Prediction of Strawberry Leaf Color Using RGB Mean Values Based on Soil Physicochemical Parameters Using Machine Learning Models

Bolappa Gamage Kaushalya Madhavi 1, Jayanta Kumar Basak 2, Bhola Paudel 1, Na Eun Kim 1, Gyeong Mun Choi 2 and Hyeon Tae Kim 1,*

1 Department of Bio-Systems Engineering, Institute of Smart Farm, Gyeongsang National University, Jinju 52828, Korea; kaushalyamadhavi81@gmail.com (B.G.K.M.); bholapaudel@gnu.ac.kr (B.P.); na.eun@gnu.ac.kr (N.E.K.)
2 Institute of Smart Farm, Gyeongsang National University, Jinju 52828, Korea; jk.basak@gnu.ac.kr (J.K.B.); ne9179@gmail.com (G.M.C.)
* Correspondence: bioani@gnu.ac.kr; Tel.: +82-55-772-1896; Fax: +82-55-772-1899

Abstract: Intensively grown strawberries in a greenhouse require frequent and precise soil physicochemical constituents for optimal production. Strawberry leaf color analyses are the most effective way to evaluate soil status and protect against excess environmental nutrients and financial setbacks. Meanwhile, precision agriculture (PA) endorsements have been utilized to mimic solutions to these problems. This research aimed to create machine learning models such as multiple linear regression (MLR) and gradient boost regression (GBR) for simulating strawberry leaf color changes related to soil physicochemical components and plant age using RGB (red, green, and blue) mean values. The soil physicochemical properties of the largest varied colored leaves of strawberry were precisely measured by a multifunctional soil sensor from the rooting zones. Simultaneously, 400 strawberry leaflets were detached in each vegetative and reproductive stage, and individual leaves were captured using a digital imaging system. The RGB mean values of colored images were extracted using the image segmentation algorithms of image processing technique. Consequently, MLR and GBR models were developed to predict leaf RGB mean values based on soil physicochemical measurements and plant age. The GBR model vigorously fitted with RGB mean values throughout the growth stage, with $R^2$ and RMSE values of ($R=0.77, 7.16$, $G=0.72, 7.37$, and $B=0.70, 7.37$), respectively. Furthermore, the MLR model performed moderately with $R^2$ and RMSE values of ($R=0.67, 8.59$, $G=0.57, 9.12$, and $B=0.56, 6.81$) when consecutively predicting RGB mean values in strawberry leaves. Eventually, the GBR model performed more effectively than the MLR model with high-performance metrics. In addition, the leaf color model uses visualization technology to measure growth progress, and it performs well in predicting dynamic changes in strawberry leaf color.

Keywords: strawberry leaf color; multiple linear regression; gradient boost regression; RGB mean values; soil physicochemical parameters

1. Introduction

Strawberry (Fragaria × ananassa) cultivation is markedly increased in South Korea. Optimal strawberry yield in greenhouse cultivation is inducible by favorable nutrient availability in the soil. Simultaneously, greenhouse strawberry output has surpassed traditional soil growing in recent years all across the world [1]. Correspondingly, the strawberry cultivation leaf is considered a vital vegetative organ that acts as a reservoir for phytochemicals and other bioactive compounds that lead to the growth of the reproductive structures [2]. Therefore, leaf color is a typical visual character of the strawberry plant impacted by the growing environment and can help understand plant
growth status, as reported by Zhang et al., 2014 [3]. Concurrently, leaf color simulation is the significant index for the verisimilitude and its application to strawberry production [4].

Digital image technology has grown in prominence in PA, with high-resolution cameras, namely, hyperspectral and multispectral, making plant phenotypic information such as shape and leaf color feasible. Simultaneously, the RGB color model and features of the red, green, and blue channels are the most used color representation for digital photographs for detecting plant chlorophyll levels. Furthermore, the scarcity of this information has become a bottleneck for maintaining strawberry production all around the world [5].

Generally, chlorophyll is the green pigment absorbs the light required for photosynthesis. Nitrogen (N) is an essential element for chlorophyll synthesis and is a component of the chlorophyll molecule that promotes photosynthesis [6]. Moreover, phosphorous (P) is of utmost importance for energy transfer and the photosynthesis process, and deficient P indicates reddish to violet color developments due to anthocyanin synthesis, as revealed by Madhavi et al., 2020 [2]. Furthermore, and even more importantly, potassium (K) is also required for sugar production, and a K shortage in plants results in the yellowing of leaf veins and purple spots [7]. Therefore, NPK nutrients influence strawberry production, and deficiencies of the abovementioned major nutrients indicate the reduction of pigment formation and subsequent leaf color changes from green to yellowish or purple. In addition to that, intensively produced strawberries require frequent and exact fertility management. Leaf analyses are the most accurate way to track the nutritional status and any deficits. Leaf analyses ensure productivity and quality and safeguard the environment by preventing the application of excess nutrients and unnecessary fertilizer [8].

