elaborate nonlinear structures in data, yet they seldom operate without an appetite for vigorous computation and oceans of labelled observations. XGBoost, in contrast, finds middle ground; it converges quickly on a solution, demands only modest hardware to stay responsive, and quietly ranks predictors according to their influence [11].

In agricultural tasks such as yield prediction [12], disease classification [13], and soil property estimation, XGBoost consistently outperforms or matches other machine learning models [14]. The models ability to produce precise forecasts while simultaneously offering clear, intelligible explanations has positioned it as a go-to tool for scientists and field operators who need trustworthy, immediately usable guidance drawn from farm-level statistics.

## Applications of SHAP in Agricultural Models

Agricultural scientists now increasingly turn to SHAP because the method converts opaque machine-learning predictions into human-readable probabilities. Field managers, policy analysts, and even extension agents can see briefly which whiskeys of nitrogen or rainfall bent the outcome. The approach moves beyond simple feature ranking, letting users interrogate a given prediction and trace its calculation back to each variable in real time [15]. Integrating SHAP explanations into the machine-learning pipelines of crop-yield forecasting systems can sharpen the quality of management decisions on the ground [16].

Beyond crop and disease management, SHAP is also applied in models for irrigation optimization [17], pest outbreak prediction [18], and risk assessment in agricultural finance. Clear, user-friendly explanations of model reasoning are one reason farmers and researchers are welcoming SHAP into their daily workflows. Helpful visual summaries turn probabilistic forecasts into concrete advice, and that transparency smooths the path from model prototype to field application.

# Methodology

This section outlines the steps taken to prepare and analyze the dataset, train the XGBoost model, and interpret its results using SHAP. Each component is elaborated below to provide a clear understanding of the methods employed.

## Dataset Description

The datasets used in this study originated from an experiment by [19]. It is a controlled-environment experiment which investigated the effects of substituting red light with green and/or far-red (FR) light on the growth of red-leaf (Rouxai’) and green-leaf (Rex’) lettuce. The study was conducted in a hydroponic setup using 12 distinct lighting treatments under constant total photon flux density. The dataset includes measurements from 192 plants per cultivar, across 23 variables, collected over two replications. Table 1 below shows the full list of features and the description of each feature.

**TABLE 1.** Full list of features and descriptions

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Replication | Experimental repetition block (1 or 2) | |
| Lighting Treatment | Name of the light spectrum treatment | |
| Blue Proton Flux Density | Intensity of blue light (400–499 nm) in µmol·m⁻²·s⁻¹ | |
| Green Proton Flux Density | Intensity of green light (500–599 nm) | |
| Red Proton Flux Density | Intensity of red light (600–699 nm) | |
| Far-red Proton Flux Density | Intensity of FR light (700–799 nm) | |
| Red+Far - Red Proton Flux Density | Combined intensity of red and FR light | |
| extended photosynthetic PFD | Photon flux over 400–799 nm range | |
| the fraction of FR light to the extended PFD | FR / (400–799 nm total) | |
| FR Fraction | FR / (Red + FR) | |
| phytochrome photoequilibria | PPE: Estimate of active phytochrome | |
| internal phytochrome photoequilibria | PPE adjusted for internal leaf distortion (iPPE) | |
| Yield photon flux density | Estimated effective photons based on McCree's curve | |
| Watt\_t\_400\_700 | Power consumption in the PAR region | |
| plant\_no | Plant identification number | |
| Fresh mass | Total shoot fresh weight (g) | |
| Plant diameter | Width of plant canopy (cm) | |
| Leaf number | Number of leaves >2 cm in length | |
| leaf\_area | Leaf area of the fifth fully expanded leaf (cm²) | |
| Soil Plant Analysis Development | SPAD index measuring relative chlorophyll content | |
| DM,fifthleaf | Dry mass of the fifth fully expanded leaf (g) | |
| Drymass(DM) | Total shoot dry weight (g) | |
| specific leaf area | Leaf area per gram of dry mass (cm²·g⁻¹) | |
| Replication | Experimental repetition block (1 or 2) | |
| Lighting Treatment | Name of the light spectrum treatment | |
| Blue Proton Flux Density | Intensity of blue light (400–499 nm) in µmol·m⁻²·s⁻¹ | |