The PA aims to reduce the unnecessary application of fertilizer and ill farming practices by controlling physicochemical parameters by the development of sensors in smart farms [9]. These advancements have made farmers more efficient by regulating the application of fertilizer and maintaining the optimum conditions. Specifically, nutritional problems are prevalent in strawberry cultivation in greenhouses and may go undetected for prolonged periods. Moreover, improper fertilizer distribution in the root zone aids in bringing down plant vigor, which may reduce plant growth and yield in greenhouse cultivation. Therefore, PA aims to test the growing medium throughout the production cycle, and to control soil physicochemical parameters such as electrical conductivity (EC) and pH to eliminate almost all problems associated with fertilization [10].

Strawberries are typically grown on soil with a pH of 4 to 8, with acid soils being less appropriate for strawberry production compared to alkaline soils with a pH of 8. According to Dixon et al., 2019 [11], the best soil pH for strawberry plant growth is 5.4 to 6.5. Hence, high soil pH levels have a negative impact on strawberry yield and leaf color development [12]. Furthermore, EC indirectly affects leaf color changes due to the magnitude of cation and anion exchange that supplies nutrients to strawberry plants [13]. According to Dixon et al., 2019 [11], a soil EC value closer to 1 dS/m is deemed optimal for plant growth. Furthermore, as shown by Sim et al., 2020 [9], soil temperature enhances the cation–anion exchange function, while lower temperatures alter the color of leaflets. Therefore, soil physicochemical parameters such as pH, EC, soil temperature (ST), and NPK level in the root zone significantly affect the leaf color changes from normal to abnormal conditions. PA was designed to forecast plant phenotypic changes (RGB color) based on soil physicochemical and plant age characteristics using the information above. Nonetheless, repeated measurement of the same leaf affects its physiological status, as mentioned by Borhan et al., 2017 [6].

According to Jaihuni et al., 2021 [8], previous studies forecasting nutrients such as NPK by utilizing statistical and machine learning (ML) methodologies were conducted. Himelrick et al., 1992 [14], discovered the extractable chlorophyll amount of strawberry leaves on a fresh weight basis using SPAD meter numerical correlation readings, and the relationship ($R^2$) was reported to be 0.92. Nonetheless, the application of photogrammetric data was cumbersome, and a considerable $R^2$ ($<0.90$) was achieved. According to the previous literature, most modelling studies on the prediction of leaf RGB color changes
have been developed based on leaf chlorophyll content or SPAD value changes in different phases [5]. Maresma et al., 2016 [15], explored an unmanned air vehicle (UAV) and various vegetation indices in maize fields using fertilizer application rates through the regression model. The study provided convincing results, with $R^2$ values of 0.92 from different indexes. However, the limitation of this study is that it did not analyze one single macronutrient effect for vegetative growth. Hence, as these studies used datasets of moderate size, there is a strong possibility that overfitting occurred. Ozreccberouglu et al., 2020 [16], explored various linear regression models to investigate the optimum pomegranate leaf chlorophyll content (of a given area) using both $G$ and $B$ color values. In the present study, ML models are used, which employ the soil physicochemical parameters with plant age to predict the RGB color of strawberry plant leaves.

ML is a promising technique for analyzing massive volumes of data and is mainly applied for prediction and classification. However, plant science, plant production, and plant phenotyping are just a few disciplines where this technology is applied [17,18]. Gradient boost regression (GBR) was used as the prediction model in this investigation. GBR is a regression machine learning technique that generates a prediction model in the form of a “decision tree”, an ensemble of weak prediction models. It creates a model, stage by stage, and then generalizes the model by optimizing an arbitrary differentiable loss function. In addition, GBR produces highly competitive, robust, and interpretable results, especially appropriate for mining data that are less than clean, as revealed by Friedman, 2001 [19]. Moreover, the GBR model is deemed to perform better with input features that are complex and nonlinear. Multiple linear regression (MLR) is a machine learning model often applied in agriculture-related research to predict the linear relationship between input variables [15].

Meanwhile, the current research aims to provide fresh insight into anticipating model leaf color dynamics in strawberries using RGB mean values based on plant age and soil physicochemical parameters. To the best of our knowledge, this study is the first of its kind to use the combination of soil nutrients and plant age to indicate strawberry leaf color. The expected results can provide a key technology to support further development of virtual strawberry production and its application in agriculture production.

2. Materials and Methods

2.1. Experimental Design

The present experiment was laid out in the controlled greenhouse at Smart Farm Systems Laboratory, Gyeongsang National University, South Korea, during the winter season in 2021. The overall experiment was 120 days (from October to the end of January). The indoor parameters such as humidity, temperature, light, and CO$_2$ were monitored daily using the specific highly accurate sensor unit MCH 383SD (Lutron Electronic Enterprises Co., Ltd., Taipei, Taiwan) [2]. In this experiment, the combination of bio plus compost and Hoagland solution was used for five rows of strawberries, with 100 strawberry plants in each row, as demonstrated in Figure 1. Moreover, bio plus compost soils consist of cocopeat (68.86%), peat moss (11.00%), perlite (11.00%), and zeolite (9.00%), as revealed by Khan et al., 2019 [20].