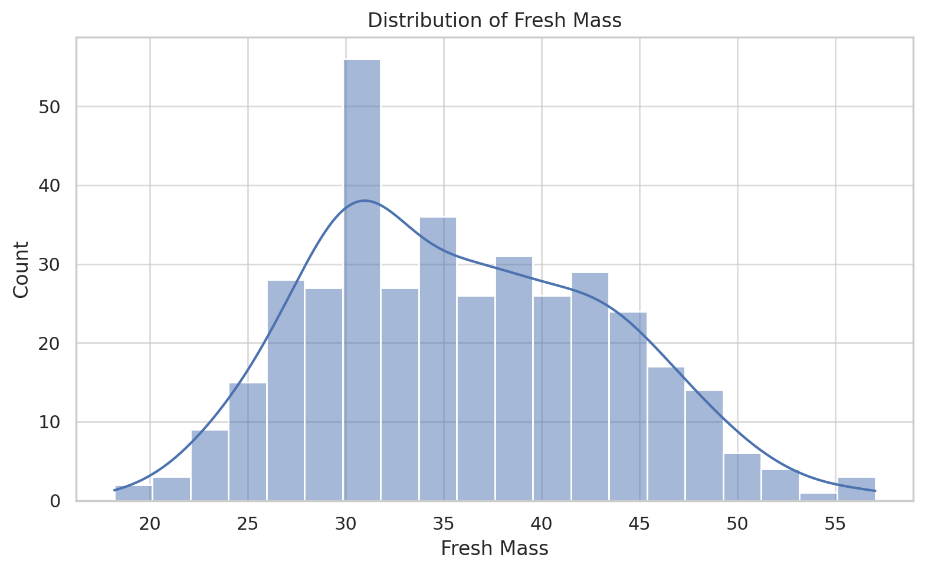
## Exploratory Data Analysis (EDA)

The dataset contains a total of 23 features, including various LED light measurements and plant growth metrics. Referring to Figure 1, the target variable, Fresh Mass, has a mean of 35.90 grams, with a standard deviation of 7.48. The minimum observed Fresh Mass is 18.19 grams, while the maximum is 57.04 grams. There are 12 unique lighting treatment classes, each representing a different LED light combination. The class distribution is relatively balanced, with each treatment having roughly similar sample counts. There are 0 missing values in the dataset, indicating that the data is complete and does not require imputation.

The distribution plot of fresh mass reveals that most plants exhibited moderate growth, with the most frequent fresh mass observed around 30 grams. The shape of the distribution is slightly right-skewed, indicating that while most plants fall within the 30–35g range, a smaller number achieved higher mass values exceeding 50 grams. This spread suggests that different LED light treatments have varying effects on plant growth. The presence of both low and high fresh mass outliers highlights the importance of identifying which specific light variables contribute to optimal or suboptimal growth. Overall, this distribution supports the use of XGBoost and SHAP to further analyze and interpret the influence of light-related features on plant development.