2.1.1. Leaf Sample Collection and Soil Physicochemical Parameter Measurement

Normal and colored strawberry leaves (biggest leaf) were collected in consistent normal growth, with no signs of pests and disease, and soil physicochemical measurements, namely, soil pH, EC, ST, and NPK content, were performed near to rootzone using a multifunctional soil sensor (JXBS-3001-SCY-PT, High-precision Environmental Sensors, Weihai JXCT Electronics Technology Co., Ltd., Weihai, China). A total of 400 leaves were collected in normal and colored leaf samples of different ages starting after 60 days of transplanting. Generally, normal leaves are green in color, whereas colored leaves consist of green and non-green parts that are differently colored [21]. A total of 400 leaves were collected every week, 40 leaves for one week, 20 normal leaves and 20 colored leaves,
respectively [5]. The leaf color extraction, model development, and performance analysis procedures were implemented according to the following diagram, as illustrated in Figure 2.

![Figure 1](image1.png)

**Figure 1.** The strawberry experiment in the greenhouse under controlled environmental conditions.

![Figure 2](image2.png)

**Figure 2.** The flow chart of the procedure for the models’ development and analysis of performance using strawberry leaf R, G, and B mean values.

2.1.2. Image Acquisition

Each leaf was placed on a smooth rectangular light chamber (80 cm × 80 cm × 80 cm), which consists of a black-colored surface directly under the white light-emitting diodes (LEDs). The outputs of these LEDs lamps were two 20 W strips. The lamp positions were adjusted so that the leaves were evenly illuminated with no shadows, as indicated in Figure 3. The reflectance of the light band was observed at 450 nm at the upper edge. Concurrently, images were captured using a high-resolution RGB camera with a resolution of 5472 × 3648 pixels (SONY DSC-RX100 vii, Seoul, Korea. The camera was fixed on a tripod 80 cm above the platform’s top, at the nadir position [6].
2.1.2. Image Acquisition

Each leaf was placed on a smooth platform and adjusted so that the leaves were evenly illuminated with no shadows, as indicated in Figure 3. The images were edited using remove.bg software, saving the PNG image as the transparent background and adjusting the image size to 612 × 408. Eventually, for each normal and colored leaf image, the mean value of red, green, and blue channels was computed using an image segmentation algorithm developed by a python program from Google Colaboratory, as shown in Figure 4. Eventually, histograms were developed for leaf R, G, and B mean values at two different ages, 60 and 123 days after transplanting, to obtain the leaf color distribution pattern.

Figure 3. Schematic diagram of the image acquisition system.

2.1.3. Leaf Image Segmentation, Denoising, and Color Feature Extraction

The images were edited using remove.bg software, saving the PNG image as the transparent background and adjusting the image size to 612 × 408. Eventually, for each normal and colored leaf image, the mean value of red, green, and blue channels was computed using an image segmentation algorithm developed by a python program from Google Colaboratory, as shown in Figure 4. Eventually, histograms were developed for leaf R, G, and B mean values at two different ages, 60 and 123 days after transplanting, to obtain the leaf color distribution pattern.

Figure 4. (a) Normal and (b) colored segmented strawberry leaf images for RGB mean extraction under different soil physicochemical conditions.

2.1.4. Data Preprocessing and Models Building

Initially, before data fitting to the machine learning models, a Pearson correlation coefficient heatmap was developed to recognize the magnitude and association among independent variables, such as soil pH, EC, ST, N, P, K, and plant age, and dependent variables, namely, R, G, and B mean values of each strawberry leaf. The scikit-learn library was used to perform data preprocessing and develop the MLR and GBR models, and figures were created using a python program from the Google Colaboratory notebook.

Subsequently, a standard scaler was applied to standardize features by removing the mean and scaling to unit variance to make prediction models based on the results. Thus, before applying the machine learning models, the standard scaler normalized the features of each independent variable. Standard scaling was performed according to Equation (1) [22].

\[ Z = \frac{x - \mu}{\sigma} \]  

(1)

where Z is the standard score, x is the feature value, \( \mu \) is the mean value, and \( \sigma \) is the standard deviation.

Eventually, all variables equally contributed to the model fitting and standardization, and the removal of bias. Subsequently, principal component analysis (PCA) was applied to reduce the dimension of the data sets and reduce the component size by up to 0.95.
2.1.5. Development of MLR and LGBM (GBR) Models

MLR has been more extensively applied in agricultural fields than other prediction techniques [23,24]. The primary goal of the MLR model is to create a linear relationship between the explanatory (independent) and response (dependent) variables. As a predictive analysis, MLR is based on the linear association with more explanatory variables and a response variable, as revealed by Abdipour et al., 2015 [25]. MLR was developed according to Equation (2) [26].

\[
y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon_i
\]  

(2)

where \(y_i\) is the R/G/B mean value, \(\beta_0 - \beta_n\) are the regression coefficients, \(X_1 - X_n\) are the input variables, and \(\epsilon\) is the error associated with \(i\)th observation.

LGBM is a decision tree-based machine learning algorithm that was released by Microsoft in late 2017. The advantages of this model are low memory usage and high convergence speed, and the model has gained increasing popularity in the machine learning field, especially in data science, as reported by Cai et al., 2021 [27]. The LGBM model has used histogram-based algorithm buckets to divide continuous feature values into discrete bins, which fasten the training process, and it splits the tree leaf-wise with the best fit. In contrast, other boosting algorithms split the tree depth-wise or level-wise rather than leaf-wise [28].