## Data Preprocessing

Referring to Figure 2, the preprocessing phase began by merging data from two sheets, Rouxai data and Rex data, each containing 192 samples. From the original 23 features, non-informative columns such as Replication, plant\_no, and Lighting Treatment were removed, as they served only as identifiers or non-numeric labels not directly relevant to modeling. The remaining features included LED-related measurements such as Blue Proton Flux, Red Proton Flux, Far-red Proton Flux, FR Fraction, Extended PFD, and phytochrome-related metrics, all of which were retained for their direct relevance to light exposure and plant physiological response. Growth outcome features such as Fresh Mass, Leaf Number, Plant Diameter, and Leaf Area were extracted and used to compute a composite score representing overall plant performance. This score was standardized using Z-score normalization. Plants with a composite score above the mean were labeled as successful (Growth = 1), while those below were labeled unsuccessful (Growth = 0), thus enabling binary classification. All input features were also standardized to ensure uniform scale. The resulting clean dataset consisted of 96 samples with 13 numerical input features and one binary target label (Growth), with no missing values. This dataset was then used to train the XGBoost model with a binary classification objective to investigate the relationship between LED lighting characteristics and plant growth success.



**Figure 1.** Distribution of Fresh Mass among all lettuce samples

A diagram of a data processing process

AI-generated content may be incorrect.

**Figure 2.** Data preprocessing pipeline

## Model Training and Evaluation

The model training process began by splitting the preprocessed dataset into training and testing subsets. Three machine learning algorithms were selected for performance comparison: Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and K-Nearest Neighbors (k-NN). XGBoost was chosen for its strong predictive performance, ability to handle missing values, and built-in regularization, making it well-suited for structured tabular data. SVM was included due to its effectiveness in handling high-dimensional spaces, while k-NN was selected as a simple, instance-based learning method for benchmarking purposes. All models were trained using the same preprocessed dataset. To improve generalization and prevent overfitting, 5-fold cross-validation was applied to each model.

Hyperparameter tuning was conducted using grid search for all three models: for XGBoost, parameters tuned included learning rate, maximum depth, number of estimators, and subsample ratio; for SVM, kernel type and regularization parameter (C) were optimized; and for k-NN, the number of neighbors (k) was selected based on cross-validation performance. Model evaluation was based on standard classification metrics including accuracy, precision, recall, and F1-score. These metrics were calculated under both the hold-out test set and 5-fold cross-validation schemes to ensure fair comparison and model robustness. To further refine model performance, a grid search with 5-fold cross-validation was applied to identify the optimal hyperparameters for XGBoost. Table 2 summarizes the best configuration selected based on cross-validation accuracy.

**TABLE 2.** Final XGBoost hyperparameter settings

|  |  |
| --- | --- |
| **Hyperparamete** | **Value** |
| Learning Rate | 0.1 |
| Max Depth | 5 |
| Number of Estimators | 100 |
| Subsample Ratio | 0.8 |
| Colsample\_bytree | 0.8 |
| Gamma | 0 |
| Min Child Weight | 1 |

## Model Interpretation with SHAP

To enhance the interpretability of the XGBoost model, the SHAP (SHapley Additive exPlanations) technique was applied. SHAP provides a unified framework for explaining model predictions by assigning each feature a contribution value, indicating its impact on the final output. SHAP values allow for both global and local interpretability, showing how features influence the model’s predictions across the entire dataset and for individual instances. This approach helped in understanding the importance of various LED light parameters in predicting plant growth, offering transparent insights into the model’s decision-making process. By providing clear explanations for predictions, SHAP improved stakeholder trust and facilitated more informed agricultural decision-making.

# Results

## XGBoost Results

To evaluate the performance of different machine learning models, accuracy scores were first measured on a single train-test split. The initial results (Table 3) showed that XGBoost achieved perfect accuracy (1.000), significantly outperforming Support Vector Machine (0.7368) and k-Nearest Neighbors (0.6842). However, to ensure these results were not due to overfitting or a favorable data split, a 5-fold cross-validation (Table 4) was conducted across all three models. The cross-validated results confirmed the superiority of XGBoost with a mean accuracy of 0.9895 and a low standard deviation of 0.0211, indicating both high performance and consistency. In comparison, SVM and k-NN recorded similar mean accuracies of 0.7826 and 0.7832 respectively, but with higher standard deviations, especially for k-NN (0.1410), suggesting fewer stable predictions. These findings establish XGBoost as the most effective model for predicting plant growth outcomes based on LED light parameters.