For LGBM modelling, GBR, a feature importance method, was performed. GBR is an ensemble learning method that combines multiple weak learners to overcome the overfitting of the model. GBR summation was developed according to Equation (3), as reported by Jiao et al., 2006 [29].

\[
F_m(x) = F_{m-1(x)} + \rho_m h(x; \alpha_m)
\]  

(3)

where \(F_m(x)\) refers to the output and \(m\) represents the number of iterations, \(h(x; \alpha_m)\) refers to the decision tree, \(\alpha\) is the parameter vector of the decision tree, and \(\rho_m\) means the weight parameter of the regressor.

GBR has many parameters that need to be tuned, such as boosting type, max depth, learning rate, and the number of leaves. In the GBR model, max depth is the parameter for the maximum depth of the individual regression estimators. The greater the value is, the more complex the features that the models describe are [30]. However, a high value might result in overfitting the training dataset. The parameter learning rate shrinks the contribution of each tree by the value of the learning rate. The parameter subsample is the fraction of samples that fit the individual base learners. Moreover, the bagging fraction is important to specify the fraction of data for each iteration and is generally used to speed up the training and avoid overfitting. Eventually, the max bin is significant to select the number of bins in which to bucket the feature values [30]. The parameters were adjusted by the trial and error method, and they are listed in Table 1.

| Parameters               | R Mean Value | G Mean Value | B Mean Value |
|--------------------------|--------------|--------------|--------------|
| Estimators               | 1000         | 1000         | 400          |
| Learning rate            | 0.1          | 0.1          | 0.01         |
| Bagging seed             | 0.75         | 0.75         | 0.75         |
| Subsample                | 0.75         | 0.75         | 0.75         |
| Max number of leaves     | 80           | 80           | 80           |
| Max depth                | 6            | 6            | 10           |
| Max bin                  | 10           | 10           | 100          |

2.1.6. Statistical Analysis

In this study, three statistical performance metrics were utilized: coefficient of determination \(R^2\), root mean square error (RMSE), and mean absolute error (MAE). In
regression analysis, $R^2$ is a crucial statistic that indicates how near the predictions are to the actual values, as well as the extent of the regression model bias–variance trade-off [8]. Controversy, the RMSE is sensitive to large perturbations in prediction errors and measures their variations. The third metric, MAE, is a useful illustrator that reveals the average distribution of errors across all model predictions. Furthermore, it also shows how widely the anticipated values are scattered across the entire model, as mentioned by Jaihuni et al., 2020 [31]. The formulas for the metrics are as below.

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (y_{t \text{ actual}} - y_{t \text{ predicted}})^2}{\sum_{t=1}^{n} (y_{t \text{ actual}} - y_{t \text{ mean}})^2}$$  \hspace{1cm} (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_{t \text{ actual}} - y_{t \text{ predicted}})^2}$$  \hspace{1cm} (5)$$

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} (y_{t \text{ actual}} - y_{t \text{ predicted}})$$  \hspace{1cm} (6)$$

3. Results and Discussion

The use of RGB models for strawberry leaf color analysis had shown clear drawbacks in the past. The model’s major flaw was that it had too few parameters to forecast RGB color. The physiological importance of soil physicochemical factors in characterizing leaf color change was not explained [32].

Generally, the parameters that indicate some attribute or trait of leaf color are soil pH, EC, ST, N, P, K, and plant age. The RGB mean value is extracted based on the normality assumption; the leaf color heterogeneity is ignored. Moreover, the mean value can only describe the leaf color state quantitatively [5].

Color alterations from green to yellow are prominent characteristics of leaf senescence [33]; therefore, the color model in this study could be used to explore an approach to prolonging the duration of the functional leaf period and delaying leaf senescence by adjusting the fertilizer rate for enhancing strawberry productivity. Moreover, the color model could be applied to recognize the growth status of strawberries based on the RGB values, which would facilitate the potential application of virtual strawberry production [34].

3.1. Color Feature Extraction

The mean values of strawberry leaf $R$, $G$ and $B$ are plotted at two different ages, 60 days (vegetative stage) and 123 days (reproductive phase) following transplanting, in Figure 5a–c. The skewed pattern shows different leaf ages with increasing and decreasing trends. Two skewed distribution patterns are observed in the last stage of the strawberry lifetime.

3.2. Data Preprocessing Results

The heat correlations between soil physicochemical properties and plant age, and dependent variables such as leaf $R$, $G$, and $B$ mean values are illustrated in Figure 6. The Pearson correlation coefficient technique was used in this investigation, which quantifies the strength of a linear relationship between two variables. It has a range of values from $-1$ to $1$, with $-1$ indicating a total negative correlation, $0$ indicating no connection, and $+1$ indicating an absolute positive correlation. Such kind of correlation aids in denoting the relationship between those dependent and independent variables as strong or weak. According to the heatmap results, positive and negative correlations were observed. As the color becomes darker in either red or blue, those variables are more highly correlated. Based on the heatmap results, color values such as $R$ with $B$ mean, $R$ with $G$ mean, and $G$ with $B$ mean had strong correlation coefficients of 0.76, 0.75, and 0.73, consecutively. Moreover, a high positive correlation was observed with P and K, whereas a strong negative correlation was exhibited between soil pH and $R$ mean, K and plant age, and soil pH and $B$ mean.
values, with 0.82, −0.73, −0.71, and −0.68, respectively. All independent variables were selected for the ML model’s development according to the correlation coefficient values.