**TABLE 3.** Model accuracy comparison (without cross validation)

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| XGBoost | 1.00 |
| SVM | 0.7368 |
| k-NN | 0.6842 |

**TABLE 4.** Model accuracy comparison (5-fold cross validation)

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Accuracy** | **Std Dev** |
| XGBoost | 0.9895 | 0.0211 |
| SVM | 0.7826 | 0.0969 |
| k-NN | 0.7832 | 0.1410 |

## SHAP Interpretations

Referring to Table 5, the SHAP analysis for both models provides critical insights into the most influential features driving the predictive outcomes. In the Lettuce Growth model, specific leaf area emerged as the dominant feature, with a mean SHAP value of 3.122147, significantly higher than the other variables. This suggests that leaf development, a key physiological characteristic, is highly predictive of lettuce growth under the studied lighting conditions. Leaf area and plant diameter also demonstrated notable impacts, with mean SHAP values of 0.613571 and 0.289712, respectively. The importance of Far-red Proton Flux Density and FR Fraction highlights the role of light quality, particularly far-red light components, in influencing plant growth responses.

**TABLE 5.** SHAP interpretations

|  |  |
| --- | --- |
| **Feature** | **Mean Shape Value** |
| Specific leaf area | 3.122147 |
| Leaf area | 0.613571 |
| Plant diameter | 0.289712 |
| Far-red Proton Flux Density | 0.111419 |
| FR Fraction | 0.072446 |

# discussion

The SHAP results showed that specific leaf area, total leaf area, and stem diameter were the key variables for forecasting how plants grew under the different LED setups tested. These predictors fit established agronomy literature, which cites leaf traits and stem thickness as drivers of shoot expansion, especially when light spectrum changes. Measurements related to far-red light also played a meaningful role, reinforcing past studies that link the red-to-far-red ratio with foliar elongation and morphology. Running the data through an XGBoost ensemble produced over 98 percent accuracy with low cross-validation variance, indicating that the model is both precise and stable across samples. Alongside high predictive power, the feature-attribution maps built with SHAP made the algorithms reasoning easier to grasp, providing concrete guidance for fine-tuning LED recipes and boosting crop yields.

# CONCLUSION

The research confirmed that XGBoost reliably forecasts plant growth, while SHAP boosted interpretability. The lettuce-growth and aphid-control models exceeded 90 percent accuracy, underscoring XGBoosts strength with intricate farm datasets. SHAP spotlighted critical predictors-leaf characteristics and light spectrum for lettuce, and biomass for aphid control-emphasizing the need for tailored insights. In summary, SHAP rendered the results clearer and more useful, underpinning decisions guided by data.

# Acknowledgments

This research work was supported by the Ministry of Higher Education (MOHE) under the 2023 Translational Research Program for the Energy Sustainability Focus Area (Project ID: MMUE/240001), the 2024 ASEAN IVO, and Multimedia University, Malaysia.

# References

1. P. Kamath, P. Patil, S. S, Sushma, and S. S, “Crop yield forecasting using data mining,” Global Transitions Proceedings **2**(2), 402–407 (2021).
2. O. M’hamdi, S. Takács, G. Palotás, R. Ilahy, L. Helyes, and Z. Pék, “A Comparative Analysis of XGBoost and Neural Network Models for Predicting Some Tomato Fruit Quality Traits from Environmental and Meteorological Data,” Plants **13**(5), 746 (2024).
3. Z. Ma, J. Guo, S. Mao, and T. Gu, “An interpretability research of the Xgboost algorithm in remaining useful life prediction,” in 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), (IEEE, Bangkok, Thailand, 2020), pp. 433–438.
4. C. Maraveas, “Incorporating Artificial Intelligence Technology in Smart Greenhouses: Current State of the Art,” Applied Sciences **13**(1), 14 (2022).
5. T. Verma, N. T. Zhi Wei, F. Gao, H. Yu, and R. S. Filho, “Optimizing Indoor Farming: Deep Learning for Predicting Plant Growth under LED Light Treatments,” in 2024 IEEE Conference on Artificial Intelligence (CAI), (IEEE, Singapore, Singapore, 2024), pp. 1051–1056.
6. Y. Park, and E. S. Runkle, “Spectral effects of light-emitting diodes on plant growth, visual color quality, and photosynthetic photon efficacy: White versus blue plus red radiation,” PLoS ONE **13**(8), e0202386 (2018).
7. A. A. Alrajhi, A. S. Alsahli, I. M. Alhelal, H. Z. Rihan, M. P. Fuller, A. A. Alsadon, and A. A. Ibrahim, “The Effect of LED Light Spectra on the Growth, Yield and Nutritional Value of Red and Green Lettuce (Lactuca sativa),” Plants **12**(3), 463 (2023).
8. D. H. Park, D. B. Lee, E. R. Seo, and Y. J. Park, “A parametric study on heat dissipation from a LED-lamp,” Applied Thermal Engineering **108**, 1261–1267 (2016).
9. D.D. Avgoustaki, and G. Xydis, “Energy cost reduction by shifting electricity demand in indoor vertical farms with artificial lighting,” Biosystems Engineering **211**, 219–229 (2021).
10. T. Chen, and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM, San Francisco, California, USA, 2016), pp. 785–794.
11. J. Wu, Y. Li, and Y. Ma, “Comparison of XGBoost and the Neural Network model on the class-balanced datasets,” in 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC), (IEEE, Greenville, SC, USA, 2021), pp. 457–461.
12. S. Saiful, and N. B. Wibisono, “Crop Yield Prediction Using Random Forest Algorithm and Xgboost Machine Learning Model,” International Journal of Research and Innovation in Social Science IX(III), 1983–1994 (2025).
13. A. Y. Ashurov, M. S. A. M. Al-Gaashani, N. A. Samee, R. Alkanhel, G. Atteia, H. A. Abdallah, and M. Saleh Ali Muthanna, “Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections,” Frontiers in Plant Science **15**, 1505857 (2025).
14. M. A. Razavi, A. P. Nejadhashemi, B. Majidi, H. S. Razavi, J. Kpodo, R. Eeswaran, I. Ciampitti, and P. V. V. Prasad, “Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data,” Artificial Intelligence in Agriculture **14**, 99–114 (2024).
15. T. Abekoon, H. Sajindra, N. Rathnayake, I. U. Ekanayake, A. Jayakody, and U. Rathnayake, “A novel application with explainable machine learning (SHAP and LIME) to predict soil N, P, and K nutrient content in cabbage cultivation,” Smart Agricultural Technology **11**, 100879 (2025).
16. E. J. Jones, T. F.A. Bishop, B. P. Malone, P. J. Hulme, B. M. Whelan, and P. Filippi, “Identifying causes of crop yield variability with interpretive machine learning,” Computers and Electronics in Agriculture **192**, 106632 (2022).
17. F. Rodríguez-Díaz, A. M. Chacón-Maldonado, A. R. Troncoso-García, and G. Asencio-Cortés, “Explainable olive grove and grapevine pest forecasting through machine learning-based classification and regression,” Results in Engineering **24**, 103058 (2024).
18. E. E. Hussein, B. Zerouali, N. Bailek, A. Derdour, S. S. M. Ghoneim, C. A. G. Santos, and M. A. Hashim, “Harnessing Explainable AI for Sustainable Agriculture: SHAP-Based Feature Selection in Multi-Model Evaluation of Irrigation Water Quality Indices,” Water **17**(1), 59 (2024).
19. N. Kelly, and E. S. Runkle, “Dependence of far-red light on red and green light at increasing growth of lettuce,” PLoS ONE **19**(11), e0313084 (2024).