Figure 5. Histograms of strawberry leaf color distribution are shown at 60 and 123 days of leaf age: (a) distribution of R means values, (b) distribution of G means values, and (c) distribution of B means values.

Figure 6. Association heatmap of soil physicochemical parameters and R, G, and B mean values.
PCA was used as a statistical means of dimension reduction in feature space. The magnitude of data was reduced, as demonstrated in Figure 7. All the data were spread out among the first and second principal components. The dimensionality reduction increases the accuracy of MLR and LGBM models [30].

\[ R = 107.58 - 0.22 \ (pH) - 9.42 \ (EC) + 1.90 \ (ST) + 3.62 \ (N) - 3.07 \ (P) - 3.59 \ (K) \quad (7) \]
\[ G = 126.65 + 2.03 \ (pH) - 7.78 \ (EC) + 2.64 \ (ST) + 1.66 \ (N) - 3.56 \ (P) + 0.37 \ (K) \quad (8) \]
\[ B = 71.76 + 0.02 \ (pH) - 5.82 \ (EC) + 1.29 \ (ST) + 1.69 \ (N) - 1.91 \ (P) - 2.64 \ (K) \quad (9) \]

where \( R, G, \) and \( B \) mean values are red, green, and blue values of strawberry leaf, and other inputs are soil physicochemical parameters including soil pH, electrical conductivity, soil temperature, nitrogen, phosphorous, and potassium (K), respectively. The constants of the MLR model are 107.58, 126.65, and 71.76 for \( R, G, \) and \( B \) mean values, respectively. The plant age regression coefficient value was zero, and it was not affected the values, more outliers are shown in Figure 8. Moreover, the existence of outliers can be attributed to the inability of the model to predict the strawberry leaf color values properly.

Synchronously, the scatter plots obtained from the GBR model (Figure 9a–c) illustrate the very close distribution pattern with measured and predicted \( R, G, \) and \( B \) mean values. Moreover, the minimum outliers (unusual values of data) in the scatter plot denote the GBR model as the best to predict the strawberry leaf color.
Figure 8. Measured and predicted strawberry leaf color mean values using the MLR model. (a) The Scatter plot of measured and predicted R mean value. (b) The Scatter plot of measured and predicted G mean value. (c) The Scatter plot of measured and predicted B mean value.

Figure 9. Measured and predicted strawberry leaf color values using LGBM (GBR) model. (a) The Scatter plot of measured and predicted R mean value. (b) The Scatter plot of measured and predicted G mean value. (c) The Scatter plot of measured and predicted B mean value.

Consequently, the trained model’s generalizability in predicting RGB mean values linked with planting age and soil physicochemical features was assessed using digital photographs. According to the metrics values, the GBR model exhibited higher performance than that of the MLR model. The overall performance levels of the training and testing models are stated in Tables 2 and 3.

Table 2. Performance training metrics of multiple linear regression (MLR) and gradient boosting regression (GBR) models.

| Color Mean Value | MLR | GBR |
|------------------|-----|-----|
|                 | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE |
| $R$              | 0.68 | 8.67 | 6.73 | 0.99 | 1.40 | 1.07 |
| $G$              | 0.62 | 9.69 | 7.70 | 0.99 | 1.43 | 1.08 |
| $B$              | 0.58 | 7.95 | 5.88 | 0.83 | 5.11 | 3.73 |
Table 3. Performance testing metrics of multiple linear regression (MLR) and gradient boosting regression (GBR) models.

| Color Mean Value | MLR | GBR |
|------------------|-----|-----|
|                  | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE |
| $R$              | 0.67 | 8.59 | 6.63 | 0.77 | 7.16 | 5.53 |
| $G$              | 0.57 | 9.12 | 7.49 | 0.72 | 7.37 | 5.55 |
| $B$              | 0.56 | 6.81 | 5.23 | 0.70 | 5.68 | 4.47 |

Most of the models performed well in RMSE and $R^2$ during the training time. However, the MLR model training phase RMSE values are slightly higher than the testing results due to the randomly allocated training set, and the test set contains data that has not been seen before [15]. The training accuracy was much higher in the GBR model than in the MLR model, and the results were slightly raised in the testing phase. In terms of the percentage difference between GBR training and testing per $R$, mean values were 22.22% less in $R^2$, 80.45% less in RMSE, and 80.65% less in MAE. Regarding the $G$ mean values, the percentage difference between training and testing of $R^2$, RMSE, and MAE was 27.27%, 80.60%, and 80.54%, respectively. Furthermore, the percentage difference between $B$ mean values when training and testing the GBR model were $R^2$, RMSE, and MAE values of 15.66%, 10.04%, and 16.55%, respectively.

Based on statical qualitative metrics ($R^2$, RMSE, and MAE), the results of the study denoted that the GBR model provided a more powerful tool compared to the MLR for forecasting strawberry leaf color, as seen in Table 2. Furthermore, $R^2$ measures the goodness of fit and strength of the relationship. The MLR model is moderately fitted with data, whereas the GBR model is substantially suited with data based on $R^2$ values [35].

Keskin et al., 2018 [36], explored the effect of leaf moisture content on predicting nutrition stress using chromameter color values. In the study, the leaf sample’s color was highly correlated with N, Calcium (Ca), and water content estimated from color data ($R^2 = 0.66$, $R^2 = 0.70$, and $R^2 = 0.65$, respectively) [36].

Previous researchers developed a deep neural network to identify the nutrition stress in plant canopies using spatiotemporal information. Abdalla et al., 2020 [37], proposed the long short-term memory and convolutional neural network (CNN) combined model to classify oilseed rape crops according to nutrition status. The Inceptionv3-LSTM obtained the highest overall classification accuracy of 95% when tested on the dataset of 2017/2018, and it also provided an excellent generalization when using a cross-dataset validation, with the highest overall accuracy of 92%.

Jaihuni et al., 2021 [8], explored cornfield normalized vegetative index (NDVI) information during vegetative and reproductive stages using the UAV and captured plants’ reflectance information. Synchronously, the field’s soil samples N, P, K, and carbon (C) were examined, and a CNN model was developed to predict the infield NPKC spatiotemporal variations. The model performed vigorously with $R^2$ values 0.93, 0.92, 0.98, and 0.83 in predicting N, P, K, and C levels in soil, respectively.

In the current study, the GBR model was more effective in regressing the $R$ mean values related to soil physicochemical and plant age parameters, followed by $G$ and $B$. On the other hand, the RMSE results demonstrated small perturbations in the difference between the predicted and actual RGB mean values. Meanwhile, the MAE values reiterated that the tested model was stable in keeping error rates in predictions to under 10%. Inclusively, it can be deduced from the metrics that the regression process successfully preserved a balance between variance and bias in the model. The results show that the GBR model efficiently imitated and predicted the strawberry leaf color values based on soil physicochemical and plant age parameters.

According to the results of both models, RMSE and MAE metrics are lower for $B$ mean value than for $R$ and $G$ mean values. Therefore, the lower values of RMSE and MAE imply higher accuracy of a regression model [38].
Furthermore, and even more importantly, the novelties of the current work lie in many aspects. Previously, extensive studies were conducted to predict the RGB model relevant to soil macronutrients. The difference is that our concept looks at the relationship between strawberry leaf RGB color and soil physicochemical characteristics and plant age. The developed models were able to quantify the soil pH, EC, ST, N, P, and K optimum values for altering the leaf color. Hence, farmers can determine fertilizer demands by the leaf RGB color information. Moreover, the color-altering problems deriving from fertilizer misuse can be easily monitored and controlled. It facilitates the optimum production of strawberries in the greenhouse under controlled environmental conditions. On the other hand, small datasets and manually acquired images were used to generate comparable RMSE levels in estimating strawberry RGB color related to soil chemical components, which is unlikely to nullify the risk of overfitting in the models; hence, in such ML models, the question of generalizability and reliability need to be scrutinized further with extensive and different datasets.

4. Conclusions

This study developed applicable and stable machine learning models, namely, multiple linear regression (MLR) and gradient boost regression (GBR) models, that vigorously predicted the strawberry leaf color from plant age and soil physicochemical measurements including soil pH, electrical conductivity (EC), soil temperature (ST), N, P, and K when compared to captured digital images from the vegetative and reproductive growth phases of strawberry growth. The GBR model performed better than the MLR model with high-performance metrics. From the results, the GBR provided $R^2$ levels of 0.77, 0.72, and 0.70 for $R$, $G$, and $B$ mean values, respectively. Simultaneously, MLR moderately fitted with datasets with $R^2$ levels of 0.67, 0.57 and 0.56 for $R$, $G$, and $B$ mean values, respectively. Plant age also affected the skewed color pattern. Furthermore, the results indirectly revealed that with an increase in plant age, the strawberry leaf $R$ mean value appreciably increased concerning $G$ and $B$ mean values, which led to an increase in the model performance of $R$ followed by $G$ and $B$ in both models.

As seen in the results, the MLR model was unable to make predictions when the data distribution was beyond the limit, and it focused only on the linear relationship between variables. Nevertheless, the GBR model performs better with input variables that are complex and nonlinear due to its self-adaptive nature.

Our proposed technique has outperformed the benchmark studies while adding some innovative features. This research can be expanded upon by analyzing soil nutrients and mapping them against vegetative and reproductive indicators. Adding diverse soil types and fertilizer levels to a study will also help to offer additional value to it. In addition, it is suggested that future studies also consider the seasonal changes that affect the strawberry leaf color changes.

Author Contributions: Conceived and designed the experiment, performed the experiment, analyzed and interpreted the data, wrote the paper, B.G.K.M.; supervision, J.K.B.; project administration, H.T.K., B.P., N.E.K. and G.M.C. contributed materials and reagents. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Agriculture, Food and Rural Affairs Convergence Technologies Program for Educating Creative Global Leader, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (717001-7).

Data Availability Statement: The datasets generated during and/or analyzed in the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.
References

1. Prasad, R.; Lisiecka, J.; Antala, M.; Rastogi, A. Influence of Different Spent Mushroom Substrates on Yield, Morphological and Photosynthetic Parameters of Strawberry (Fragaria × ananassa Duch.). Agronomy 2021, 11, 2086. [CrossRef]

2. Madhavi, B.G.K.; Khan, F.; Bhujel, A.; Jainihi, M.; Kim, N.E.; Moon, B.E.; Kim, H.T. Influence of Different Growing Media on the Growth and Development of Strawberry Plants. Heliyon 2021, 7, e07170. [CrossRef] [PubMed]

3. Zhang, Y.; Liang, T.; Liu, X.-J.; Liu, L.; Cao, W.-X.; Yan, Z.H.U. Modeling Dynamics of Leaf Color Based on RGB Value in Rice. J. Integr. Agric. 2014, 13, 749–759. [CrossRef]

4. Yang, W.-H.; Peng, S.; Huang, J.; Sanico, A.L.; Buresh, R.J.; Witt, C. Using Leaf Color Charts to Estimate Leaf Nitrogen Status of Rice. Agron. J. 2003, 95, 212–217. [CrossRef]

5. Chen, Z.; Wang, F.; Zhang, P.; Ke, C.; Zhu, Y.; Cao, W.; Jiang, H. Skewed Distribution of Leaf Color RGB Model and Application of Skewed Parameters in Leaf Color Description Model. Plant Methods 2020, 16, 1–8. [CrossRef]

6. Borhan, M.S.; Panigrahi, S.; Satter, M.A.; Gu, H. Evaluation of Computer Imaging Technique for Predicting the SPAD Readings in Potato Leaves. Inf. Process. Agric. 2017, 4, 275–281. [CrossRef]

7. Zhang, B.; Thornburg, T.E.; Liu, J.; Fontana, J.E.; Wang, G.; Li, Q.; Davis, K.E.; Zhang, Z.; Liu, M. Potassium Deficiency Significantly Affected Plant Growth and Development as Well as MicroRNA-Mediated Mechanism in Wheat (Triticum aestivum L.). Front. Plant Sci. 2020, 11, 1219.

8. Jainihi, M.; Khan, F.; Lee, D.; Basak, J.K.; Bhujel, A.; Moon, B.E.; Park, J.; Kim, H.T. Determining Spatiotemporal Distribution of Macronutrients in a Cornfield Using Remote Sensing and a Deep Learning Model. IEEE Access 2021, 9, 30256–30266. [CrossRef]

9. Sim, H.S.; Kim, D.S.; Ahn, M.G.; Ahn, S.R.; Kim, S.K. Prediction of Strawberry Growth and Fruit Yield Based on Environmental and Growth Data in a Greenhouse for Soil Cultivation with Applied Autonomous Facilities. Hortic. Sci. Technol. 2020, 840–849. [CrossRef]

10. Smith, J.L.; Doran, J.W. Measurement and Use of PH and Electrical Conductivity for Soil Quality Analysis. Methods Assess. Soil Qual. 1997, 49, 169–185.

11. Dixon, E.K.; Strik, B.C.; Fernandez-Salvador, J.; DeVetter, L.W. Strawberry Nutrient Management Guide for Oregon and Washington. In Proceedings of the Northwest for Small Fruits Research, Pacific Northwest, Ferndale, WA, USA, 2–4 December 2019; Oregon State University Extension Service: Corvallis, OR, USA, 2019.

12. Milosevic, T.M.; Milosevic, N.T.; Glisic, I.P. Strawberry (Fragaria × ananassa Duch.) Yiel Das Affected by the Soil pH. An. Acad. Bras. Cienc. 2009, 81, 265–269. [CrossRef] [PubMed]

13. Bagale, K.V. The Effect of Electrical Conductivity on Growth and Development of Strawberries Grown in Deep Tank Hydroponic Systems, a Physiological Study. J. Pharmacogn. Phytochem. 2018, 7, 1939–1944.

14. Himelrick, D.G.; Wood, C.W.; Dozier, W.A., Jr. Relationship between SPAD-502 Meter Values and Extractable Chlorophyll in Strawberry. Adv. Strawb. Res. 1992, 11, 59–61.

15. Maesma, À.; Ariza, M.; Martinez, E.; Lloveras, J.; Martinez-Casasnovas, J.A. Analysis of Vegetation Indices to Determine Nitrogen Application and Yield Prediction in Maize (Zea mays L.) from a Standard UAV Service. Remote Sens. 2016, 8, 973. [CrossRef]

16. Özreçberoğlu, N.; Kahramanoğlu, İ. Mathematical Models for the Estimation of Leaf Chlorophyll Content Based on RGB Colours of Contact Imaging with Smartphones: An Example. Folia Hortic. 2020, 32, 57–67. [CrossRef]

17. Singh, A.; Ganapathysubramanian, B.; Singh, A.K.; Sarkar, S. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends Plant Sci. 2016, 21, 110–124. [CrossRef]

18. Zhang, N.; Rao, R.S.P.; Salvato, F.; Havelund, J.F.; Möller, I.M.; Thelen, J.J.; Xu, D. MU-LOC: A Machine-Learning Method for Predicting Mitochondrially Localized Proteins in Plants. Front. Plant Sci. 2018, 9, 634. [CrossRef]

19. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 2001, 29, 1189–1232. [CrossRef]

20. Khan, F.; Okyere, F.G.; Basak, J.K.; Qasim, W.; Park, J.; Arulmozhi, E.; Lee, Y.J.; Kim, H.T. Comparison of Different Compost Materials for Growing Strawberry Plants. In Proceedings of the International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019, Angers, France, 16–20 June 2019; pp. 869–876.

21. Shimoji, H.; Tokuda, G.; Tanaka, Y.; Moshir, B.; Yamashita, H. A Simple Method for Two-Dimensional Color Analyses of Plant Leaves. Russ. J. Plant Physiol. 2006, 53, 126–133. [CrossRef]

22. Loukas, S. How and Why to Standardize Your Data: A Python Tutorial. Available online: https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c382 (accessed on 4 April 2022).

23. Basak, J.K.; Qasim, W.; Okyere, F.G.; Khan, F.; Lee, Y.J.; Park, J.; Kim, H.T. Regression Analysis to Estimate Morphology Parameters of Pepper Plant in a Controlled Greenhouse System. J. Biosyst. Eng. 2019, 44, 57–68. [CrossRef]

24. Basak, J.K.; Okyere, F.G.; Arulmozhi, E.; Park, J.; Khan, F.; Kim, H.T. Artificial Neural Networks and Multiple Linear Regression as Potential Methods for Modelling Body Surface Temperature of Pig. J. Appl. Anim. Res. 2020, 48, 207–219. [CrossRef]

25. Abdipour, M.; Ebrahimi, M.; Izadi-Darbandi, A.; Mastrangelo, A.M.; Najafian, G.; Arshad, Y. Variability and Association Grain Weight with Grain Size (and Shape) and Grain Quality, and Stepwise Regression Analysis on Thousand Grain Weight in Iranian Durum Wheat Landraces. Not. Bot. Horti Agrobot. Cluj-Napoca 2015, 7, 944.

26. Darlington, R.B.; Hayes, A.F. Regression Analysis and Linear Models: Concepts, Applications, and Implementation; Guilford Publications: New York, NY, USA, 2016; ISBN 1462521134.

27. Cai, W.; Wei, R.; Xu, L.; Ding, X. A Method for Modelling Greenhouse Temperature Using Gradient Boost Decision Tree. Inf. Process. Agric. 2021. [CrossRef]
28. Analytics Vidhya Which Algorithm Takes the Crown: Light GBM vs. XGBOOST? Available online: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/ (accessed on 10 March 2022).
29. Jiao, F.; Xu, J.; Yu, L.; Schuurmans, D. Protein Fold Recognition Using the Gradient Boost Algorithm. In Proceedings of the Computational Systems Bioinformatics; World Scientific, University of California: San Diego, CA, USA, 2006; pp. 43–53. [CrossRef]
30. Nagano, S.; Moriyuki, S.; Wakamori, K.; Mineno, H.; Fukuda, H. Leaf-Movement-Based Growth Prediction Model Using Optical Flow Analysis and Machine Learning in Plant Factory. Front. Plant Sci. 2019, 10, 227. [CrossRef] [PubMed]
31. Jaihuni, M.; Basak, J.K.; Khan, F.; Okyere, F.G.; Arulmozhi, E.; Bhujel, A.; Park, J.; Hyun, L.D.; Kim, H.T. A Partially Amended Hybrid Bi-GRU—ARIMA Model (PAHM) for Predicting Solar Irradiance in Short and Very-Short Terms. Energies 2020, 13, 435. [CrossRef]
32. Wenting, H.; Yu, S.; Tengfei, X.; Xiangwei, C.; Ooi, S.K. Detecting Maize Leaf Water Status by Using Digital RGB Images. Int. J. Agric. Biol. Eng. 2014, 7, 45–53.
33. Qiuxia, L.; Gangqiang, C.; Mingjie, S.; Guangyong, Q. Research Progress on Plant Leaf Senescence. Chin. Agric. Sci. Bull. 2006, 22, 282–285.
34. Yadav, S.P.; Ibaraki, Y.; Dutta Gupta, S. Estimation of the Chlorophyll Content of Micropropagated Potato Plants Using RGB Based Image Analysis. Plant Cell Tissue Organ Cult. 2010, 100, 183–188. [CrossRef]
35. Garson, G.D. Partial Least Squares. In Regression and Structural Equation Models; Statistical Associates Publishers: Asheboro, NC, USA, 2016.
36. Keskin, M.; Sekerli, Y.E.; Gunduz, K. Influence of Leaf Water Content on the Prediction of Nutrient Stress in Strawberry Leaves Using Chromatometer. Int. J. Agric. Biol. 2018, 20, 2103–2109.
37. Abdalla, A.; Cen, H.; Wan, L.; Mehmood, K.; He, Y. Nutrient Status Diagnosis of Infield Oilseed Rape via Deep Learning-Enabled Dynamic Model. IEEE Trans. Ind. Inform. 2020, 17, 4379–4389. [CrossRef]
38. Chugh, A. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared—Which Metric Is Better? 2020. Available online: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e (accessed on 13 March 2022).